Google Cloud

Future of AI: Perspectives for Startups

2025

Foreword

Thank you for interest in this eBook.

As a Google Cloud partner, we at [Partner company] are committed to providing you with the resources and insights you need to make informed decisions about your cloud journey.

We believe that this eBook will provide valuable insights into the benefits and capabilities of Google Cloud. [Partner company] is equipped to support you with [value proposition related to the report].

If you have any questions or would like to discuss your specific requirements, please do not hesitate to contact us.

Sincerely,

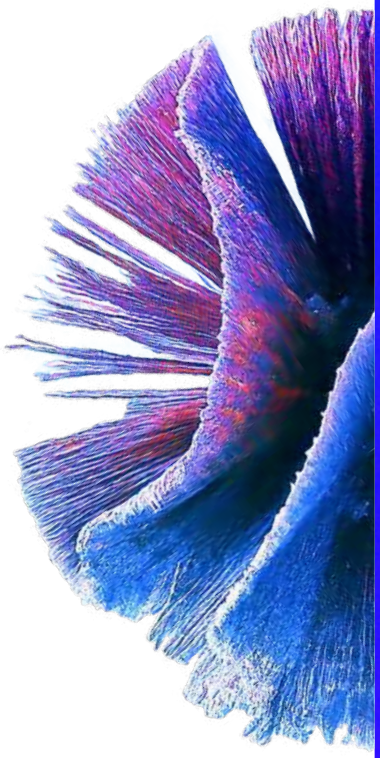
Cirrog

Google Cloud
Partner

Learn more at cirrog.com

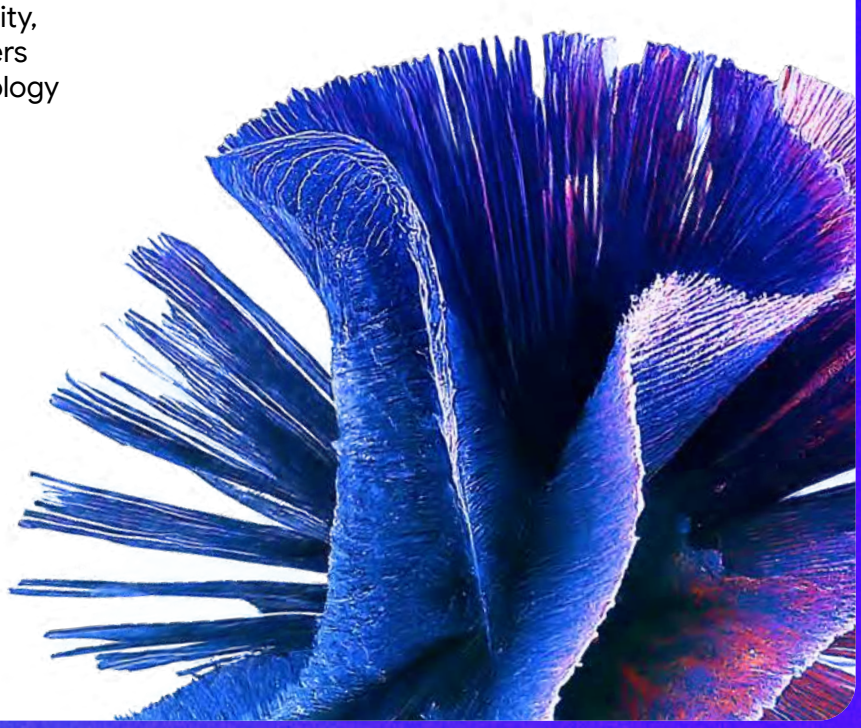
Table of contents

Foreword



AI is transforming every organization around the world and represents an unprecedented opportunity to solve complex problems, drive growth, create efficiencies, and open up new business opportunities. This is particularly true for startups, who are moving very quickly to address new market opportunities with AI. Google Cloud is at the center of AI innovation, and we're proud of our technology leadership that continues to push the boundaries of what's possible for our customers, including more than 60 percent of all funded generative AI startups globally. We are excited to partner with startups, the venture capital community, and industry leaders to ensure that founders and their teams have access to the technology that will help them redefine the future.

Thomas Kurian
CEO, Google Cloud



Google is building all the components of the AI technology stack, from custom chips, to data centers to frontier models. As a result, our new Gemini 2.0 models are more capable, faster and more efficient than previous versions. These models are natively multimodal—they are able to process text, images, audio and video. They can also generate images and steerable text-to-speech audio. With long context windows of up to 2 millions tokens, Gemini can power advanced applications that require deep understanding and memory.

Additionally, Thinking model is capable of showing reasoning skills for solving complex problems, which is especially useful in math and science. Gemini can also natively use tools like Google Search to access real-time information, and DeepMind's Project Mariner has demonstrated that agents built with the Gemini model can complete tasks using a web browser. Conversational experiences can now be built with the Gemini Multimodal Live API, which accepts audio and video streaming input. The combination of these capabilities enables a new class of agentic experiences and we're excited to see what startups build with Gemini in 2025.

David Thacker
VP, Product, Google DeepMind

AI predictions

Amin Vahdat
VP/GM ML, Systems, and Cloud AI, Google Cloud

Tight synchronization and massive compute requirements will push infrastructure to never-seen-before levels of compute density and capability.

Apoorv Agrawal
Partner, Altimeter Capital

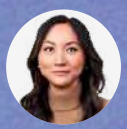
By combining voice, vision, and natural language, multimodal AI will reduce the need for devices like computers and cell phones and make interacting with the digital world more seamless.

Arvind Jain
Founder and CEO, Glean

AI will continue to be a tool to augment human capabilities, not replace them. The concept of AI-based employees perpetuates a limited perspective that hinders the true potential of both AI and human intelligence.

Chamath Palihapitiya
Founder and CEO, Social Capital, and Co-Founder and CEO, 8090

The future of software is about doing more with less. With AI automation, the software industry will get more efficient and the average profit margin of the S&P 500 will double as companies get more and pay less.



Crystal Huang
General Partner, GV

Micro-booms and busts in AI will be inevitable as the tooling for building generative AI applications becomes more readily available and therefore commoditized.

David Friedberg
CEO, Ohalo Genetics

The industries I think are most susceptible to AI-driven transformation are media, SaaS and biology (therapeutic drugs and agriculture). For example, genome language models will be able to predict the exact DNA sequence needed for any desired plant trait or biologic drug, revolutionizing agriculture and human health.

Douwe Kiela
CEO, Contextual AI

Long context and RAG will converge. Models may learn how to decide when to use long context and when to use RAG to optimize both accuracy and efficiency.

Dylan Fox
Founder and CEO, AssemblyAI

The timeline for widespread enterprise adoption of AI will be slower than people think. A lot of last-mile issues need to be solved that aren't obvious until you're deep into them.

Edo Liberty
Founder and CEO, Pinecone

The fastest ROI in AI is in agents, but the biggest opportunity is in enterprise search.

Elad Gil
CEO, Gil Capital

AI is massively underhyped. We are in the early stages of a huge transformative wave driven by AI, with traditional machine learning, such as self-driving cars, finally hitting its stride, and generative AI in its infancy.

Harrison Chase
CEO and Co-Founder, LangChain

To truly leverage agentic systems, we will need them to scale beyond just what we can ask them to do—to be 'ambient agents', running in the background, always on, monitoring streams of events, and alerting me only when something interesting happens.

James Tromans
Managing Director, Web3, Google Cloud

You may end up trusting your web3-enabled AI agent more than you trust anyone else. You'll tell it way more things than you would tell someone else. It will know all about you and use that information to assist you.

Jennifer Li
General Partner, a16z

It will take longer than people expect for AI agents to truly go mainstream. Model reasoning capabilities will need to improve, but it's really the infrastructure that will be leveraged by agents and deep-rooted systems integration problems that will need to be solved.

Jerry Chen
Partner, Greylock

The final business models for AI have not been pioneered yet, and it will take time for investors to understand them.

Jia Li
Co-Founder, President and Chief AI Officer, LiveX AI

I envision a future where AI agents seamlessly integrate into daily life and our interactions with AI become more human-like—with agents that can not only understand language but also perceive and respond to visual cues.

Jill Greenberg Chase
Investment Partner, CapitalG

Foundation models will remain fairly static in their capabilities for the next 18 months, leaving ample opportunity for startups to build specialized AI solutions that deliver a clear return on investment.

Matthieu Rouif
Co-Founder and CEO, Photoroom

AI will understand and adapt to human emotion. AI will get better at understanding what triggers emotions in humans, allowing for stories and content to be personalized and adapted to individual emotional responses.

Mayada Gonimah
CTO and Co-Founder, Thread AI

We're seeing a sudden explosion of models, solutions and tooling that are being rushed to market when they haven't proven to actually move the needle —whether with ROI or by fundamentally changing the developer ecosystem. A lot of these libraries are going to die or get less and less funding.

Raviraj Jain
Partner, Lightspeed

The most significant technological changes will occur in foundation models for the physical world, especially in robotics.

Salim Teja
Partner, Radical Ventures, and Board Member, Aspect Biosystems, Promise Robotics, Intrepid Labs

The landscape will shift from building bigger and better models to prioritizing the widespread implementation of AI across industries to solve real-world problems and generate tangible value for businesses and society.

Sarah Guo
Founder and Partner, Conviction

Mike Vernal
Partner, Conviction

When it comes to AI, the most useful data is often tied to specific, real-world use cases. In many instances, that leaves the field open to new entrants.

Yoav Shoham
Professor Emeritus of Computer Science, Stanford University, and Co-Founder, AI21 Labs

The early days of "prompt and pray" AI—simply feeding a language model a prompt and hoping for a good result —are over. To meet the critical demands of enterprises, we need robust "AI systems" that orchestrate multiple models and tools for reliable results.

Advice for founders

Computation will be much faster, cheaper, and more reliable as time goes on.

Let's say you have this amazing idea, all you need is to reduce compute costs by a factor of 10 or even 100 for it to be profitable. That is now within reach.

Amin Vahdat
VP/GM ML, Systems, and Cloud AI, Google Cloud

Align pricing with value delivered.

Don't just use a per-seat model, but consider usage-based or value-based pricing. Your pricing should reflect the value your product provides to users.

Apoorv Agrawal
Partner, Altimeter Capital

View AI as a way to increase topline revenue and open up new opportunities for innovation, not just as a tool to drive efficiency and cut costs.

Arvind Jain
Founder and CEO, Glean

Focus on building point features, not point products.

Prioritize developing high-value, specific functionalities instead of creating comprehensive software. This allows competition in markets where larger companies could replicate features using AI.

Chamath Palihapitiya
Founder and CEO, Social Capital, and Co-Founder and CEO, 8090

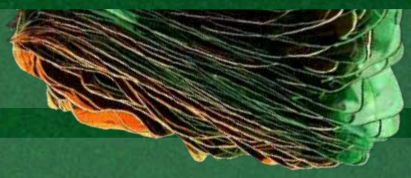
The most impressive founding teams are fanatical about documenting, codifying, and measuring their processes.

They want to be the best at testing out best practices, writing down what works and sharing it across the organization. I've found that having a culture that appreciates documentation (within reason!) is related to operational excellence in other ways.

Crystal Huang
General Partner, GV

You need some engine of value creation that gives you the ability to persist an initial advantage following the launch of an AI-enabled offering, either data generation for continuous performance improvements, or network effects. It's not enough to just be an LLM wrapper.

David Friedberg
CEO, Ohalo Genetics



Move fast. Time to market is the most important thing for a startup, and the market is now.

Things are moving incredibly fast in AI, so you have to make sure you're not left behind.

Douwe Kiela,
CEO, Contextual AI

Instead of aiming for general-purpose AI, focus on solving specific niche problems with a lot of depth.

Dylan Fox
Founder and CEO, AssemblyAI

Start with out-of-the-box solutions.

Instead of building everything from scratch, startups can use existing tools to quickly get a production-ready product that they can test and monetize.

Edo Liberty
Founder and CEO, Pinecone

Find something people really care about, ship it as fast as you can, test whether people care or not, then iterate from there.

Elad Gil
CEO, Gil Capital

Prioritize evaluations. Develop evaluations immediately to better scope the problem. Having clear metrics and ways to assess the performance of an AI system are crucial from the outset.

Harrison Chase
CEO and Co-Founder, LangChain

How you monetize and sell AI is as important as how you build it.

If you can flip the business model against your competitors, that's powerful because it's hard for incumbents to switch.

Jerry Chen
Partner, Greylock

From the outset demonstrate strong product-market fit, a distinct competitive advantage, and a clear path to profitability.

Jill Greenberg Chase
Investment Partner, CapitalG

Companies do not need decentralization for decentralization's sake.

Instead, apply the benefits of web3 —such as cheaper payment rails, auditability, and transparency—to AI agents solving customer use cases.

James Tromans
Managing Director, Web3, Google Cloud

Understand model capability and marry it with what users want, that's the magic.

Jennifer Li
General Partner, a16z

Do not underestimate the power of data and focus on collecting diverse, high-quality data while ensuring security and privacy.

Jia Li
Co-Founder, President and Chief AI Officer, LiveX AI

Design your user experience in a way that helps people accomplish their goals without having to use prompts.

Matthieu Rouif
Co-Founder and CEO, Photoroom

Take an opinionated view of where AI capabilities, tooling and infrastructure are going to go and build a product that bridges the gap on that capability today, keeping in mind that the underlying engine will improve significantly.

Raviraj Jain
Partner, Lightspeed

Build with 'agnostic' infrastructure that lets you take advantage of the best-in-class models and databases that are constantly changing.

Mayada Gonimah
CTO and Co-Founder, Thread AI

Invest in applied AI research capabilities, understand the AI technology landscape, identify key opportunities for differentiation, focus on solving real-world problems, and be prepared for the rise of agentic AI.

Salim Teja
Partner, Radical Ventures and
Board member, Aspect Biosystems, Promise Robotics, Intrepid Labs

Be cautious in the middle:

Be aware of the different layers of the AI stack (foundation models, middleware, dev tools, and applications) and consider where your company fits. Be cautious about building a company in the middle layer, because it can be under pressure if the foundation models evolve rapidly.

Sarah Guo
Founder and Partner, Conviction

Mike Vernal
Partner, Conviction

Focus on "product-algo fit".

Understand the strengths and weaknesses of the current technology and build your startup around what the AI technology can do well today, not on some idealized future version.

Yoav Shoham
Professor Emeritus of Computer Science,
Stanford University, and Co-Founder, AI21 Labs

No matter where you are with AI adoption, we're here to help.

Book your generative AI consultation today.

Sign up now

Get up to $350K in cloud credits with the Google for Startups Cloud Program.

