



A 5-Step Formula for Safe Generative AI

Dr. Seth Dobrin
CEO, Qantm AI





Generative AI has the potential to revolutionize every industry.

**What's
your plan?**

Others are starting already ...



AI teams are identifying use cases



Fitting GenAI into existing governance and accountability structures



Using exclusively commodity models



Leveraging existing human resources



Leveraging existing technical capabilities

But their approach isn't working ...



Needs to be tied to business goals



Unique challenges associated



Selection of models is use-case specific



Existing talent lacks the understanding

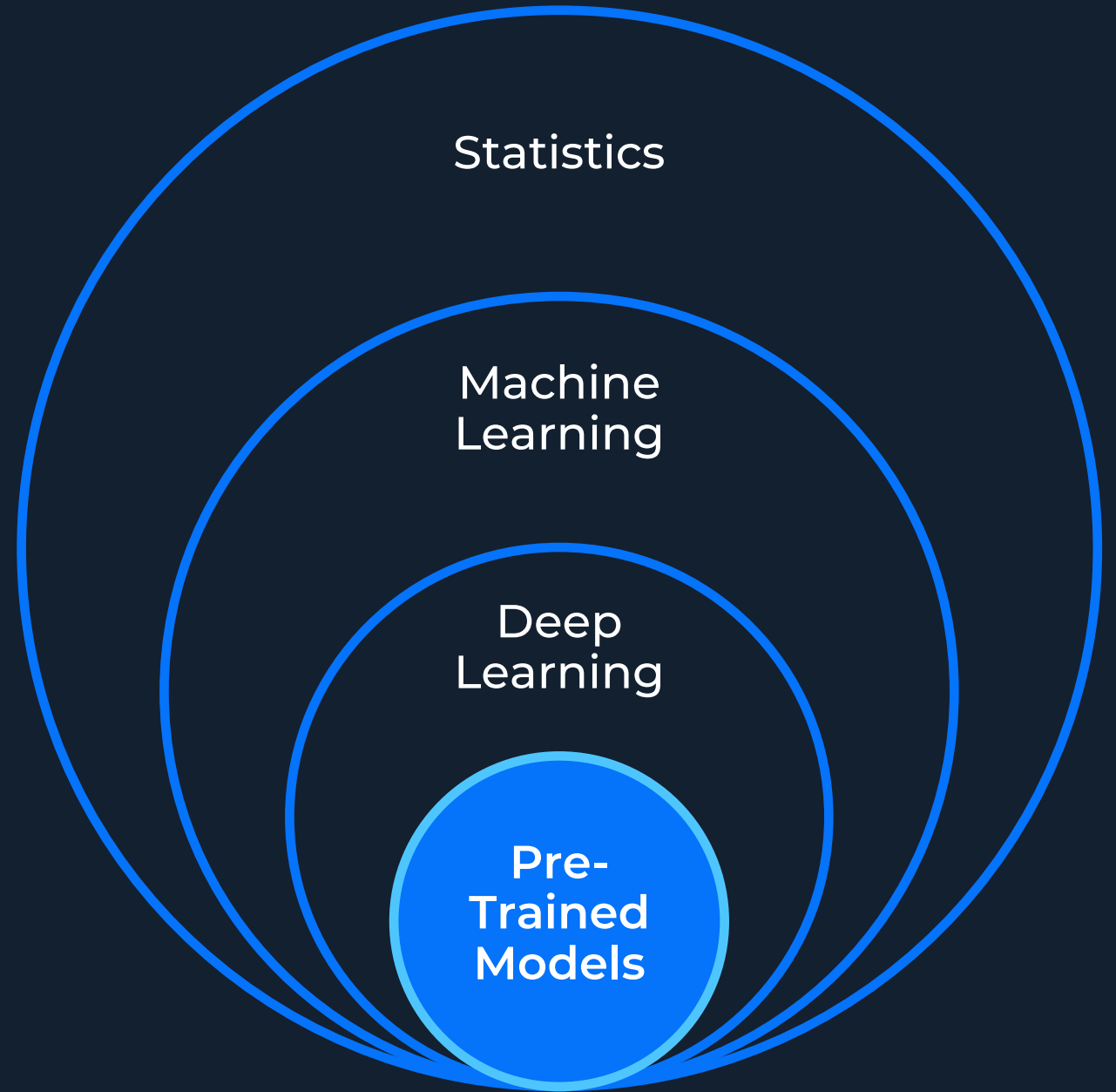


New architectures are needed

What
is AI
Anyway ...