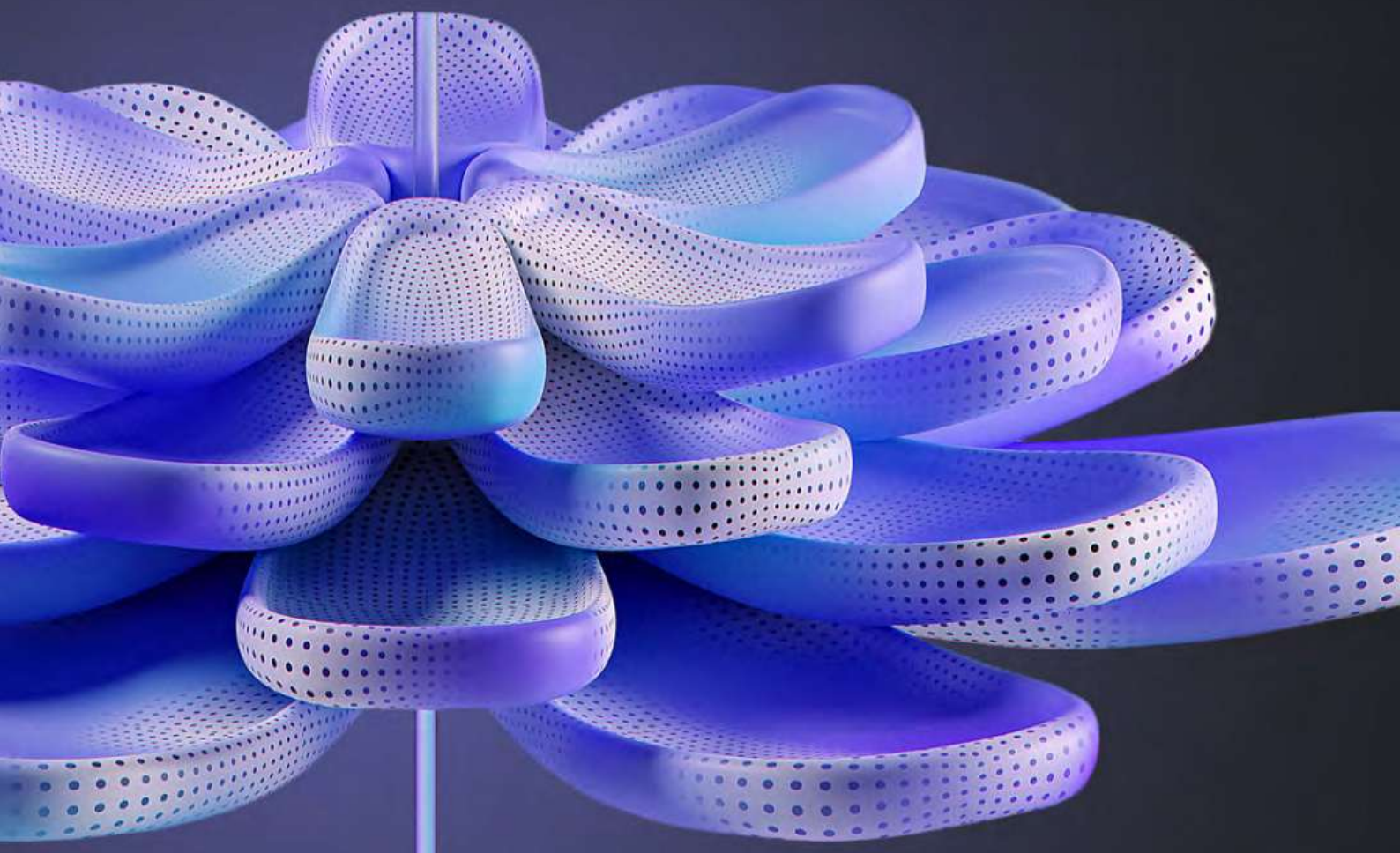
Apply now

Contact our Startup sales team.

Get in touch

What's next in AI

Perspectives of industry leaders



Amin Vahdat

VP/GM ML, Systems, and Cloud AI, Google Cloud

Amin is an Engineering Fellow and Vice President for the Machine Learning, Systems, and Cloud AI team.

Before joining Google nearly 15 years ago, Amin was the SAIC Professor of Computer Science and Engineering at UC San Diego. He is an ACM Fellow and a member of the National Academy of Engineering.

10 years from now our infrastructure will be unrecognizable

From software to hardware, fundamental shifts across compute, networking, storage, and beyond are underway

Today, we take our existing architecture for granted as it represents the "conventional wisdom" for how systems should be built. Changes underway over the past five years or so mean that we are in the early-to-middle stages of another revolution that will render the emerging computing platform unrecognizable by the end of this decade.

One of the biggest drivers of this transformation is the shift away from real-time information access and delivery to the proactive generation of insights and interpretation driven by generative AI.

The tremendous computation and communication demands required by generative AI necessitate corresponding shifts in the design and build of underlying infrastructure up and down the stack, including:

01 Application-specific computation

Generative AI computation consists of immense volumes of matrix multiplications and related numeric computations. This specialization means that purpose-built hardware for a very specific subset of general purpose computation can deliver a factor of ten or more compute efficiency when measured in performance normalized to cost or performance per Watt.

02 Networking specialization

We're seeing a sea change from standards based, commodity network hardware, protocols, and software to specialized networks, such as Inter Chip Interconnect (ICI) for TPUs and NVLink for GPUs, designed for the very specific task of supporting the higher-level compute primitives, such as all-reduce in hardware with little-to-no layering, no OS management, and essentially direct memory-to-memory transfers in support of specific computation primitives.

03 Memory wall

High Bandwidth Memory (HBM), where RAM is stacked in 3D on the same package as the compute to dramatically reduce latency and increase bandwidth, is hitting its own fundamental limitations—the memory wall. Growing compute without commensurate ability to feed the compute with the required data, either from local memory or from across the network, will result in the expensive stranding of both compute and power—the need for breakthroughs in compute and memory architecture to maintain the needed system balance points.

04 Compute packing

Tight synchronization and massive power requirements push infrastructure to never-seen-before levels of compute density. Computation must be across homogeneous elements, communication must be pre-planned and coordinated, and fault tolerance must be efficient. From a latency and power/cost perspective, packing ever larger amounts of computation power into smaller, more power-dense environments can have significant benefits.

05 Liquid cooling

The sustained rate of intense computation for ML compute shifts the equation such that improved performance efficiency of running chips faster easily justifies higher power density and liquid cooling. Liquid cooling in turn requires a redesign of data center buildings and cooling infrastructure, racks, and more end-to-end to deliver the highest levels of efficiency, alongside off-grid power generation technologies (including wind, solar, hydro, and battery arrays) to enable the delivery of ML compute in various locations across the planet.

> “Out-of-the-box capabilities with advanced reasoning capabilities, like orchestrating LLMs with retrieval-augmented generation (RAG) and function-calling, will enable startups to access models and AI infrastructure without massive capital expenditure or engineering resources.

AI infrastructure evolution will impact startups in different ways

This tremendous growth in computing capability is accelerated by continuous improvements in model reliability and decreasing computation costs, as a result:

01 More progress will be made on model quality, safety, and latency

If your startup is focused on veracity and your goal is to bring high-quality, low-or zero-hallucination offerings, then you should still focus on this issue. But if your goal is to deliver some other service, leverage the rising tide of quality that all the models are now providing.

02 Computation costs will plummet

The cost of computation is dropping quickly. If I were an AI startup, I'd bet on the future of computation being much faster, cheaper, and more reliable. Let's say you have this amazing idea and all you need is to reduce computation costs by a factor of 10 for it to be profitable. To me, that's an easy bet—even if you need a factor of 100 reduction.

As we've seen across the industry, model builders will continue to optimize for cost savings, utilizing bare-bones infrastructure with tiny margins. Availability of a powerful and diverse set of accelerators (GPUs and TPUs) will continue to be key, as well as high-throughput networking capabilities highlighted earlier.

The vast majority of startups, focused on building innovative and differentiated applications, will shift further up the stack to more software-based services. Out-of-the-box capabilities with advanced reasoning capabilities, like orchestrating LLMs (Large Language Models) with retrieval-augmented generation (RAG) and function-calling, will be in higher demand as they allow access to models and infrastructure without massive capital expenditure or engineering resources.

Apoorv Agrawal

Partner, Altimeter Capital

Apoorv leads investments in software and AI startups with notable investments in OpenAI, Glean, Parloa, and Everest.

An engineer by training, Apoorv started his career as a Forward Deployed Engineer at Palantir and still codes.

Computer automation and multimodal AI are the future of personal computing

The next inning of AI growth will automate the mundane and humanize technology

We're in a period where we've become addicted to computers and phones. I believe we'll look back on it as a strange blip in human history. In a hundred years, people won't be glued to their screens. AI will liberate us. Particularly, the rise of multimodal AI—or AI that combines voice, vision, and natural language processing—is going to make interacting with the digital world completely seamless.

Imagine a world where you can control your entire digital life conversationally—speech, visual cues, touch, and when all else fails, typing. That's the future we're building, and it's going to unlock big opportunities for startups in the years ahead.

Betting on automation, efficiency, and human-centered experiences

As an AI-focused investor who believes in a future driven by multimodal AI, I've identified three core themes that guide my decision-making:

01 Automating the mundane

Customer service, back-office functions, and anything that's data-heavy and rules-based is ripe for AI disruption. I'm interested in making AI investments in areas where humans are doing work that does not fulfill them, does not bring them joy, and is not creative. AI excels at automating unfulfilling, data-rich tasks, freeing up humans for more meaningful work.

02 Augmenting human capabilities

AI is about making us smarter and more efficient. Think of tools that can synthesize vast amounts of information and deliver actionable insights to unlock tangible productivity gains. I'm interested in startups developing AI solutions that help improve what humans can do.

03 Humanizing technology

I think voice AI in particular offers a more natural and intuitive way to interact with technology. I have focused on voice AI infrastructure providers because I believe this area is exploding. Startups that can create truly human-centered AI experiences will have a big advantage in the market. And another interesting area to explore is how we can use AI to better understand and connect with each other, improving relationships beyond productivity gains.

Similar to these themes, I'm also interested in agentic workflows, where AI agents handle entire tasks or processes for us. It's no surprise to me that some of the smartest minds are building tools to free us from clicking buttons and being hunched over a screen.

> "I'm interested in making AI investments in areas where humans are doing work that does not fulfill them, does not bring them joy, and is not creative. AI excels at automating unfulfilling, data-rich tasks, freeing up humans for more meaningful work.

Shifting focus to the application layer

As the cost of infrastructure and models plummets, the opportunity is shifting towards the application layer. Think of it like this: the AI stack is currently an inverted triangle, with most of the value concentrated in semiconductors and infrastructure. But as we move into the next innings of AI growth, that triangle will flip. The application layer will become where innovation and value creation happen.

For startups, this means focusing on building AI-powered products and services that solve real-world problems. Don't get bogged down in the complex world of foundational models. Instead, leverage existing infrastructure and focus on creating innovative applications that deliver tangible value to users. And build for durability, so you're not just surviving but winning with the next generations of models and supporting technology.

I think this is where the real opportunity lies and see two big trends emerging:

01 Horizontal applications

AI-powered tools that improve productivity across all job functions.

02 Vertical applications

Specialized AI products built to cater to specific roles like engineers or customer service representatives.

One key debate is whether horizontal AI will become so good that it replaces the need for vertical solutions. Another question startups need to consider is, should you pursue per-seat, usage-based, or value-based pricing for your product? There's no easy answer, but I advise startups to align their pricing with the value delivered.

For my final thoughts, I'll leave you with what I tell founders who ask me for advice on using AI effectively:

Data

There's no AI strategy without a data strategy. Understand what data you have and how you can use it to train effective AI models.

Integration

Consider how you can integrate your AI solution with other platforms and data sources to create a seamless user experience.

Workflow

Finally, think about the tasks that people don't want to do—like booking flights. Can AI automate those tasks or make them more intuitive?

Arvind Jain

Founder and CEO, Glean

Arvind is the Founder and CEO of Glean, the AI-powered work assistant that brings people the answers they need to be more productive and happier at work.

Prior to Glean, Arvind co-founded and led R&D at Rubrik, a publicly traded leader in the data security space and also spent over a decade at Google as a distinguished engineer, where he led teams in Google's Search, Maps, and YouTube products.

Prioritize topline growth with AI, ROI can wait

AI is about building products and providing experiences that were not possible before

The conversation around AI agents has become unrealistic. At some point last year, what started with 'AI can answer this question' quickly became 'agents are going to do our work for us so we can go to the beach,' with little proof or thought on how we get there.

Almost every Founder or CEO I speak to knows that we are in AI's early stages and that it will fundamentally change their business over the years. I also believe that humans will remain in control, mastering this technology for the foreseeable future. Individuals will become significantly more productive, perhaps even achieving ten times more than they could before. It's crucial to understand that AI should be viewed as a tool to augment our capabilities, not replace them.

In this world of human and AI collaboration, I have a few pieces of advice for founders and CEOs trying to find the right balance between the two:

01 Stop obsessing over ROI

AI is still in its early stages, and this is not the time to think about a hard ROI from your investment in it. While you might have some projects focused on the bottom line to achieve savings, you should reinvest those savings into projects that can generate substantial growth. AI's sole destination is not to achieve 20% efficiency by having fewer customer service agents or by automating some business processes. AI is also about increasing your topline by doing things and building products you were never able to do or build before. The opportunity is clear: Use AI to unearth savings through some efficiency work, and then reinvest that into new innovations that open up new opportunities for you.

02 Ensure AI fluency across your team

As a startup, not only do you have to make sure your team is at the forefront of the latest AI tools and technologies, but that your employees feel comfortable using them. This isn't just about adopting new technologies, it's about fundamentally changing how your company is organized and built. Your employees need to be educated on how to use AI, and see it as a partner. You also should look to hire people who are already comfortable using AI while ensuring that all of your employees are empowered with the right tools and education to successfully integrate AI into their workflows.

03 Approach AI as a tool, not as a product

Any product you build needs to embrace AI to be smarter, more powerful, and modern. As a startup you need to think of AI first as a tool. If you develop a product that uses AI to achieve 90% of its function, you will set your startup up to be replaced completely. Start with the problem (rather than the function) and deploy AI to solve that problem better.

04 Account for the rate of change

AI technology is constantly evolving. When you build your product roadmap, ask whether the problems are something you should solve yourselves, or whether AI will solve them itself in that timeframe or foreseeable future. Make those calculated decisions on where to invest R&D to anticipate those advances in core technology, rather than spending time on things that will soon become obsolete. The same applies to how you build. Design your infrastructure to be more agnostic to model and tooling advancements, predicting what's coming and enabling plug-and-play where feasible to take advantage of updates without massive overhauls or disruptions.

> "AI is not just about efficiency. The bigger opportunity is about increasing your topline—doing things and building products you were never able to do before.

Chamath Palihapitiya

Founder and CEO, Social Capital, and Co-Founder and CEO, 8090

Chamath is Founder and CEO of Social Capital, a technology investment company, and Co-Founder and CEO of 8090, an AI company.

Before founding Social Capital in 2011, Chamath was a Senior Executive at Facebook from 2006-2011, where he was the leader of the company's Growth, Mobile, and Platform teams.

Why a shrinking software market can fuel greater profits

The future of software is about using AI to do more with less, empowering businesses to become more efficient and profitable. In helping companies fulfil this, many will be willing to share the upside with the software companies that actually create this value

I think AI is more than just an ongoing shift: it's a hyper-disruption.

The traditional software industry, with its complex products, high cost, balkanized features, and licensing models, is being completely reimagined. AI has the power to streamline this industry and shrink the multitrillion-dollar "Software Industrial Complex" to a fraction of its size. I don't believe this is a bad thing. It's an opportunity for well-run businesses to become more profitable and potentially share some of this upside with the companies that enable this transformation.

Like many of the best-run companies today, all companies should be able to build their own custom solutions—aka their own Business Operating Systems—that are tailored specifically to their needs, rather than rely on off-the-shelf products that require expensive change management and customization.

The real value creation in AI is in creating a software factory that can take business requirements as raw material input and generate high quality production code as output. When this is possible, businesses can entirely reframe their org chart.

To me, it's a given: the future of software is about doing more with less. It's about automating, and ultimately empowering business leaders to achieve their goals with more simplicity and quality.

In the fast-evolving landscape of AI, where models and tools are constantly being updated and replaced, startups need to prioritize flexibility. Building solutions that are adaptable and future-proof is crucial, since today's cutting edge could be obsolete tomorrow.

This means adopting a modular approach to development, where components can be easily swapped out or upgraded. It also means staying informed about the latest advancements in AI to avoid the pain of constantly redoing work, and to take advantage of new opportunities as they arise.

> “
>
> **The real value creation in AI isn't in building products to address specific problems. It's in creating automated software factories that can take business requirements as raw materials and generate production code as output.**

A change to value creation

We're moving towards a world where the reward function isn't based on clicks or engagement, but on maximal truth. I expect this shift to fundamentally change how we approach problem-solving and value creation.

Imagine a future where fully autonomous agents can tackle complex tasks. For example, software that takes a company's product inventory and generates all the commerce infrastructure needed to sell those products across multiple countries, with all kinds of complicated privacy rules, in less than two hours.

We're already seeing the emergence of one-person companies working alongside AI. The people who run these businesses use AI to automate tasks, generate content, and even manage customer interactions. I interpret this as a testament to the democratizing force of AI, enabling people to build and scale businesses like never before.

The key takeaway is this: focus on what truly adds value and cut out the unnecessary. AI can help us identify and eliminate inefficiencies, streamline workflows, and optimize processes.

Opening up new paths to value

The conventional path for startups to solve specific inefficiencies and problems can still be very lucrative. That said, I see three early adopter archetypes of AI:

01 Established enterprises with significant IT budgets and low ROI from their existing systems.

02 Companies experiencing rapid growth that outpaces their ability to attract engineers or build internal infrastructure to keep pace.

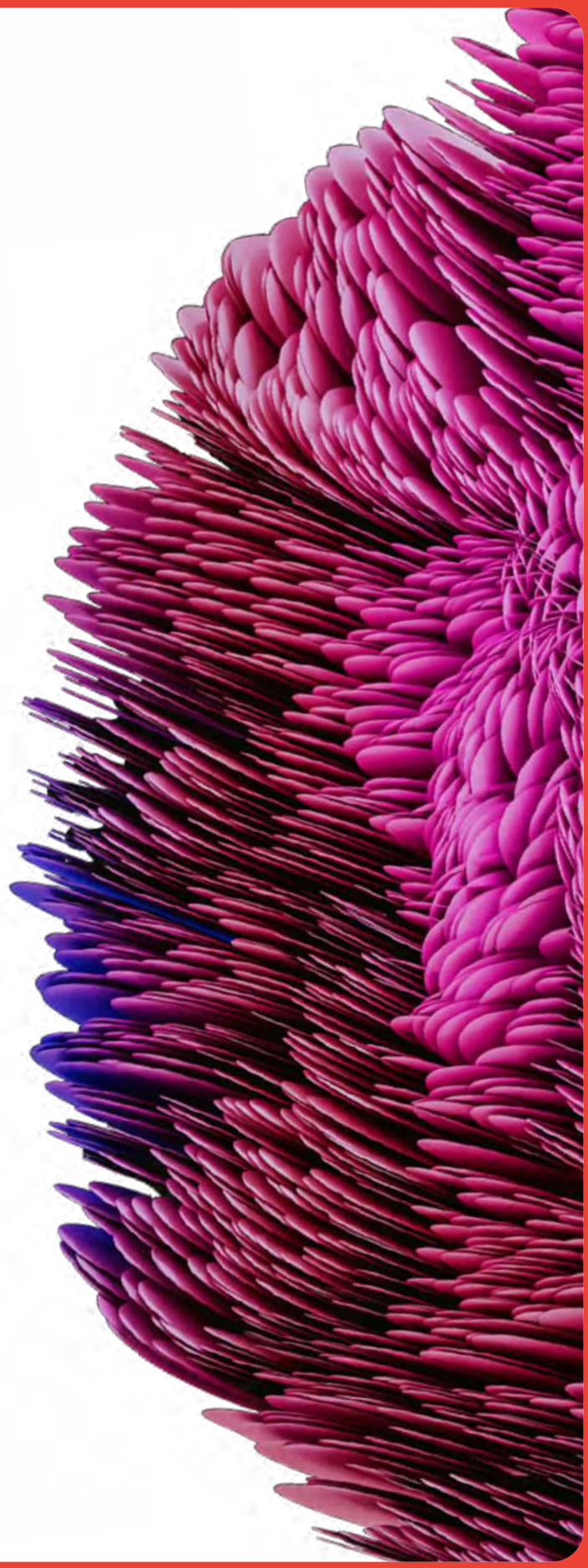
03 Turnaround companies, that are under pressure to restructure and reduce costs, making them more open to innovative solutions.

These diverse customer profiles demonstrate the expansive potential of AI across various businesses and stages of growth. This creates a fertile ground for innovation, allowing startups to leverage AI to disrupt existing markets and open up new frontiers.

Build software factories, not products

For startups, this presents a unique challenge and opportunity. The traditional model of building point products may no longer be enough. Instead, startups need to think bigger, focusing on flexible solutions that can be deployed quickly and efficiently.

AI is changing everything we thought we knew about software. It's time to embrace a new way of thinking, one that prioritizes simplicity, automation, and above all, value creation. The future of software is about doing more with less, and AI is the key to unlocking that potential.



Crystal Huang

General Partner, GV

Crystal is a General Partner at Google Ventures, where she concentrates on investments in AI, SaaS, and infrastructure, with a special interest in product and developer-led adoption strategies.

Previously, Crystal was an investor at NEA and Notable Capital (fka GGV) and began her career in technology M&A at Blackstone.

AI-powered hyperpersonalization is coming in 2025

...but will it be enough to drive stickiness?

While AI is still innovating at a breakneck pace, the hype and investment cycles for startups in this space have become the shortest of any technology trend I've ever been a part of.

I've never seen so many companies pick up crazy traction at launch, get to tens of millions of annual revenue, and then have people lose interest in such a short amount of time. It could be that a week later a competitor launches and it's a little better. It could be that a foundation model company launches the same functionality at their next big release. I think this boom and bust cycle is going to continue because the tooling is so broadly accessible to build an application.

It's difficult and costly to build a new foundation model so not many teams will be able to do it, but at the application layer, I think there's going to be tons of disruption and rebirth.

It's exciting that the 2025 landscape will look nothing like last year, and while generative AI is obviously exciting territory for investors, the standard valuation framework still applies.

Here's what I'll be looking for:

Stickiness as a metric

If your product is easy to implement, it's just as easy to uninstall. Products need to be stickier to create lasting value, which means being both indispensable and deeply integrated into the user's workflow. And while the enterprise is driving all the urgency in AI, it often requires mutual effort between the platforms and enterprises—working together to create automation or workflows, and gaining difficult access to enterprise data to significantly boost performance.

The fastest growth is stemming from individuals who are willing to experiment with something new for $15 a month, rather than an enterprise investing in a legacy system for $20m a year. But while financially accessible entry points can lead to more rapid consumer adoption, this can make it equally challenging to keep them from trying something new.

“
If your product is easy to implement, it's just as easy to uninstall. Products need to be stickier to create lasting value, which means being both indispensable and deeply integrated into the user's workflow.

Hyperpersonalization

We've been hearing about a future where every consumer will get AI-powered marketing campaigns or tutoring or healthcare services that are highly personalized and generated on-the-fly, but that hasn't happened yet. The high costs of training and inference for AI models were major obstacles, but these are rapidly decreasing at a rate that I think is quite magical. Training expenses are dropping as smaller and more domain-specific models are emerging, and inference costs are plummeting across the board—making personalized reasoning and decision making not only economically viable but potentially differentiator for AI companies looking to retain users over time.

There's one last thing I'll add: enterprise customers are surprisingly savvy about how quickly some AI capabilities are commoditizing. I've spoken to CIOs and CTOs who are onboarding one generative AI solution this year but already planning an RFP for a cheaper and more powerful solution next year. It's no longer enough to simply raise capital; startups need to demonstrate true product ROI, generate meaningful revenue, build a defensible moat, and, at the end of the day, deliver on the hype.

David Friedberg

CEO, Ohalo Genetics

David is CEO of Ohalo Genetics, having co-founded the company in 2019 while CEO of The Production Board, a venture foundry that builds and invests in technology businesses in food, agriculture, and life sciences.

Prior to Ohalo and TPB, David was founder and CEO of The Climate Corporation, the world's leading digital agronomy platform, today used by farmers across 200+ million acres worldwide. Monsanto acquired The Climate Corporation in 2013, when he joined Monsanto (now Bayer) as a member of the executive leadership team.

David is also one of four co-hosts of All-In, a top business and investing podcast. Earlier in his career, he held senior roles in Corporate Development and Product at Google and graduated from University of California, Berkeley with a Bachelor's degree in Astrophysics.

Regulation and how AI is revolutionizing hardware, media, software, and biology

What I think of regulation and how an AI-driven future includes robots everywhere, an end to traditional media, no more SaaS subscriptions, cheaper food and breakthrough drugs

Let's start with a controversial area: AI regulation. Some people believe that AI should be heavily regulated to prevent its misuse, while others believe that regulation will stifle innovation. My take? The world won't regulate AI uniformly, so the industries or jurisdictions that apply less regulation will gain advantages and win markets. I think you need to make AI free to operate for it to be competitive and successful against the bad actors that try to use AI for harm, relying on existing laws to guard against the inappropriate use of AI, rather than try to control the development of AI.

Some AI companies will push boundaries, others will prioritize safety. Every company will define its own "responsible AI," and that's healthy. We need diverse AI systems with varying constraints, just like we have diverse media outlets. Let the market, not regulators, decide what it wants.

Further, trying to regulate how models are constructed will never keep up with the pace of innovation in AI.

Huge potential for disruption: robotics, media, SaaS, and biology

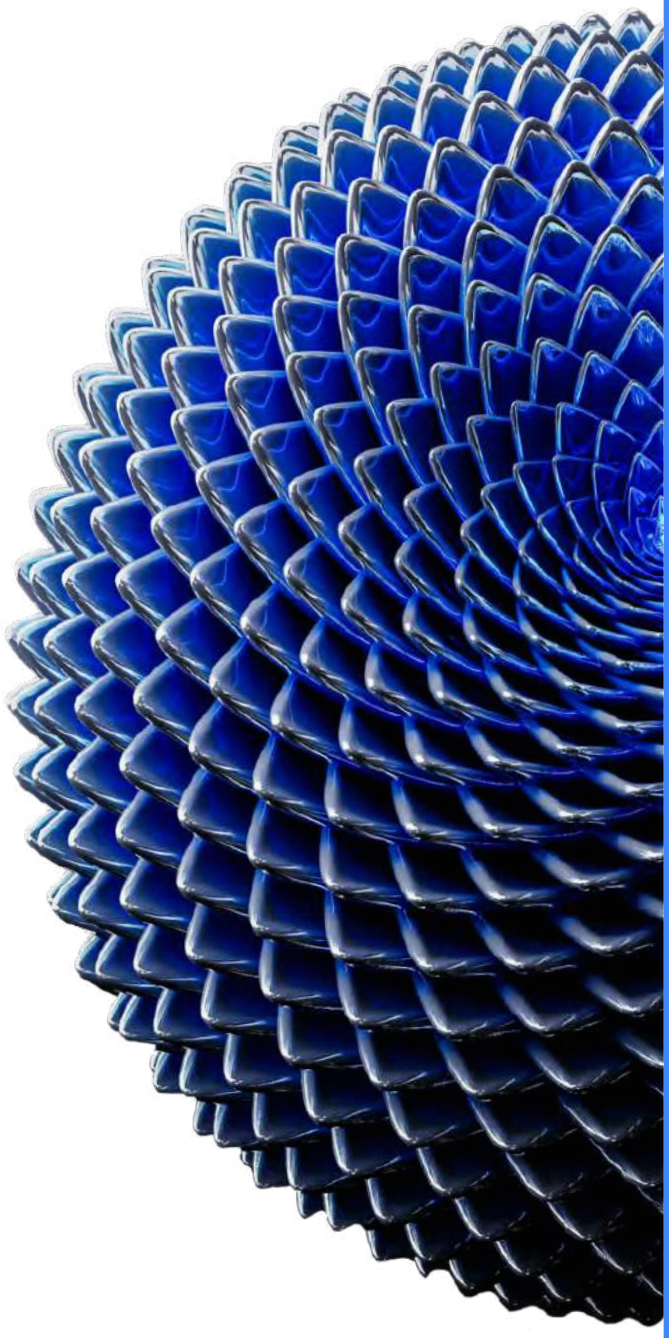
Here are a few sectors where I see massive potential for disruption:

01 Hardware

2025 will be the Year of the Robot. We're witnessing a seismic shift in AI's impact in automation through machine vision and self-learning. While 2024 focused on build-out of compute and innovation in foundational models, 2025 is going to be the year of application and utilization of real-world and simulated data to train and deploy hardware systems with breakthrough economic value creation. Rapid advancements are bringing us closer to human-like machines that can automate physical tasks, improve efficiencies, and execute physical tasks humans can't.

02 Media

AI is going to change the media landscape as we know it. Personalized movies and video games, with content generated on the fly, will radically alter the value proposition and economics of content "publishing". A basic movie story experienced from any character's perspective, in any style, and for any length of time. A video game with endless variation and storylines. A content series where the user is featured personally in the story. The dynamic possibilities of AI-driven media are endless.



> "It's possible that over the next five years, we figure out a way to take large models and reduce them to make smaller AI models work together in networks.

03 Vertical Software-as-a-Service (SaaS)

AI-powered tools will empower any enterprise or startup to dynamically, quickly, and affordably create their own workflows and software applications, replacing the need for expensive, off-the-shelf SaaS solutions, with per-seat license fees. During a recent hackathon at my company, we used AI tools to build applications and workflows that replaced existing SaaS subscriptions. It took a few hours. This kind of in-house development allows for personalized solutions, reduces costs, and offers the ability to make continuous updates —all thanks to AI-driven software development.

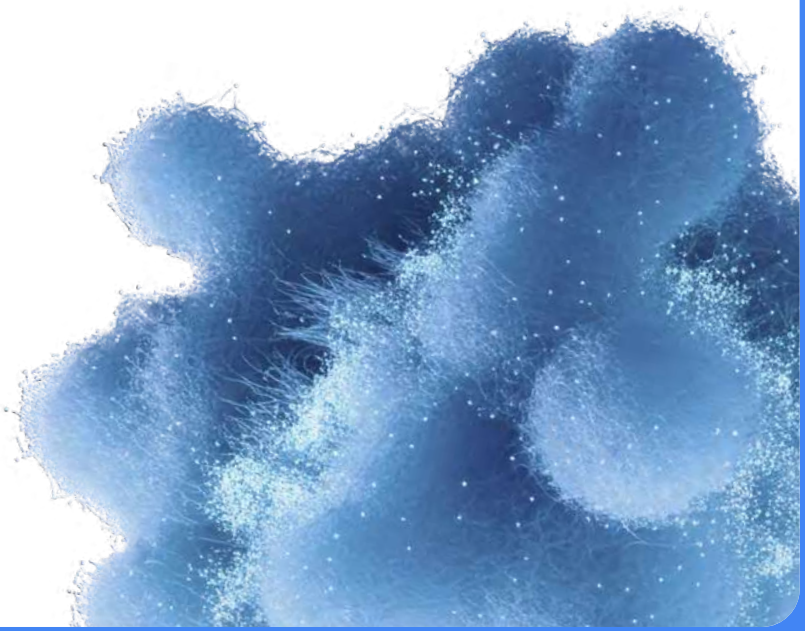
04 Biology

AI can now be used to design changes to the genome of any single-cell organism or plant, enabling the creation of novel biologic drugs and new crops that can produce more nutritious food using fewer natural resources. Thanks to low-cost DNA sequencing, we are seeing the utilization of "genome language models" (GLMs) that allow us to define a biological "goal"—an antibody drug that binds to a specific cancer protein or a corn plant that grows in a specific climate—and the software can render the exact DNA sequence(s) needed to enable that outcome. Using CRISPR gene-edited tools, those recommendations can be quickly developed and deployed. Biology has become software and AI is unlocking a new era of possibility.

An LLM wrapper is not a business

It's also not just about the technology itself; it's about how you use it. Startups need to be agile and adaptable, but also deliver unique value. If you're just an LLM wrapper, it's going to be hard to build a sustainable business—you're likely going to get commoditized away. Businesses need an engine of value creation that persists an initial advantaged offering with continuous improvements. This will typically come from proprietary data generation, which is used to continuously improve model performance, or network effects that lock-in access to data or customers.

Another key competitive advantage will arise from infrastructure costs. It is not the case that bigger models are always better. It's possible that over the next few years, large models will be reduced to smaller models that work together in networks—and they may be able to do things that larger models can't. This will reduce run-time costs, improve speed and other performance metrics, and unlock a new era in performance. I liken this to how genes are expressed and regulated in biology.



Douwe Kiela

CEO, Contextual AI

Contextual AI focuses on production-grade RAG agents that are highly accurate, auditable, and specialized for your business.

Douwe was previously the Head of Research at Hugging Face and before that a Research Scientist at Facebook AI Research. He received his PhD from the University of Cambridge.

Contextualize AI models for RAG 2.0

As AI models mature and use cases evolve, devising efficient ways to use rich context to get the right answer to complex questions becomes critical

When it comes to AI, there's a tendency to think that we are much later in the game than we are. In our company, we've learned how incredibly immature the infrastructure layer of AI technology is—things you would expect to work don't work all that well at scale.

At the same time, I feel like a lot of companies are aiming too low with their use cases, focusing on basic assistant or co-pilot apps. They're not thinking about the workflows and processes that you can automate with this technology if you aim slightly higher.

Actively incorporate structured and unstructured data in retrievals

Agentic retrieval-augmented generation (RAG) is a very interesting long-term trend for more active and accurate retrieval. AI agents try to figure out what information is needed to answer a given question, retrieve it from wherever it is stored, and even ask follow-up questions before giving a final answer.

If you can do that then you can potentially have different components like text-to-SQL models, code snippets, calculators, and structured database queries to reason on top of it to get to the right answer to the most complex question.