Level-Set on AI



Are you using AI
in your
organization ...

After more than a decade, is AI finally reaching enterprise scale?

28%

Widely implemented,
driving critical business value and will remain a major component of our strategy

16%

A single use case
in one or a few departments but is driving critical value and will remain a major component of our strategy

27%

Implemented in a few use cases
across a few departments that will be key for us to scale

30%

A minor component
of a broader strategy but will be critical in the future



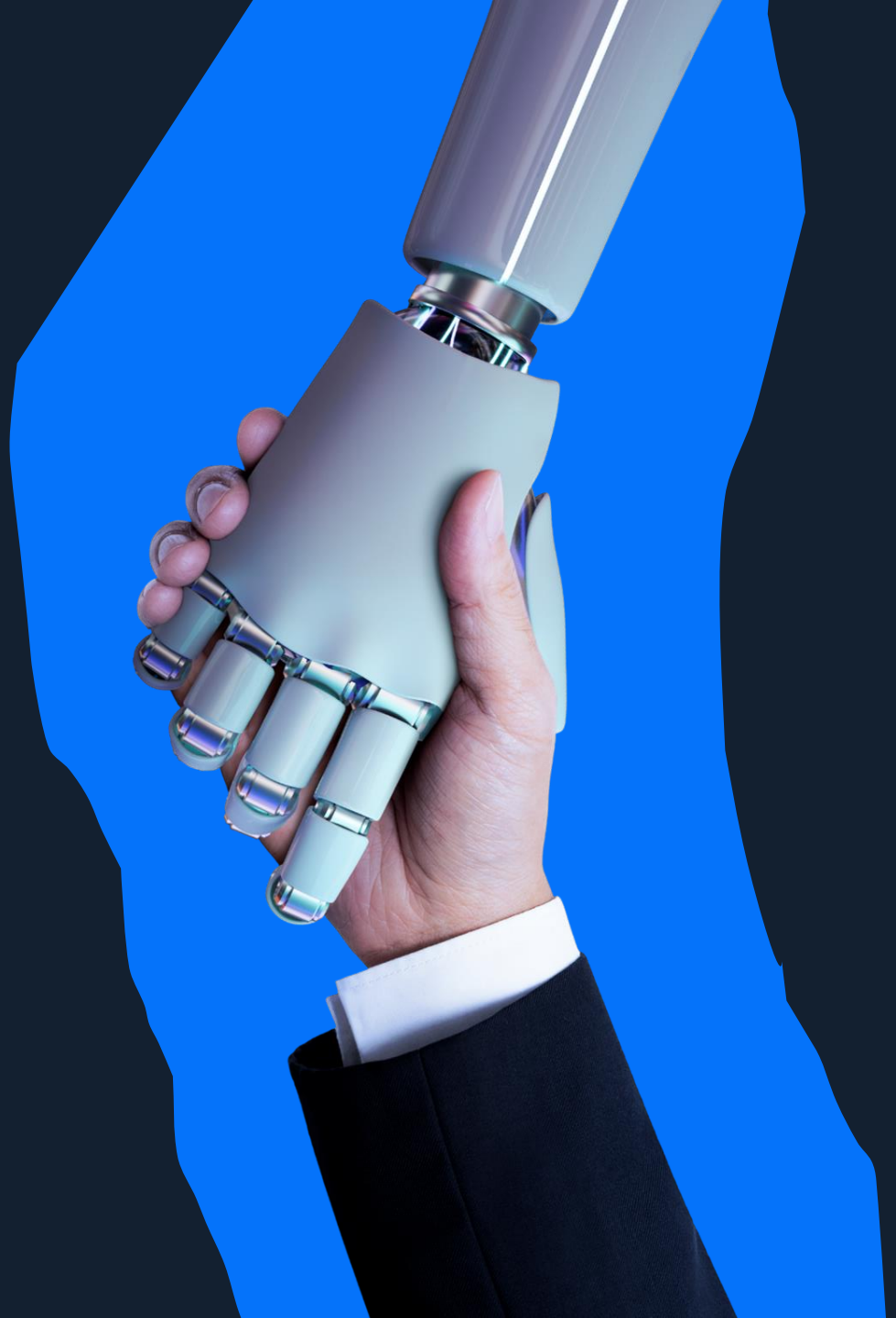
AI projects

69%

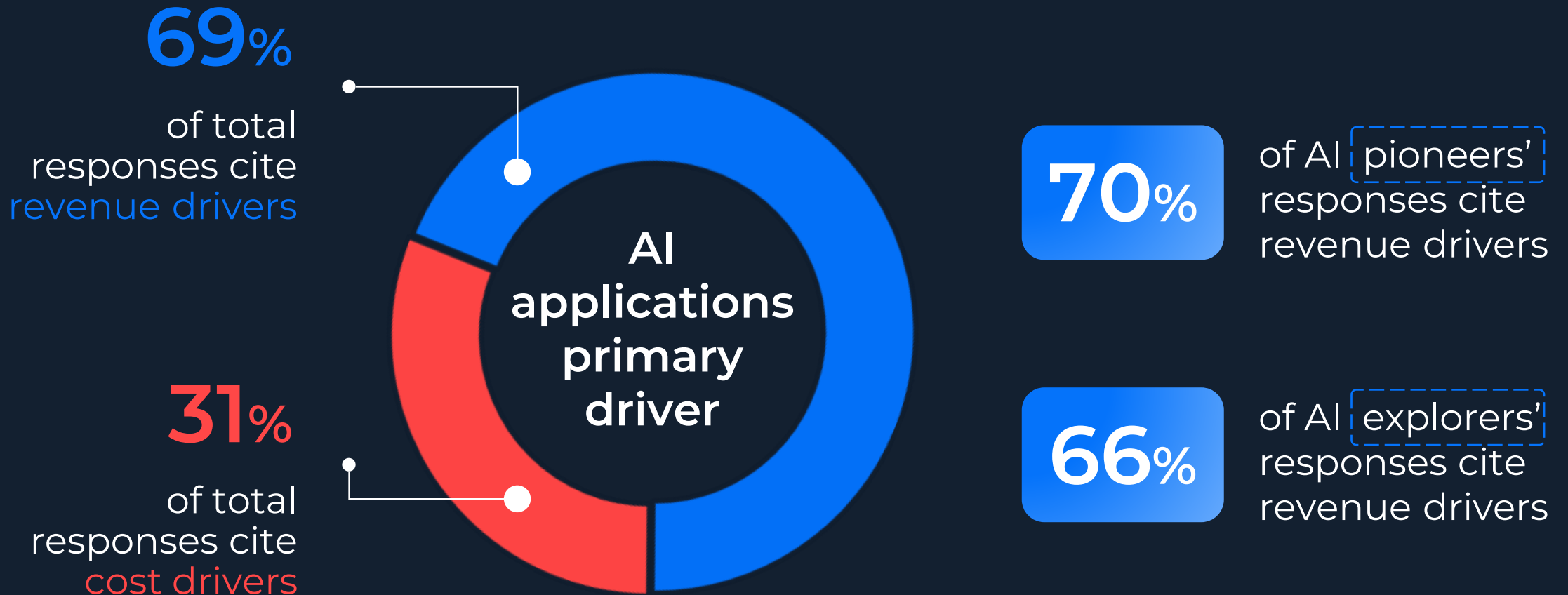
AI pioneers:
AI/ML is in production

31%

AI explorers:
AI/ML is in pilot/POC



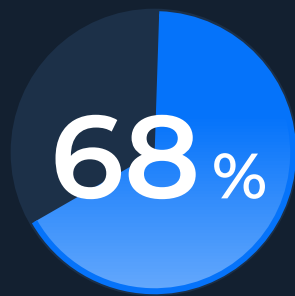
Revenue Drivers Lead Cost Drivers for AI/ML



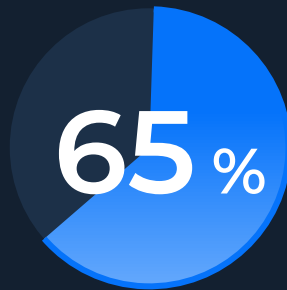
Do your
customers
know???

More than half of Americans are aware of common uses of AI ...