Right now, there's a data dichotomy separating structured data that lives in a database and unstructured data that lives somewhere in a data bucket. To do traditional RAG on top of it, you have to chunk it and put it into a vector database and then do some additional manipulation. But that dichotomy is slowly disappearing, so I think the really exciting use cases are at the intersection of structured and unstructured data.

As technology matures, I expect to see multi-agent systems combine the best of human and AI performance to solve complicated problems that require domain expertise and cultural understanding. This is where the whole field is headed.

Diversify data sources and apply hierarchy as best practice

Next generation RAG models are more tightly integrated for less hallucination, greater accuracy, and enterprise-grade performance. The idea is quite simple—train the generator and retriever together rather than as separate models where they're not really aware of each other.

Data ingested into the RAG pipeline always has an implicit hierarchy, and to maximize accuracy, you need to account for that hierarchy. This is particularly important as companies incorporate external data sources.

We see lots of customers who don't want to rely on any external information, perhaps because they have rich internal knowledge bases that they don't want to contaminate.

Yet, within these companies there are often multiple internal data sources with some kind of hierarchy that prioritizes, for example, research over email or text messages. They can apply the same logic to external sources.

It gets very interesting when there are conflicts, for example between something found on the web versus internal research. These conflicts can be handled by assigning weight to different sources or summarizing disagreements. In our own pipeline, we have an active retrieval strategy that figures out what retrieval is needed and how to weigh it.

Ground and contextualize models for most accurate and detailed results

Poor or incomplete training makes language models prone to hallucination. To avoid this, we start with a base model that may not be as grounded as we would like. We then pre-train it to be strongly grounded in the retrievals that come from our retrieval pipeline.

A well-integrated system is very good at contextualizing the language model, and the language model is strongly grounded in what contextualized it. Finding ways to enhance accuracy and detail across results will benefit companies going forward.

Aim for the quality of long context at the cost of RAG

There is some debate about whether to use RAG or long context, but if you know what you're doing you probably do both. There is a trade-off between cost and quality. If stuffing information in the context were free and accurate, then you could put everything in context, but that's not how it works. Instead, it scales with context length, so you need to use it in a smart way.

The human brain doesn't store much in working memory, instead fetching only what is needed in the moment. Similarly, in these AI systems you don't want to have too much context in working memory because that requires a lot of compute. Instead, retrieve just the relevant information needed to have the right context to do your job. If you want to know the name of the headmaster in the Harry Potter series, you don't have to read all seven books to get the answer.

The trick is to minimize the impact to compute while obtaining accurate results. Context scales quadratically in the number of tokens in your context. That's expensive. With long context models we solve that problem by making the attention mechanism very sparse. Basically, we completely ignore most of the tokens in most of the attention heads.

Take that sparsity to the limit and you zero out most of the information and you are left with contiguous blocks that you want to pay attention to. Those end up in your final answer given over to the language model. In this way, long context and RAG are the same thing.

> “Use cases that seem slightly out of reach are the ones you should focus on. You get a head start if you assume that the technology is ready, even if it isn't quite yet.

Dylan Fox

Founder and CEO, AssemblyAI

AssemblyAI is a leading AI company that has raised $115M from Accel, Insight, Smith Point and Y Combinator to build AI systems that transform human speech into meaningful outcomes and product experiences.

Dylan has a background as a Research Engineer and lives in Brooklyn, New York.

Bridge the gap of last mile issues

Differentiate your startup by addressing last mile issues and using nuanced benchmarks for AI performance

I'm excited about how real-time, AI-driven agents will evolve. We'll see multimodal agents that will include visual and audio in their interactions. For example, rather than calling a plumber for help with a leaky pipe, you'll open an app, show where the pipe is leaking, and the agent will talk you through fixing it in real time.

The timeline for widespread agent adoption and of AI in general will be longer than people think. I say this for a few reasons:

01 We clearly have to address a ton of last mile issues that aren't obvious until you're deep into production. The user experience with AI still has some clunkiness that'll need to be smoothed out—things like interruptions and delays in responses that make you realize you're not talking to a human.

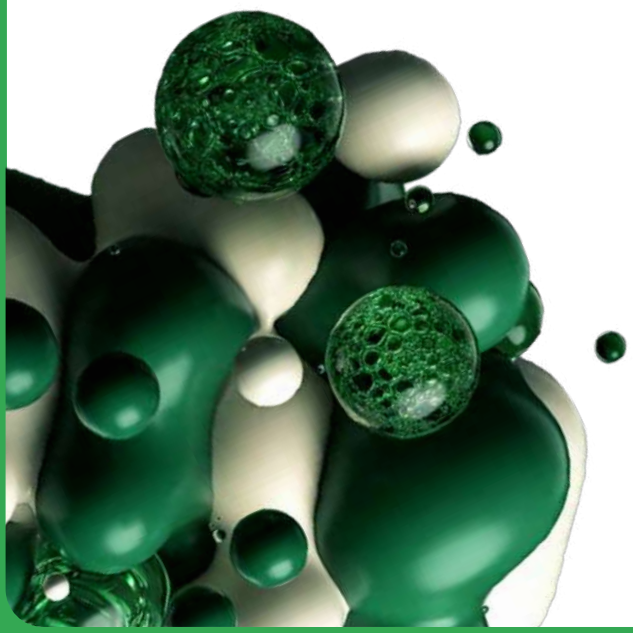
02 I also think that companies using and implementing AI tech will need to change how they evaluate it to get the results they expect. The silver lining is that these same last mile issues and difficulties evaluating AI tech offer startups the chance to differentiate themselves.

Solving last mile issues

Last mile issues with AI are things like its ability to augment models with proprietary or domain-specific knowledge, reliably follow instructions, or detect, remove, or totally avoid hallucinations. I've found that these issues are often unique to a specific industry, use case, or even customer. Each issue also has what I call a "fuzzy threshold" of how much it needs to be solved before it's no longer a barrier to AI adoption.

In speech AI tech, and specifically with real-time voice agents, one of the biggest last mile issues we encounter is the agent's ability to distinguish the voice of the person it's interacting with from background voices from a TV or other people in the room. Another is its ability to recognize rare words like a person's name, a place, or an industry or customer-specific term.

If you're an AI startup, you can leverage these last mile issues to your advantage. It's easy to throw AI in a lot of places in your products. But to build a differentiated experience, I think you really have to go into it with a clear perspective on the use case you're trying to solve for and then dig in to deeply understand what people actually care about for that use case. That'll show you what last mile issues you need to bridge to make your product the best it can be.



> "Solve last mile issues in your application and you can capture that market.

Evaluating AI with the right benchmarks

The second major barrier to AI adoption is with evaluating AI technology. I see companies making huge investments in AI based on metrics that are too generic. For companies to feel confident that the AI technology they deploy into production will do what they expect, they need to assess it based on more detailed and relevant metrics.

For example, with speech-to-text AI, the primary benchmarking metric is word error rate. Word error rate compares a human transcript to a machine transcript, removing all formatting for an apples-to-apples comparison. Yet we've found that for many businesses or use cases, how punctuation shows up in applications or how acronyms are formatted is really important. In fact, for end users and application developers, recognizing things like consecutive digits, email addresses, or rare words may be just about the only things they care about.

So I'd advise AI startups to help companies evaluate your AI products based on benchmark metrics that really indicate success for their specific use. To uncover those metrics, you'll need to understand how the company will use the AI and the workflows they'll use it for. The closer your benchmarks relate to customer expectations—which are often tied to last mile issues—the easier it is to know when your AI product has crossed that fuzzy threshold for success.

It's an exciting time for those of us working with AI, and though we need to get past some barriers before there's widespread AI adoption, it will happen. I think there's a lot of opportunity for startups—you solve those last mile issues in your application and you can capture that market.

Edo Liberty

Founder and CEO, Pinecone

Pinecone is the leading vector database for building accurate and performant AI applications, at scale, in production.

Edo was previously a Director of Research at AWS, Head of Amazon AI Labs, and a Senior Director of Research at Yahoo.

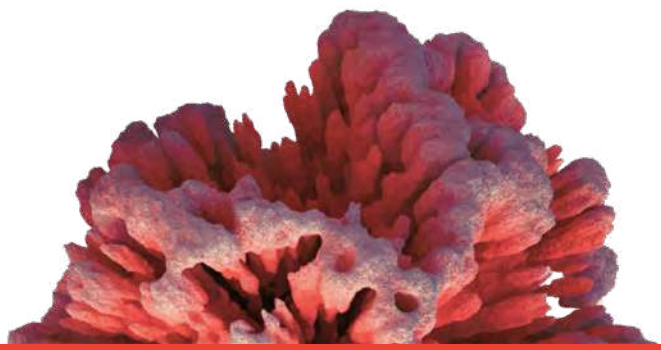
A massive AI opportunity lies in harnessing enterprise data

While the industry argues on AI's next commercial breakthrough, I believe one answer is hiding in plain sight

We hear a lot of folks in this industry talk about RAG (Retrieval-Augmented Generation) and agents, but I actually think one of the biggest opportunities in AI is to be found in search. Eighty percent of the world's data is unstructured; having an understanding of that data and the ability to efficiently retrieve it at scale can provide a ton of business value to companies with vast amounts of proprietary information.

Vector databases are essential for accurate, up-to-date information in AI applications. They store that information as embeddings, which help provide a semantic-based understanding of data that can be applied to any workload.

A lot of companies understand the value of lexical or keyword-based search, which is great for understanding documents and text. But there's a much broader opportunity from understanding data at a semantic level—whether that's through images, videos, PDF files, or tables—to leverage a centralized sense of understanding towards a myriad of applications, including search.



Of course, step one in all things data management and AI-related is keeping data secure—and then broadly how you think about scaling it, making sure it's accurate, fresh, and grounded. Many may not realize that making your product production-ready—which includes things like GDPR readiness, SOC2 Type II certification, HIPAA compliance, SSO, RBAC, CMEK, encryption at rest and in transit, etc.—can actually take up more engineering time than anything else and therefore should not be taken for granted.

Accuracy, groundedness, and low latency are baseline in 2025

Accuracy is critical. A lot of LLMs today get stuck in that 60% to 80% accuracy range. It really comes down to "garbage in, garbage out". How do you give them access to the largest corpus of data for them to be able to provide grounded generation in real time?

A vector database makes AI applications knowledgeable by understanding and retrieving the most relevant information, which, in turn, provides grounded, accurate results. Importantly, making generative AI applications knowledgeable also reduces hallucinations. The most significant use cases of the present and future that businesses can extract value out of—including AI search—are all optimally powered by vector databases built from the ground up for accuracy and performance at scale in production.

Beyond that, latency is another huge piece. A lot of developers struggle with the "sausage-making" that happens in AI. A database is just one piece of a much broader puzzle. How can we take some of those models and co-host them within our database? We need to make sure that the experience, whether it's to an end consumer or an employee, is as performant and highly accurate as possible.

> "I actually think one of the biggest opportunities in AI is in search.
>
> 80% of the world's data is unstructured; having an understanding of that data and the ability to efficiently retrieve it at scale can provide a ton of business value to companies with vast amounts of proprietary information.

Elad Gil

CEO,
Gil Capital

Elad is CEO of Gil Capital, an early investor and advisor to key technology companies including Airbnb, Anduril, Figma, Coinbase, Stripe, Square, and many more.

AI is massively underhyped

Despite all the attention, we are in the earliest days of AI's transformative impact on the world. For startups, the biggest opportunities might be hiding in plain sight

The thing that's underappreciated about AI is that the end product is a unit of cognition. Everybody is viewing it more as a software tool, but you're actually selling human-level capacity. And the current investment in GPUs makes it clear we are in the very early days of a massively transformative wave. There are a few areas that I believe have a lot of potential, and you can see them in different layers of the tech stack and in different types of companies.

If you look at the stack, there's a lot to do in terms of foundation models that are specialized beyond just language models and into areas like healthcare and physics. And then at the applications level, there is a ton of stuff that is interesting around converting services, software, and consumer products with AI—making common workflows more efficient and cost effective.

The key questions to ask as a startup are: What does your business actually do, where do your costs lie, and how will your competitive landscape be impacted by AI?

For physical industries (like shipping), AI might have minimal impact on their core operations, but will massively optimize the paperwork involved. Software companies with robust core offerings or broad, complex product suites may not see their fundamental functions disrupted by generative AI (although traditional ML may help), but there is still potential for it to optimize internal processes, workflows, and customer support.

There is a ladder of capability in AI where each iteration of a model brings a new level of possibilities, and those possibilities open up new use cases and markets. The greater portion of your cost structure tied to processes influenced by technology, the greater your opportunities for AI impact or transformation.

Here are a few areas I'm excited about:

01 Democratizing healthcare

I think one of the more interesting but underutilized areas of AI is in post-trained or fine-tuned models. Imagine a program with the expertise of a renowned specialist that could interpret images, your family history, and prior medical queries, and use all that context to diagnose and create a care plan for you on demand. While early efforts like Google's Med-PaLM and Med-PaLM 2 have explored this space, it's surprising that no one has fully productized a solution of this scope yet. And that's not to mention AI's power to dramatically reduce the cost of drug development through protein folding, as well as automated preclinical work, compliance checks, and clinical data for improved, faster outcomes.

02 Personalized education

AI can offer an unprecedented level of individual attention to students, coaching them through specific problems or learning gaps and giving them immediate access to best-in-class resources. Throughout history there have been huge societal benefits to having public libraries or public resources that are easily accessible knowledge bases. Generative AI unlocks that again.

03 Machine learning

Older areas of machine learning which are going to be enormously impactful are going under the radar. Self-driving vehicles are a great example of a technology that's finally hitting its stride and nobody's paying attention to it.

There's one last thing I'll add: In early technology shifts there is often a lot of low-hanging fruit. And oddly, founders don't go for it because they think it's too easy. But I think in new markets you can focus on the easy stuff, and those things work really well at scale and you can build the defensibility later. Find something people really care about, ship it as fast as you can, test whether people care or not, then iterate from there.

> "In new markets you can focus on the easy stuff, and those things work really well at scale. You can build the defensibility later.

Harrison Chase

CEO and Co-Founder, LangChain

Founded in early 2023, LangChain empowers developers to build context-aware reasoning applications with the LangChain open-source framework, LangSmith for LLM observability, evaluations, and prompt engineering, as well as LangGraph for agent orchestration.

Prior to starting LangChain, Harrison led the ML team at Robust Intelligence (an MLOps company focused on testing and validation of machine learning models), led the entity linking team at Kensho (a fintech startup), and studied statistics and computer science at Harvard.

AI problems require human solutions

Agents hold the key to a new level of productivity, but their success depends on our guidance

At its core, any AI agent needs to understand a problem like a human would, then replicate the solution in some form of code and prompts. So when you are building one, ask yourself two questions: Are you communicating to the agent what you want to do, and is it understanding what you're communicating?

I think evaluations are super important to make sure you're not regressing as you're changing your application. I also think it will be crucial to figure out how to evaluate not just the end result, but the trajectory to get there. Models, prompts, and RAG strategies change—how do you represent those different steps and know that your application is getting better with each one? How do I evaluate more complex agents?

How do I use those evals to improve applications programmatically while also remembering that it's just one part of a larger picture? More time spent here will force us to think better about what success actually looks like.

It's our responsibility to define and refine workflows, set evaluation metrics, provide feedback, course-correct, and mitigate the risks of models in their current state.

Here are five big areas I think will lead to big advancements in 2025 and beyond

01 UX unlocks everything

UX is a fascinating space for innovation in agents right now. To truly leverage agentic systems, I want them to scale beyond just what I can ask them to do —to be 'ambient agents', running in the background, always on, monitoring streams of events, and alerting me only when something interesting happens or when they need my help. But when you have agents running in the background, any human-level involvement requires complicated infrastructure-level programming. For example, it means the agent needs to run on a schedule; it needs to be able to alert a human, and pause until that human interacts with it (meaning you need to maintain the agent at a certain state indefinitely). That could involve code file changes, internal reasoning, browser state, documents, then to be able to save that state and resume it as soon as the human responds.

02 Human-in-the-loop becomes a priority

I'm not super bullish on fully autonomous agents. I believe the best agents will incorporate a significant human-in-the-loop component, with checks at the most insightful places to ensure they're not repetitive and that the agent learns from your feedback over time. This is the most important thing to get right when building agents because A) it gives you a nice controllable UX for interacting with these agents and B) it lets the system learn from those interactions. This approach enables complex engineering work behind the scenes while keeping us engaged when it matters most.

03 Vertical-specific beats general purpose

AI models struggle with complex reasoning tasks. They require very specific instructions and still exhibit unpredictable behaviour. Until models can reason more effectively, general-purpose agents will remain unreliable. Vertical, narrow-focused agents basically replace human workflows, and the best way to build them is to think about how a human would do something and then build a combination of code and prompts to replicate that process.

04 Context learning delivers over fine-tuning

Each agent we see built today is still bespoke—whether that's with a custom flow, guardrails, logic, or React-style loop. I think there will definitely be more model innovation, but I also think the interesting stuff will probably continue to be at the system layer. I think agents will continue to be the hot thing for the start of 2025. After that, I would guess it's something to do with memory. For example, if I have a bunch of interactions with a human, how do I remember those interactions and learn from them? I'm very bullish on using LLMs to reflect on those interactions and update their own instructions or profile of a user. As the models get better this might become more possible, but currently it's really costly and takes a lot of time.

05 Low-level frameworks hit their stride

I think there are two different styles of agent frameworks out there. There are high-level frameworks that have a built-in concept of tasks, where you don't really have control over the exact flow of how information is passed on. Then there are low-level agent frameworks that give you complete control over the exact state of the agent and what goes into the model at any point in time through things like prompts. I predict we'll see a lot more lower-level frameworks in 2025, because we learned that people need to control the internals of these systems in order to make them reliable enough to go into production.

> "I'm not super bullish on fully autonomous agents. I believe the best agents will incorporate a significant human-in-the-loop component, with checks at the most insightful places.

James Tromans

Managing Director, Web3, Google Cloud

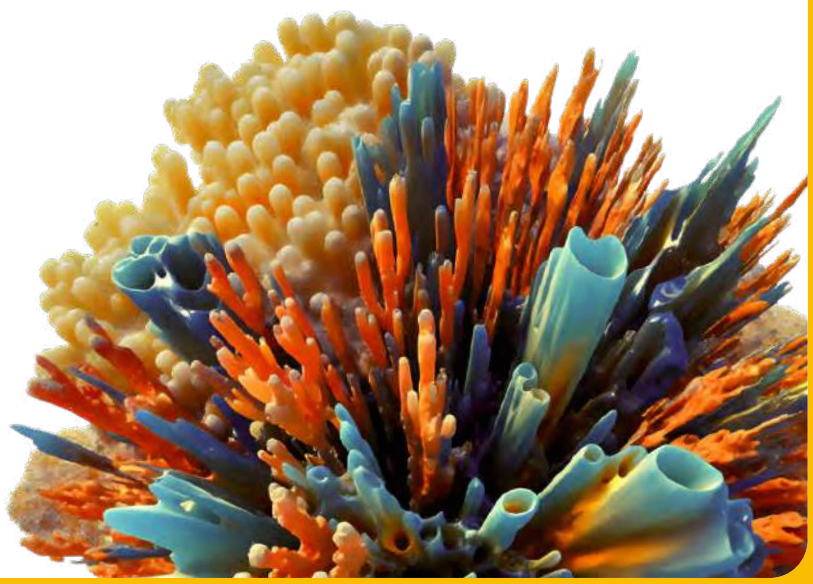
James is the Managing Director leading Google Cloud's Web3 efforts.

James previously worked as a technologist across different industries and holds a DPhil in the Computational Neuroscience of Vision from the University of Oxford.

AI and Web3: A perfect match?

Combining AI and Web3 offers opportunities to deliver unique value to consumers while building confidence in AI models and accelerating development creativity

AI agents using blockchain payment rails will become commonplace, despite formidable challenges ahead. Developer focus will center on AI agents acting as personalized assistants tailored for specific tasks like personalized investing, insurance, and mortgage management, leveraging Web3 payment rails and smart contracts. Other promising areas include developer enablement, composability, and model provenance.



Web3 decentralization: Accelerating AI agent use cases

A more compelling application of Web3 lies in empowering AI agents to conduct commerce effectively. We will see broadly useful innovations in this area. AI agents using blockchain payment rails is very much a current state capability; however, examples of agents leveraging traditional payment rails remain nascent.

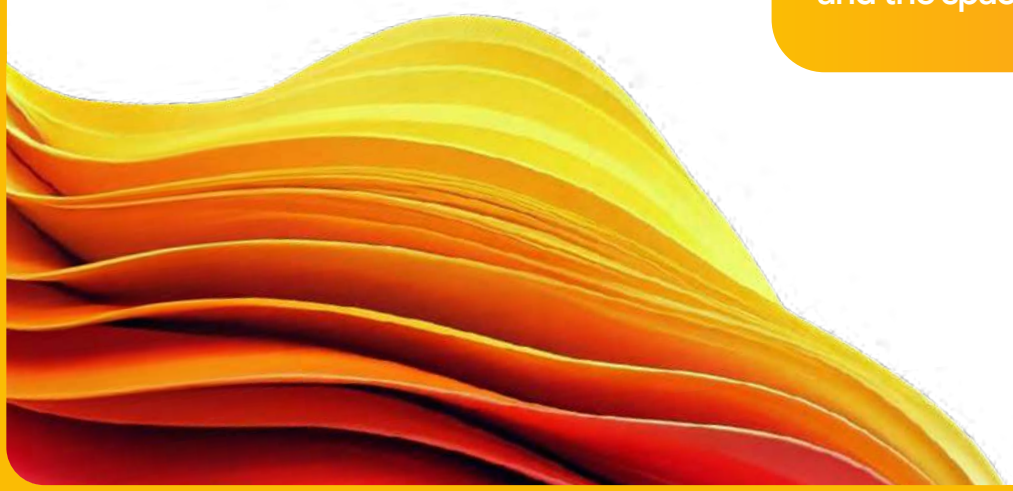
For example, an investor will talk to an agent to define a personal investment strategy, with the agent extrapolating and confirming things like risk profile, desired exposure, and volatility levels, at a cost that is significantly less than the 1.5% AUM typically associated with a Portfolio Manager. The agent will then interact with decentralized finance (DeFi) contracts, executing transactions on Web3 rails, using stablecoins. These digital assets, which have seen explosive non-cyclical growth in the past 24 months, offer a stable medium for micropayments and other transactions. We might even see the tokenized deposits or the digitization of commercial bank money being used here. Similarly, AI agents can evaluate and select insurance policies and set up payment schedules.

However, agents require access to accurate, verified user information. While verifiable credentials are gaining traction, a user-friendly, seamless implementation remains elusive—a significant opportunity for startups.

As technology advances, AI agent autonomy will increase. Agents will access LLMs with a small enough parameter count to run on personal devices, alongside users' verified credentials within a secure enclave. The agent can process this data locally, gaining user insights while employing zero-knowledge proofs to protect privacy. This technology enables the sharing of facts about information—for instance, proving someone is over 18 without revealing their full date of birth—rather than the information itself. This combination of AI and decentralized identity (part of Web3) is both achievable and highly useful.

Future applications include AI agents understanding and interpreting contracts and terms/conditions, soliciting bids, and entering reverse auctions, providing options to move forward with to save time, money, and administrative burden. The user reviews the top options and a pros/cons list generated by the agent. Creating this three-sided marketplace presents a significant challenge, but the potential rewards are substantial.

> "To act on behalf of an individual, the agent needs access to qualified, accurate information about that person. A gap still exists here, and the space is getting hotter.



AI: Enhancing Web3 development

LLM agents assist developers in writing secure smart contracts, mitigating risks such as re-entry attacks. This is particularly valuable in DeFi contracts handling escrowed funds.

More broadly, AI can detect errors, automate coding tasks, and promote best practices. Widespread AI adoption will lead to higher-quality code, increased developer productivity, and ultimately, more innovation.

Web3 composability: A catalyst for innovation

Web3's composability principle—inherent in blockchain technology—enables developers to create smart contracts as self-contained, often open-source modules of business logic. Their transparent nature allows anyone to understand and reuse them. At scale, this fosters a rich ecosystem where developers can easily combine and adapt existing contracts to create new applications with less coding effort. AI agents will further reduce development costs and complexity.

As we build out more advanced multi-agent approaches that collectively 'think' to solve a problem, the mapping of discrete smart contract functionality to agent speciality is a natural fit: an agent can go hunting for a smart contract that offers specific functionality, evaluate the options, and pick the best one. Doing this at scale to solve a complex multi-step problem is a strong fit.

Blockchain: Validating AI model training and testing data

While blockchain technology has been proposed for verifying model training data provenance, its large-scale adoption remains to be seen. The need for a robust system to track and validate training data is clear, but which specific user base will demand this capability and drive product-market fit is yet to be established. This remains an area of exploration and experimentation within the Web3 space. Furthermore, while closed-source models often lack real-time data and inference validation, some models deployed in the crypto context are already demonstrating this capability.

Jennifer Li

General Partner, a16z

Jennifer is a General Partner at Andreessen Horowitz, specializing in enterprise and infrastructure investments in data systems, developer tools, and AI, while serving on multiple boards, including ElevenLabs, Ideogram, and Pylon.

Previously, Jennifer held product leadership roles at Solvvy (acquired by Zoom) and AppDynamics (acquired by Cisco) and holds an MS in Software Engineering from Carnegie Mellon University and an MS in Technology Management from Rensselaer Polytechnic Institute.

The opportunities around AI agents

Investing in next-generative AI infrastructure

It will take longer than people expect for AI agents to truly go mainstream. Model reasoning capabilities will need to improve. And I've seen way more smaller size models being used in production than the most capable models, even though that's still the frontier that all the big labs are pushing for—so we'll see how the whole model landscape shakes out.

But I don't think the block is just on the model side—the infrastructure that will be leveraged by agents and deep-rooted systems integration problems will need to be solved as well. I'm also thinking about the data management tools AI agents will need, the real-time capabilities they will need, to better interact with users.

There needs to be a state management and memory management system—which could be in the agent itself or live outside of the models—to really orchestrate this workflow.

What humans are really good at is the last mile: thinking things through and making that insight-based big decision. Almost anything else—the low-hanging fruit—is, I would say, ripe for startups to improve.

Rapid evolution of the developer workflows

I'm particularly excited about AI applications that transform the developer experience.

Coding has turned out to be one of the most mature spaces in AI because it has more easily verifiable results and fast feedback loop. You can check if the results are correct, and there's already a lot of tooling, from testing to validation, to help figure out if the model has performed the way you want. It can be as simple as autocomplete or as involved as an agent building a full-blown app.

There are ways to integrate AI capabilities to improve developer productivity, to help developers build that next application or product—but developer behavioral habits will be harder to revamp. So if you really guide developers step by step to adopt AI, then you can bring about a 10x productivity gain.

I think we're already seeing adoption and ROI. And it's not just about coding. AI is also making it easier for developers to consume and learn new tools and technologies. Documentation is becoming more accessible and interactive, and developers can quickly get answers to questions without getting lost in a sea of text.

So I'm interested in companies that can seamlessly integrate AI capabilities into existing (such as IDEs) or new developer tools to demonstrably enhance productivity. And startups that are using AI to generate user interfaces and components, make it even easier for everyday users to build and deploy applications.

An incremental approach is more rewarding than 'everything AI'

Startups and founders seem to understand that AI is the future. But I'm finding an almost intrinsic fear of bringing in new capabilities, a hesitation to move too big or too fast when it comes to AI. For founders, I'd say really focus on where AI capabilities can move the needle and try to incrementally make changes and shifts. I think that approach has been more fruitful and rewarding than any sort of bigger "everything AI" shift.

The most important thing founders can do is be very close to the ground and very close to the new AI research and how existing behaviors are slowly shifting. That can help bring new products into formation, put products out there, and find out how the market and users will react.

Ultimately I think great teams are amazing product builders. Investors are leaning to great product tastes, and great engineering talent that can express their ideas into intuitive, sleek products . To understand model capability and marry it with what users want, that's the magic. So we're leaning towards high-velocity, fast-building product teams as much as we're leaning to research and machine learning capabilities.

> "The most important thing founders can do is be very close to the ground and very close to the new research and how existing behaviors are slowly shifting.

Jerry Chen

Partner, Greylock

Jerry is a seasoned technologist and investor specializing in enterprise software, cloud infrastructure, data products, and AI, with a focus on partnering with ambitious founders to create transformative businesses.

A Partner at Greylock since 2013, Jerry brings extensive experience from his leadership roles at VMware and Accel Partners, supporting companies from inception to scaling, and serves as a board director for multiple innovative companies.

Creative AI founders turn the tables on incumbent players

AI startups have the opportunity to reset the rules for how they build, sell, and monetize applications and how buyers in turn fund and monetize them

Your competition is no longer against the incumbent, it is against the incumbent's business model. I'm sure that the final business models for AI are not yet pioneered. We don't know the unit of value for AI companies—the founders I meet are still figuring it out. So that's a space to watch because it is important to how you build, sell, and monetize your product.

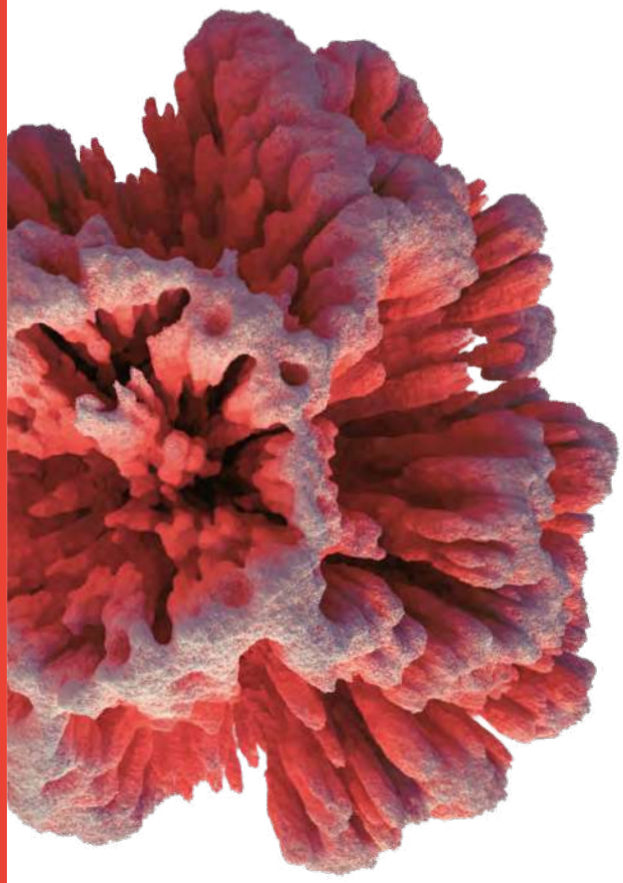
I think we will see a lot of new business models as AI permeates different verticals and lines of business, which is the fun part. It's been done before. SaaS and cloud companies changed the model from applications sold by seat to subscription or ad-driven sales.

Here are four things to think about

01 Look for budget replacement opportunities

When targeting operating expenses rather than the traditional software budget, it is incredibly interesting because instead of competing for funding against an incumbent software provider, you're competing based on efficiencies you can bring to a customer's internal operating models.

This approach is particularly effective for a new category of applications for human-AI interaction such as an AI call center, AI therapist, or AI auditor. These applications are more disruptive because you're creating something that doesn't yet exist—which we're excited about.



02 Serve highly personalized needs across verticals and workflows

We see AI serving a long tail of specific needs in different verticals in companies big and small. Before businesses had to pick a generic solution and customize it for their business. With AI, you can cater to unique workflows more easily and at a lower cost.

We expect AI applications to leverage data on personal devices to understand individual interests, preferences, and access rights to company data such as system records and workflows. We already have a couple of companies in our portfolio that serve as a bridge to help large models access personal data. AI can then learn how each business behaves and use this personal data to build a business-specific application. That unlocks value and creates a defensible position by constantly improving workflows based on evolving personal data. We think startups have an advantage here against very large incumbents.

> “Your competition is no longer against the incumbent, it is against the incumbent's business model. I'm sure that the final business models for AI are not yet pioneered.

03 Target new buyer personas across the organization

We encourage startups to think about every function across the business and really understand their different personas. Look for signals that an AI agent or application is needed for a particular vertical or knowledge worker category like developer, accountant, and lawyer. Look for high wages relative to value created, and repetitive workflows or skills. If there is a mismatch, then market demand probably exists.

It's both interesting and scary because so much time has been spent understanding what the CIO or VP of engineering wants. Sometimes for certain SaaS companies, it's what the CFO or VP of Sales or HR wants. Now you have to think about every function in the business. Founders that have some fundamental insight into the pain point of a specific persona will have a competitive advantage.

In the end, we still have to find the customer and deliver value to them. Whoever nails the user experiences better and at velocity is going to win.

04 Experiment and codify your AI application stack

No one knows what the canonical AI application stack will be for models, embeddings, vector databases, security, agentic frameworks, and so on. I think we'll first see developers build AI apps as they do today and then quickly figure out how to take advantage of a new stack of tools to make their AI-powered applications work really well.

We see developers using open source and cloud to play around with new technologies and adopt piecemeal before buying en masse. We're seeing a lot of experimentation within our startups and ultimately developers vote with their feet. They go to the best tools, so we watch where they head.

We also see startups building synthetic data and other models for post-training evaluation to make sure models perform with great accuracy. There's so much that we don't know right now and until the technology matures, people will have an important role in evaluation.

Jia Li

Co-Founder, President and Chief AI Officer, LiveX AI

LiveX AI delivers AI agents for happy, loyal customers.

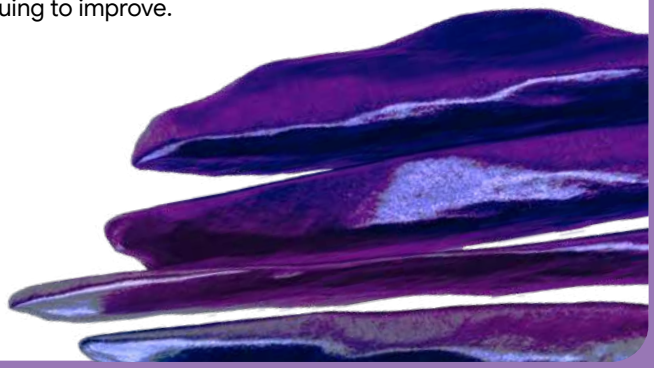
Jia has been elected as an IEEE Fellow for Leadership in Large Scale AI.

Human-like AI transforms customer experience

AI is evolving to drive more efficient, personalized, and human-like experiences

One of the most exciting developments I'm observing in AI today is the rise of AI agents that can simulate human-like interactions. As humans, we can observe, we can communicate, we can interact, we can take action, and we can show. Although right now it might be challenging for one model and one type of data to have all these effects, we're making progress—and human-like AI agents are the way of the future.

AI agents today have the power to understand directly from natural language, or even from multimodal data: if you show your product to AI and, for example, prompt it to "show me some problem with the product," it can immediately understand your prompt and complete your request. AI is becoming much more helpful—and much more personalized—because AI's human-like reasoning, understanding, and interaction is continuing to improve.



Reshaping customer experiences with AI agents

This shift towards human-like interaction has the potential to revolutionize the customer experience. I feel that this is the beauty and power of the generative AI-powered AI agent: it can understand the intent and emotion of individual customers, and then help those customers navigate based on their intent.

I believe that this progression towards more efficient, more personalized interactions exists because AI is not only about the size of the model, or different architectures, or compute power. It's also about the data. People often underestimate the power of data, but it's the key to unlocking the true potential of AI.

Many older, rule-based AI systems, what I'd call the "AI 1.0" experience, relied on pre-defined responses. Think about customer service experiences with a lot of yes or no questions, where we might press some numbers for the agent to then direct us along a predetermined route.

Newer AI agents, however, can understand natural language and even interpret multimodal data, like images or videos. This allows for more personalized customer experiences because the AI can recognize a person's problem, intent, and emotion, leading to more natural and helpful interactions.

Addressing trust and security challenges of AI agents

As AI agents integrate into our lives, it's crucial for us to address new challenges around trust, privacy, and security. It's important for us to think about security and privacy, to ask ourselves: how do we build trustworthy products? How do we handle all the privacy and security going into large language models?

In the past, AI might have solved simpler yes or no type questions. But now, large language models are trying to answer more complex questions, although they might not yet be able to provide us with clear or beneficial solutions. So more sophisticated AI can sometimes lead to hallucinations and concerns about compliance requirements around generative AI, posing new challenges about trust, privacy, and security.

One approach to trust and privacy that works, is to leverage models to handle the privacy protection —for example, to strip out personally identifiable information or sensitive data before the model even starts to learn. Also specifically trained models can handle prompt injection and deal appropriately with unsafe types of questions.

> "I feel that right now, people focus a lot on the size of the model, different model architectures, and compute power. One controversial point I always try to emphasize is the importance of data. People often underestimate the power of data and the impact of different data types on generating exceptional output for the models.

Improving accuracy, latency, and understanding to build human-like AI

We need to be mindful of how AI agents are trained and how they use data. It's important to learn how people are going to interact with AI agents through different types of modality and different types of behavior. Sometimes people might prefer that all the actions be taken by the AI agent, and sometimes people might like to have an agent that's more like a co-pilot or sidekick.

I believe the model architecture for AGI doesn't quite exist yet, and the data hasn't gotten there yet. Instead, we're trying to use individual specialized types of models and mini agents to learn specialized tasks, and we're pulling them together to have that holistic effect.

Real-time, human-like interactions won't be possible without improving accuracy and latency. An AI agent is a complex framework. Each step can take time and there are multiple steps stacked together. So now, that's why many AI agents are taking minutes, if not tens of minutes, depending on the complexity of the product.

Taking advantage of improved computation

Computation will be much more efficient and scalable. There will be new model architectures that can focus on solving specific problems with much more efficient structures.

We're getting there. For example, on one of my projects, we were able to speed up performance of average token generation on the structure of the models over six times. That becomes very powerful. If a task is going to take more than 10 seconds, then customers will get impatient, and they'll give up. So because of new speed around software, algorithms, and data, startups can now empower businesses to scale their services and serve every consumer.

Jill Greenberg Chase

Investment Partner, CapitalG

Jill leads AI investing for Alphabet independent growth fund, CapitalG. Investments include Magic.dev and /dev/agents.

A former CEO of a private equity-backed business and founder of a Y Combinator-backed startup, Jill is also a guest lecturer at the Stanford University Graduate School of Business since 2019.

Moving from a world of co-pilots to agents

I believe we're quickly entering an era where AI is intelligent enough to manage entire workflows. So how can businesses adapt?

While many in the AI space anticipate endless scaling leading to artificial general intelligence (AGI), I think foundation models are going to hit a plateau over the next few years. It's not that scaling is ineffective; it's the technical impracticalities that hold us back.

Imagine building a multi-hundred-trillion-parameter model. I think it would be an extremely compelling model that would probably achieve AGI. But practically, it's very hard to train that scale of model. We're talking about hitting barriers at every magnitude of scale—10 trillion tokens, 100 trillion tokens, and so on. Each fix requires retraining, making it a resource-intensive, time-consuming climb.

This limitation, however, presents an interesting opportunity for startups. While the big players wrestle with scaling, startups can focus on delivering tangible ROI to customers. This could be on the infrastructure side, making training more efficient, or on the application side, building solutions that don't require massive models.

How I'm investing in the age of AI agents

Looking ahead, I see two major trends emerging, and I plan to make investments in startups that are exploring these spaces

01 The rise of agentic workflows

I believe we're moving beyond AI co-pilots and into the era of autonomous agents. This shift is fueled by advancements in foundation models, particularly in reasoning capabilities. The budgetary implications for companies are huge, allowing them to save on not just software costs, but also labor.

Imagine AI agents that can fully handle customer service inquiries, not just assist human agents, or a pair programmer, to make developers dramatically more effective. This translates into a market worth billions of dollars. Startups have a key role to play in this agentic future. They can build full-stack agents for specific markets like customer support, healthcare, finance, and legal, focusing on the complex control and connectivity challenges. Or, they can create platforms that empower individuals to build their own agents, catering to personalized needs.

> "I genuinely believe that when all of this agent stuff works, it's going to allow human beings to spend so much more time on the things that are of high value.

02 Direct-to-consumer agents

This is where things get really exciting for consumers. Imagine an AI agent that manages your daily tasks. I could just say to my AI agent, "Hey, make sure I get to my meeting on time". Your agent checks your calendar, calculates travel time, and schedules an Uber or Lyft, all without you having to navigate to different apps.

This has the potential to revolutionize app development. If we're interacting with an AI agent, the value of an app shifts from its interface to its functionality. We might even see the emergence of AI super apps that act as a central hub for managing our entire day, further changing how we interact with technology.

Note that the emergence of direct-to-consumer agents also raises questions about data privacy and control. How comfortable are you with different apps accessing and sharing your information? This is where startups can step in, creating frameworks that allow users to define their preferences and control how their data is used by these interacting agents.

Building your startup's strategy in the agentic era

Investors are looking for startups that can demonstrate strong product-market fit, a distinct competitive advantage, and a clear path to profitability from the outset. These fundamentals remain crucial for navigating the evolving AI landscape.

I genuinely believe that when all of this agent stuff works, it's going to allow human beings to spend so much more time on the things that are of high value. Imagine a world where anyone can bring their ideas to life, and the only limiting factor is their imagination. That's the kind of future that excites me—and I'm looking forward to partnering with more incredible founders who are making this future possible.

Matthieu Rouif

Co-Founder and CEO, Photoroom

Photoroom is an AI-powered photo editing app that instantly resizes your images for any platform, removes and applies backgrounds, and edits hundreds of photos in seconds with one-click AI tools.

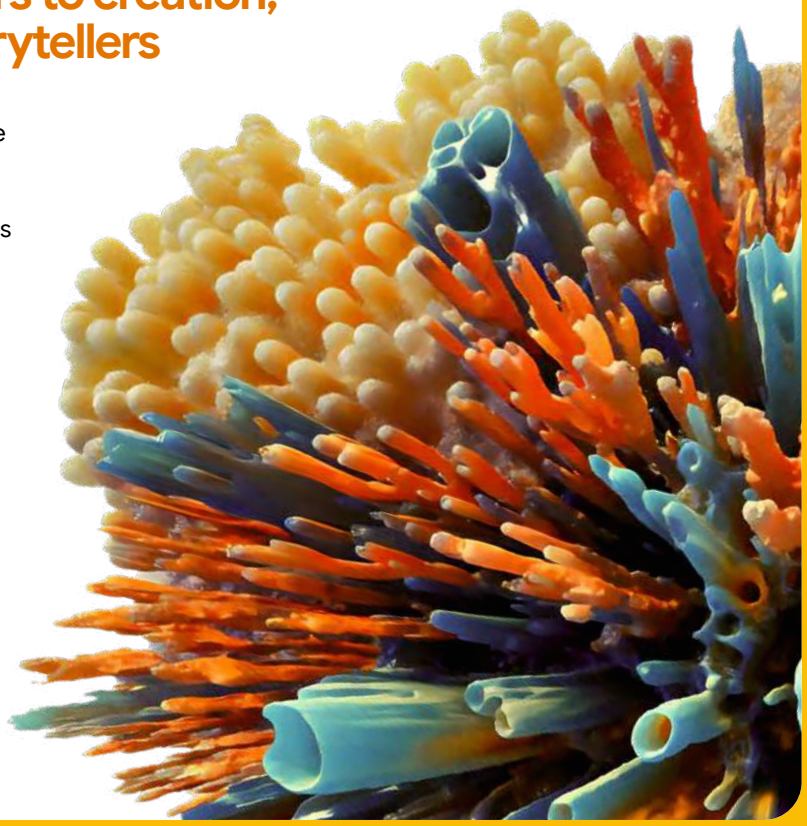
A graduate of Stanford, Matthieu previously led product management at Replay, a video editor, which was acquired by GoPro.

AI will free creators to create

AI agents are reducing barriers to creation, supporting creatives and storytellers

As someone who focuses daily on enabling people to use AI to edit photos, I'm seeing all types of people adopting and using AI today. Less tech-savvy people who couldn't use photo editing apps before can now create the images they want with AI.

While there might be a gap between experts, well-versed in AI prompts, and the rest of the population, AI use is already ubiquitous. I think it's our job as startups to continue making AI more accessible to all people.



Unlocking unique stories

One of the things I'm most excited about is how AI will allow many more unique stories to be shared with the world by anyone and from anywhere. For example, creatives traditionally needed publishers, producers, and promoters to get their songs, books, or other creations out to the world, and those barriers can be especially difficult to overcome when first starting out.

I see AI agents now reducing barriers to creation and supporting creatives and storytellers as they complete all the activities necessary to share their work. Think of a chef blogging about a new recipe. The creative work is the actual recipe, but the chef still has to write an excellent copy to accompany the recipe along with high-quality photos. AI can help with all these pieces that could delay the chef from publishing new recipes.

I believe that AI startups that focus on assisting creatives with the more tactical elements of improving their stories will enjoy increased demand for their work. Similarly, AI opens up a world of possibilities for AI tech startups—both as consumers and purveyors of AI technology. Just as AI agents can automate business tasks to help creatives get their stories out into the world, agents can automate marketing, finance, and customer support tasks to help businesses get their products out to the world. All this will free founders to focus on innovating.

> "AI will allow many more unique stories to be shared with the world by anyone and from anywhere.

Advice from one AI tech startup founder to another

I've learned a few lessons along my journey as an AI startup founder that others may be able to benefit from:

01 Develop guardrails for content

Content creation with AI is moving quickly, and millions of people use it for many different things. This is challenging traditional notions of authorship and even intellectual property. I recommend using common sense to outline how both your employees and customers of your AI tech can produce content with AI—and don't take any shortcuts. For example, we only train our models on licensed content purchased from creators to avoid legal and ethical issues.

02 Understand what's important for your AI tech users

Today, we mostly see generic AI that's not adapted to specific use cases. Think about it, if you're an office designer, you wouldn't create the same office design for a five-person tech startup as you would for a medical practice. Get something useful or usable into people's hands, interview them to understand what's missing or how to bring extra value, and adapt your AI accordingly.

03 Take an AI-native approach

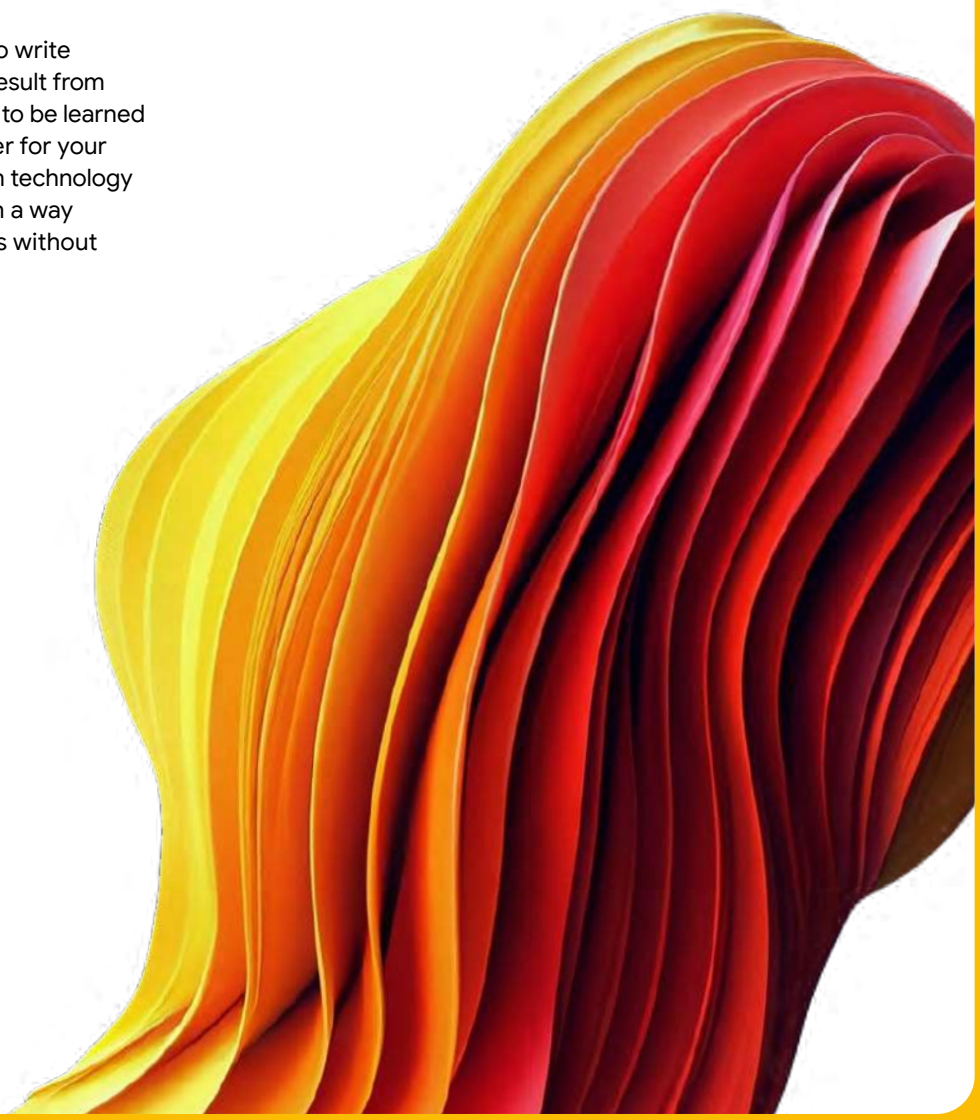
Build your company as an AI-first organization—not just with your AI products and services, but also with your company culture. When you define the big pillars of your company's culture, think about how AI will play a role in that culture. For example, at Photoroom, we adopted an open organizational culture to support remote work during COVID. We don't have communications via direct messages. Because nothing is siloed, it all feeds our company's AI model. Everything is accessible. Nobody has to ask the same question twice, and we get excellent summaries around various topics. It's a powerful asset that just gets better over time.

04 Design prompt-less AI

Assume that people have no clue how to write and use prompts to pump out the right result from your AI. AI prompting is a skill that needs to be learned and developed and shouldn't be a barrier for your potential user base. To get your AI-driven technology adopted, design your user experience in a way that helps people accomplish their goals without having to use prompts.

Inspiring new design trends

Looking ahead, I'm particularly interested in how AI will unleash new design trends. Consider the advent of font design—people wrote text by hand for hundreds of years. The invention of the printing press led to the start of font design. With the computer, the number of available font designs exploded. Similarly, I believe that AI technology will spur the beginnings of new design trends and art movements that we can't even begin to imagine today. With AI, the only real limits are our own creativity.



Mayada Gonimah

CTO and Co-Founder, Thread AI

Thread AI is a composable AI orchestration platform for organizations to design, implement, and manage their workflows.

As CTO, Mayada's focus and primary responsibilities include the development, architecture, and execution of Thread AI's technical strategy.

Most recently, Mayada served as a distributed systems engineer at Palantir, spearheading the development of a real-time AI inference platform. During her tenure at Palantir, she built intricate AI infrastructure systems, including the design of comprehensive model orchestration and evaluation frameworks, as well as complex containerization systems.

Workflow and infrastructure matter more than shiny tools

Focus on distributed systems fundamentals and build for evolving AI layers

We're seeing the sudden explosion of tooling that's trying to solve for a very small portion of the value chain. While it's great that everyone's excited about AI, and it's starting to solve some actual problems, AI models, half-baked solutions, and some pretty brittle tooling are being rushed to market when they haven't proven that they actually move the needle—whether with ROI or by fundamentally changing the developer ecosystem.

Durability, workflow orchestration, and observability as key enablers

There's this move back to fundamentals as companies are realizing, "Okay, I didn't have to blow my entire stack to embed an AI-based workflow."

My thesis is that AI should never be a workflow in and of itself—it should always be part of a workflow. So integrations and infrastructure for observable workflow management will become even more critical parts of the AI stack. Companies will need interfaces that let them insert AI in their existing workflows and test them in parallel to existing processes.

I think being more intentional around where you embed AI and having the tooling around observability is going to be key moving forward as we build these production-worthy AI-native workflows. Bringing back the focus on distributed systems fundamentals and understanding that just because something is probabilistic doesn't mean it's a completely new paradigm. You just need the right checks and balances for AI or you don't put it in the middle of a critical path where you can't handle any probabilistic outcomes.

> "Companies are realizing they don't need to blow their entire stack to embed an AI-based workflow. AI should never be a workflow in and of itself—it should always be part of a workflow.

Building for evolving infrastructure and databases

A lot of the layers of AI—the infrastructure and databases—are evolving. So as you build out your roadmap and your business as an AI startup, how do you think about those layers? How do you build that infrastructure layer knowing that these models are going to improve pretty rapidly over time?

I think you want to make sure that if a new version of a model comes along, you don't have to blast your infrastructure. You need a plug-in system where things are versioned, and maybe you integrate with models at the protocol level rather than with every single SDK. You add the bindings for gRPC, REST, or GraphQL, so you can just add that latest and greatest model. The same thing is happening with different database techniques.

Taking a composable approach to AI

I really believe that it will be important to develop AI models and infrastructure using a composable approach that takes the best use cases and techniques across AI and optimizes processes, like serving and combining different smaller models across different data modalities. So, I'd advise AI startups to take that approach as they develop their products and services. I'd also recommend investing in data readiness with data cleaning, curation, and hydration. That'll help you be best equipped to quickly adopt AI into your business processes and get the most out of the models you use.

The need for better explainability tools

I wish explainability tools were further along. The way a lot of these generative AI models are trained and later operationalized is completely severed. Once the model is trained, weighted, and later hosted for operationalizing, it's difficult to go back to some of the data. In some companies, models have been trained on data that they weren't supposed to train on, which introduced potential risk. It would be great to have a way to track some of the provenance or governance around the training. There's research into using ledger-like technologies, and I think that will all be important, especially as we're now seeing a lot of legal cases around data.

Raviraj Jain

Partner, Lightspeed

Raviraj is a Partner at Lightspeed Venture Partners, focusing on investments in artificial intelligence, frontier technologies, and broader space of b2b/enterprise software.

Raviraj serves on the boards of companies including Snorkel AI, Skild AI, Typeface AI, and Nirvana. Before joining Lightspeed in 2017, he was a product leader at LinkedIn and holds an MBA from Harvard Business School and a B.Tech from IIT Bombay.

Ready or not. Ramp up investment in AI.

Delaying AI adoption puts companies at risk of becoming irrelevant

I don't believe that AI is ready for large-scale implementation across the board - it's simply moving too fast.

Generative AI lends itself really well to creative use cases, but the core issues of hallucination and a lack of performance and security will require many new solutions over the next few years. On top of that, there's a lot of unstructured data, undefined processes, and end goals that are not well understood. We need specific tooling that will take a long time to emerge. And I'm not saying it won't be implemented, I'm just saying it's much harder than people give it credit.

But if companies choose to ignore it for now because it's moving too fast they will actually run the risk of being irrelevant in their core business—not just in terms of adoption of AI, but because it can fundamentally change the cost structure, output, and quality of different aspects of what matters for the company.

Here are a few things we care about when it comes to AI, and would encourage you to care about as well.

01 Become AI-native

Businesses underestimate the potential of AI, even if their products aren't AI-driven. To succeed, you need to prioritize becoming AI-native across all functions. You have to be truly AI-native to say, "Yeah, I understand what things are available today and where they are going so that I can build something that's durable". And there will be a lot of companies that'll build a product for customers based on where the technology is today and will become irrelevant two years down the line.

02 Don't wait for perfection

The progress of models is constantly accelerating. One can argue, "Why should I even try to implement AI today? Why not wait for two more years and let it get better?" I don't think that's the answer, because we are in the early stages of an evolution and things will keep evolving. You cannot afford to wait for these models to get better to start using them; you have to continuously invest and evolve with them yourself.

03 Embrace a venture capitalist mindset

In some sense, all organizations have to think like VCs (Venture Capitalists)—even though you're beholden to quarterly results, you still have to think five to ten years ahead and how your organization and market structure will change, and how critical it will be to leverage AI as part of that change. It's not about simply adding AI to your product, but your organization. The vector of differentiation lies in domain understanding and adoption, and this is a disproportionate opportunity for nimble, forward-leaning organizations to win.

The pace is accelerating

I'm often asked which industries will be particularly disrupted by AI, and frankly, the answer is all of them. But I prefer to say that what AI can do best is in sorting massive amounts of unstructured data, and that's applicable to any industry. From legal to insurance to healthcare, anywhere where unstructured data or voice-related interactions happen will see a lot of opportunities for automation.

I do believe that while the application layer will see far more activity, the foundation layer will see the greatest leaps in technological change. I'm excited by the potential of these models in the physical world, through things like robotics, as well as how they will upend our understanding of maths and physics. When you look at the volume of investment in AI, it's clear that we are going to get to a tipping point here sooner rather than later.

This work is not just about training the biggest model you can. It's about building the model, the enablement layer, and the product layer to solve a specific problem in a big way. There will be a few high-quality, massive models, but there will also be a lot of vertical models that can do specific jobs more effectively. How do you build the last mile for these models to connect to your specific problems, and figure out a really smart distribution strategy to solve people's problems? A lot of folks who haven't used software before in a real way will soon start using AI in a real way. Someone has to build the boat to help them sail it—whoever does that will build a very phenomenal company.

> "You cannot afford to wait for these models to get better to start using them; you have to continuously invest and evolve with them yourself.

Salim Teja

Partner, Radical Ventures, and Board Member, Aspect Biosystems, Promise Robotics, Intrepid Labs

Salim is a Partner with Radical Ventures, a leading venture capital firm focused on artificial intelligence.

With over 25 years in the technology industry, Salim is an active investor and leads the firm's Velocity Program, a specialized team helping founders to accelerate the growth of their AI-first ventures.

Bridging the gap: Turning AI into real-world solutions

The focus is shifting from what's possible to creating tech that actually helps solve today's biggest challenges

As an early investor in the AI space, I've seen remarkable progress. AI is already disrupting industries like sales, marketing, HR, and software development by automating tasks. This trend will only accelerate as companies recognize the substantial return on investment that AI tools offer.

I also foresee AI expanding into the physical world to move beyond just our screens and interact with the environment around us, revolutionizing areas like:

01 Robotics

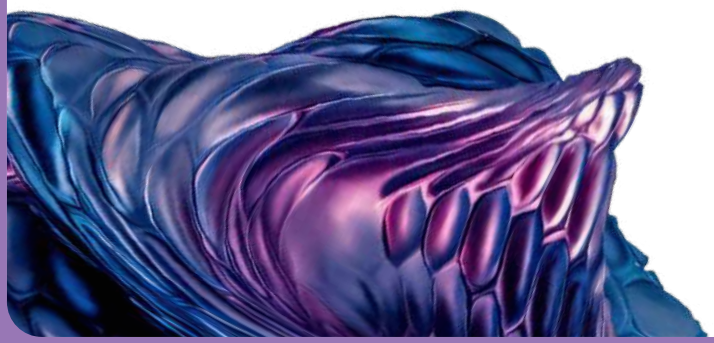
I see the convergence of AI and robotics as a major leap forward in how technology can be applied in the physical world, and I think it will have profound impacts. I believe robotics can be used in construction, for example, to make housing more accessible and affordable.

02 Materials discovery

I believe this can lead to innovations in various sectors, from manufacturing to construction. AI can accelerate the discovery of materials with specific properties, helping us develop new solutions to many pressing problems.

03 Spatial computing

This is an area of groundbreaking research right now with the development of 3D representation in the physical world and how to generate information from this in order to drive an LLM type of capability around spatial computing.



Striking a balance between research and application

There is some incredible groundbreaking research going on in AI right now, though significant challenges exist in translating research breakthroughs into practical solutions. Investors are looking for startups that have identified industry-specific problems solvable with AI and developed innovative solutions that deliver tangible value to customers. I'm paying close attention to:

01 Agentic operating systems

How will AI agents interact and collaborate within a larger system to revolutionize software development and technology? The agentic world is going to touch every single application in a pretty meaningful way as we design, deploy, and deliver on value propositions, and I think it's important for startups to build with interoperability in mind and be prepared to capitalize on the monetization opportunities it presents.

02 Mathematical reasoning

I believe it's important to keep advancing AI's mathematical reasoning capabilities. This will unlock new possibilities in scientific discovery and drive further innovation.

> "The true power of AI will be realized when it goes beyond data processing and starts interacting with us and our physical world in real and tangible ways.

Understanding AI developments and building your moat

I see too many startups that don't have product or technical moats—a competitive advantage that's hard to replicate. So as an investor, I start to worry about the durability of value propositions and how these startups can scale over time. And so I'm not saying startups should ignore research or that every startup needs to have Nobel Prize winners inventing new fields of AI research.

I think it's wise for startups to focus on research that relates to their technology area where they can find slivers of technology that can be helpful in creating competitive moats. This means having someone in-house who understands the AI technology landscape and can identify opportunities for differentiation. I think it's important to stay informed about technology developments in your field, so you can leverage new discoveries to your advantage.

Focusing investment on implementation

In general, we're shifting our focus from simply being fascinated by what's possible with AI to how we can implement it in industry at scale. In AI, it's easy to get caught up in the research and the art of what's possible, but we can sometimes lose sight of how hard the adoption journey is—it's not trivial.

My focus is looking at startups who can help deliver value, build real businesses, and use AI to address real problems such as improving health, fighting diseases, combating climate change, and addressing the affordable housing crisis through robotics in construction. These are all meaningful problems that will affect everyone whether they're in our industry or not.

Sarah Guo + Mike Vernal

Founder and Partner, Conviction

Partner, Conviction

Sarah Guo is the Founder and a Partner at Conviction, an AI-native venture capital firm founded in 2022. Prior, she was a General Partner at Greylock. She has been an early investor and partner to 50+ companies including Figma, Harvey, Sierra, HeyGen, Mistral, Cognition, Stackblitz and more. Sarah is from Wisconsin, has four degrees from University of Pennsylvania, and lives in the Bay Area with her husband and three kids. She co-hosts the AI podcast "No Priors" with Elad Gil.

Mike Vernal is a Partner at Conviction, a venture firm founded in 2022 to serve the next generation of founders. Previously, he was an investor at Sequoia, where he partnered with 20+ companies, including Rippling, Notion, Verkada, Clay, Statsig and more. He started his career in product and engineering roles at Microsoft and was then an early member of the Facebook team, eventually leading large parts of product and engineering. He studied Computer Science at Harvard where he met his wife and he spends his free time playing baseball with his three young kids.

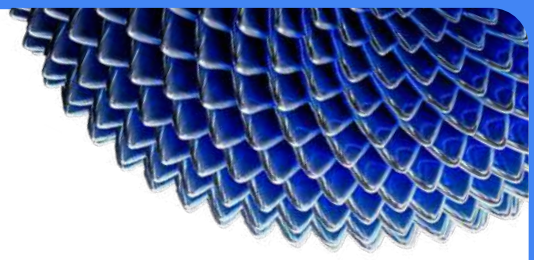
AI is leveling the playing field

AI is rewriting the rules of the tech industry, offering agile and creative startups a once-in-a-lifetime chance to compete with today's leaders

AI's potential for value creation is so large that it is accelerating economy-wide adoption of technology and changing the dimensions of competition in many industries.

We're seeing it now in areas like law, healthcare, pharma, and defense. These industries face a great deal of change. Their inputs are intelligence and knowledge work. There is a high volume of paper pushing. The opportunity for disruption, automation, and democratization is very large. Much of the accepted wisdom in venture capital is being challenged—from which markets are attractive to the business models and even scope of technology companies.

Increasingly, technology companies are taking on services industries or giving their customers more capability, rather than squeezing more efficiency out of an existing workflow. The changing basis of competition makes every industry more dynamic. It's a huge opportunity for startups. When the enabling layer of the technology stack evolves at an incredibly fast clip, as it currently is with foundation models, then in order to make investment decisions, one must take a point of view on what the 'governor' is—the limiter to development and adoption.



Some core tenets of our thinking:

01 Project capability sufficiently

To determine if a company is attractively positioned with AI, we suggest a simple test: when a new model is released, are you happy or sad? If you're sad, you're probably doing something wrong—if you are just building at the boundary of what the models can do today, there's a reasonable chance that the next models will subsume that. You should aim to design your product and in a way that eagerly anticipates and leverages new capabilities.

02 Follow the data

While popular, the assumption that large companies with vast datasets will automatically dominate the AI landscape will often prove false. The data held by incumbents may not be suitable for training AI models, and legal constraints might limit its use. The most useful data is often reasoning traces and evaluation methods for a specific, real-world use case. In many instances, this data has not yet been collected by any player, and the field is open for new entrants.

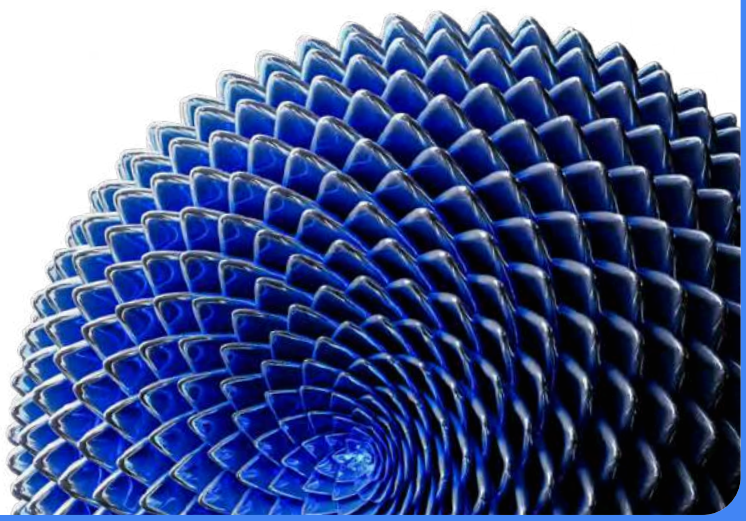
03 Prioritize first-principles thinking

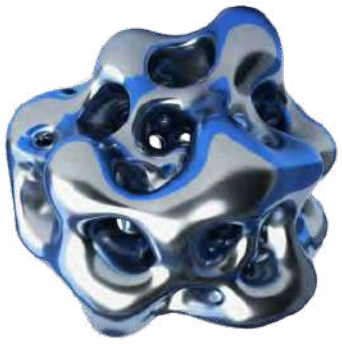
In a fluid environment like AI, there is a lot of discourse where perspectives change quickly. Instead of blindly following dominant narratives from investors or research labs, ground your decisions in fundamental understanding of technology and markets.

04 Great founders and products, forever

There is a bifurcation of potential investments we look for: there is one class which is just great products as evaluated by our taste and understanding. Then there are earlier-stage investments in people that move quickly, have good taste, understand customers, and iterate quickly to solve problems. That's been a recipe for success for a long time, and now with AI we just have a new set of tools for them to build products around.

We get particularly excited when somebody holds a combination of world class understanding of a particular domain, and is high-velocity with all the core entrepreneurial traits you want. The most important thing is to be able to iterate quickly, especially in a context that is evolving very quickly as well. You want people who are creative and experimental, who can figure out where the edges of technology are, learn them, and keep on pushing.





Domain-specific models unlock new opportunities

Everyone is (appropriately) fixated on the magic of foundation models right now. But we are at the beginning of a 10-to-20-year wave of applications, many of which we cannot yet predict.

We expect to see many domain specific models used over time, with a spectrum of approaches: pre-trained, post-trained, and fine-tuned, to a specific domain. Some domains, such biology, materials science, or robotics, will require breakthroughs in efficient data collection and generation.

> When the bottom layer of the technology stack evolves at an incredibly fast clip, as it currently is with foundation models, the middle layer ends up being a 'governor', limiting the speed of application development.

The 100-person, 100-billion dollar company

AI acts like a force multiplier, enhancing the capabilities of humans while still needing them to steer it. While we don't believe that entire functions in growing companies will yet be outsourced, we believe this force multiplier can fundamentally improve the quality of businesses. Many founders in our portfolio are aggressively leaning into leveraging AI. Our approach is to encourage people to be unreasonably ambitious about what small teams can achieve.

Yoav Shoham

Professor Emeritus of Computer Science, Stanford University, and Co-Founder, AI21 Labs

Yoav is a leading AI expert and founder of three previous AI companies, acquired by Ariba/SAP and Google (x2).

Yoav has won multiple academic awards and Fellow of the Association for the Advancement of Artificial Intelligence, Association for Computing Machinery and the Game Theory Society.

Stop prompting and praying, and start building complete AI systems

That is, if you care about controllability, reliability, and efficiency

Frankly, I hate the term AGI, or Artificial General Intelligence. It's not a thing. I've been around the block enough to know that intelligence is multifaceted. Machines will undoubtedly be able to automate more and more functions, but there's a false sense of there being a discrete point at which that mythical AGI will have been reached. I believe this loose thinking and hype around AGI is a distraction.

Instead, let's talk concretely about AI technology, its strengths and weaknesses.

Why businesses are experimenting madly with AI, but are cautious about deploying it at scale

You see this in the enterprise. While consumer adoption of AI has set a record pace, business has been slower to adopt it. Certainly, CEOs and boards everywhere are presenting their company as being "AI first" (or planning to become that), and are experimenting heavily, sometimes with hundreds of use cases.

But for all of the mass experimentation going on in enterprise, only a fraction of AI projects actually reach deployment. This boils down to two key challenges:

01 Cost

The high cost of running LLMs challenges the economics of business software.

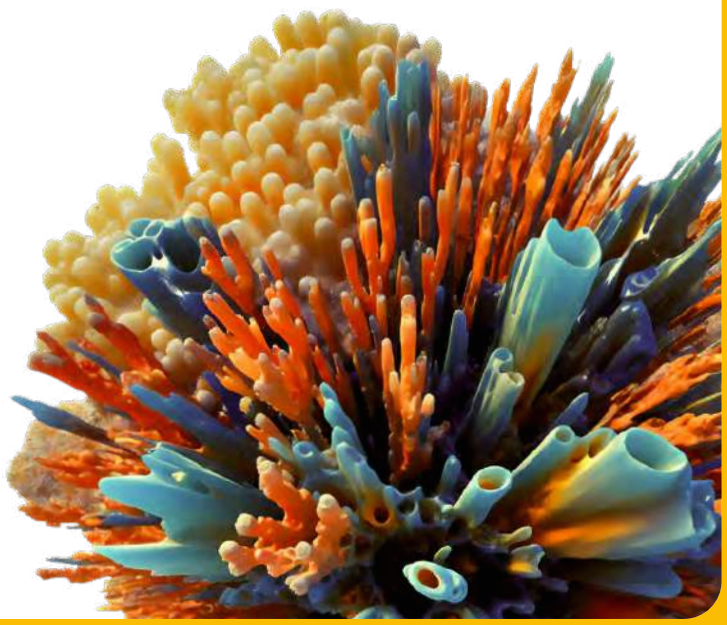
02 LLMs produce wonderful output alongside utter nonsense

Imagine writing an investment memo or responding to a customer, and your AI is brilliant 95% of the time but produces garbage the other 5%. That is a showstopper in the enterprise.

> "AI systems will vastly outperform basic LLMs in reliability and efficiency, but they won't be perfect. Smart entrepreneurs will build products that leverage these systems' strengths while accounting for their limitations.

The first challenge is due to the inherent cost of serving (let alone training) LLMs, and will be dealt with by a combination of two methods. One is using smaller LLMs (the term Small Language Models, or SMLs, is making the rounds), those "tiny" sub-7B and even sub-3B parameter models. The other method is using different, more efficient architectures than the standard transformer architecture; AI21's Jamba family of models is an example.

The second challenge is particularly acute, and more challenging. The phenomenon is inherent to the probabilistic nature of LLMs, and it's delusional wishful thinking that this will go away, no matter how much effort is placed on things like "alignment", "guardrails", and such. I believe that we're going to wean ourselves from what I call our current "prompt and pray" modus operandi. The industry will realize that LLMs are an element of a more comprehensive AI system, which can seamlessly integrate and leverage the strengths of various AI technologies, including LLMs, retrieval, tools, and other traditional code. AI systems will offer greater control, efficiency, reliability, especially when tackling tasks that require nontrivial reasoning, which most tasks do.



A healthy practice for startups: Use AI systems, but focus on "product algo" fit

AI systems that combine multiple LLMs and other tools offer a compelling solution. These systems allow for better cost and compute management by intelligently routing tasks to the most suitable resources. For example, a smaller LLM could act as a "router," directing tasks to specialized LLMs or tools for optimal efficiency. These systems can also enhance reliability and quality by incorporating checks and balances during the computation.

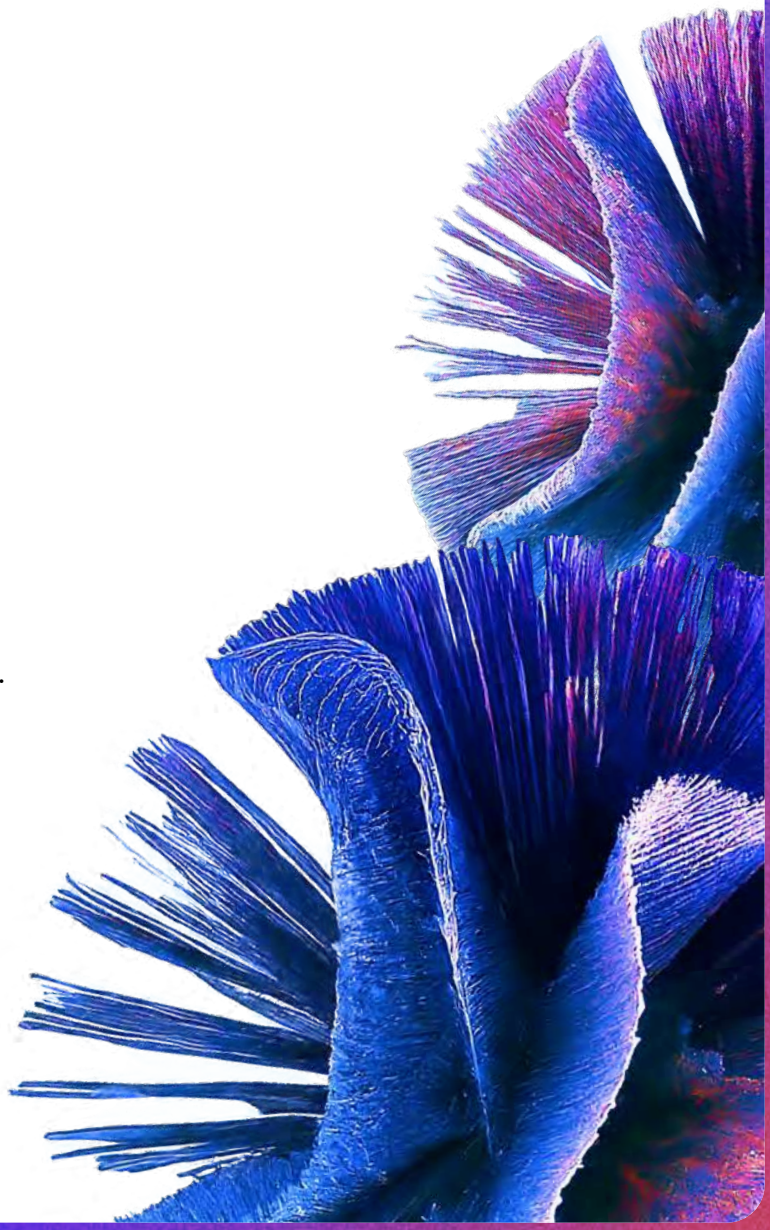
But as you apply these AI systems to real-world problems, you should approach this wisely. My common advice to AI startup leaders is to strive for "product-algo" fit. What I mean by that is while AI systems will be a dramatic improvement over barebones LLMs in terms of reliability and efficiency, they will still be imperfect; the underlying uncertainty involved in LLM calls, search and retrieval will not completely go away. So as an entrepreneur creating a new product, understand the strengths and weaknesses of the technology, and craft that product in a way that leverages their strengths and compensates for their imperfections. This is what I call "product-algo fit".

Build the future with Google Cloud

Today, more than 60% of all funded generative AI startups globally and 90% of generative AI unicorns globally are Google Cloud customers. We're proud to bring our long history of AI leadership and innovation to bear, delivering cutting-edge cloud AI solutions that support our customers' unique opportunities and challenges.

At Google Cloud, we're collaborating with researchers, founders, startups, enterprises, partners, and public sector agencies to think critically about how our responsible AI solutions can continuously meet the needs of employees, customers, patients, and citizens. This includes providing world-class infrastructure and full-stack capabilities at the forefront of innovation, engaging deeply with inventors on data, agents, and applications to help bring new outcomes to life, and partnering with enterprises to evaluate how AI advances can modernize experiences both inside and outside organizations. The breakneck pace of change being driven by generative AI means that startups are facing unprecedented challenges—and we're here to help.

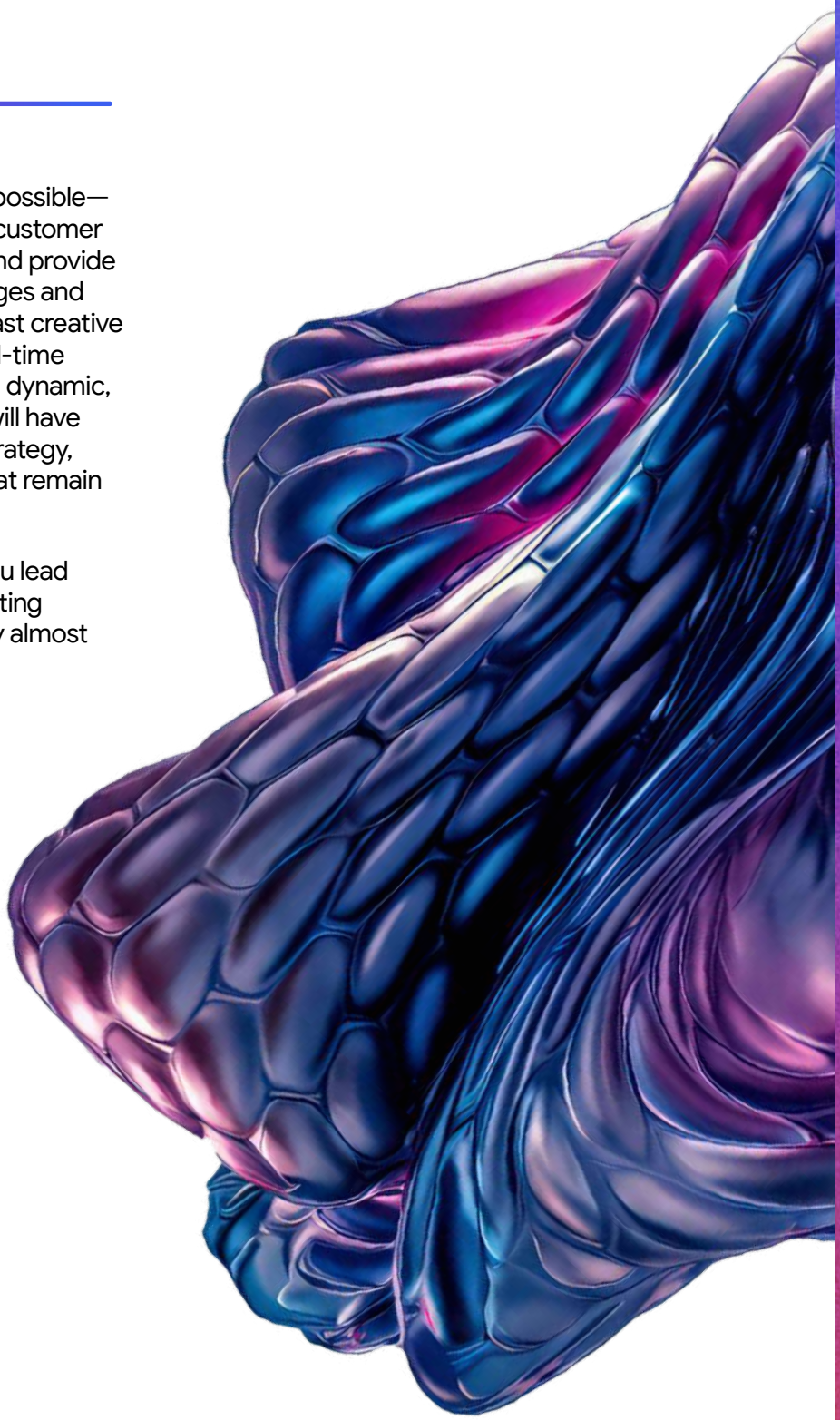
Darren Mowry
Managing Director, Global Startups,
Google Cloud



AI allows for a level of go-to-market personalization that was never before possible—the creation of intuitive, individualized customer journeys that analyze user behaviour and provide customized recommendations, messages and offers in real-time. The future will see fast creative production, tailored activation, and real-time measurement, all working together in a dynamic, interconnected system. And startups will have more time to focus on the creativity, strategy, and authentic customer connection that remain central to driving engagement.

We're excited to partner with you as you lead the generative AI charge in a rapidly shifting environment where what's true one day almost certainly isn't true the next.

Alison Wagonfeld
CMO, Google Cloud



Google Cloud

Accelerate your startup journey with Google Cloud.

Take the next step

Please contact us if you have any questions about the research or if you'd like to set up time to talk about how we can customize Google Cloud solutions to fit your needs.

Contact us