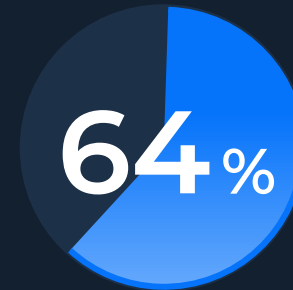
% of U.S. adults who identify that the following use AI in multiple choice questions



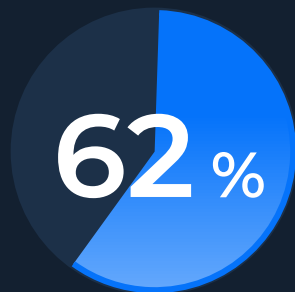
Fitness trackers



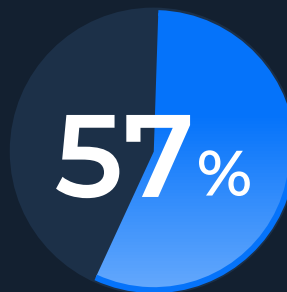
Chatbots



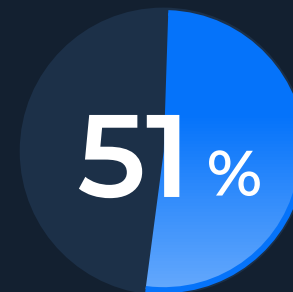
Product recommendations



Security cameras



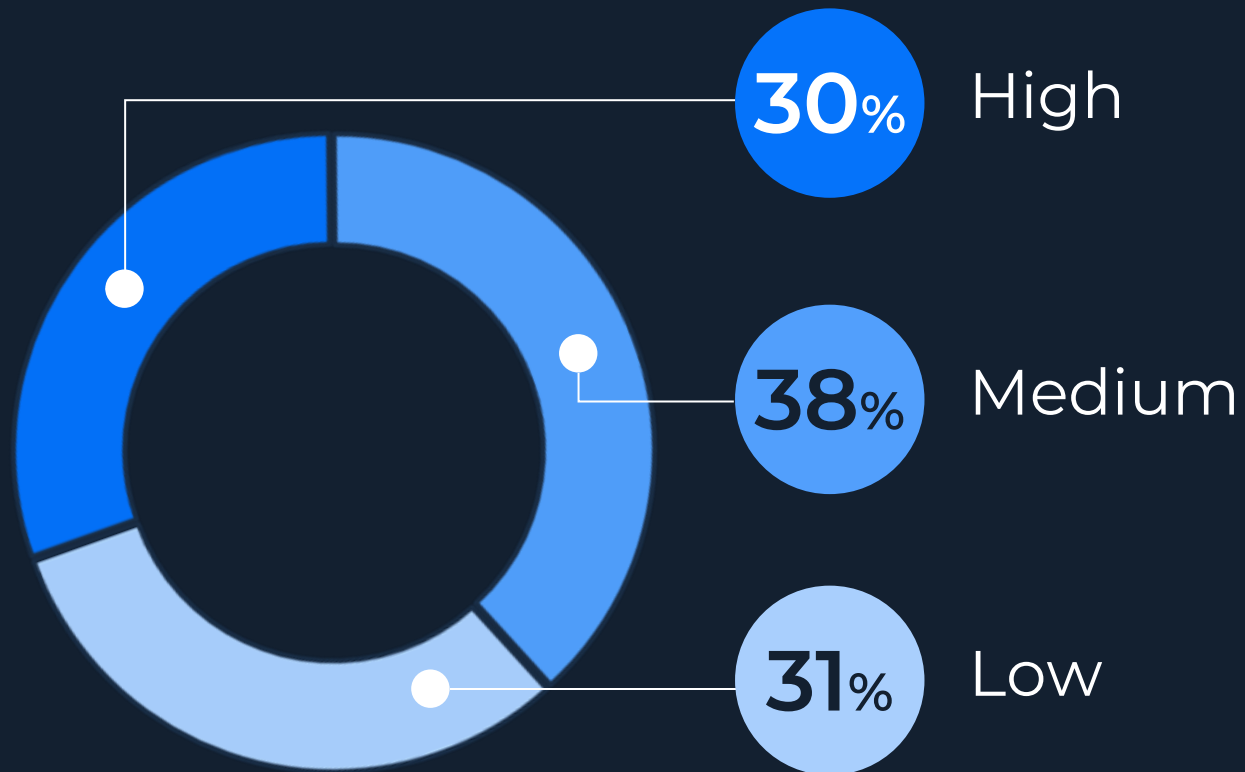
Music playlists



Email service

... but fewer can identify AI's role in all 6 examples

% of U.S. adults who correctly identify ... as using AI



Worldwide retail AI economic impact through 2029

\$9.2
trillion



54%

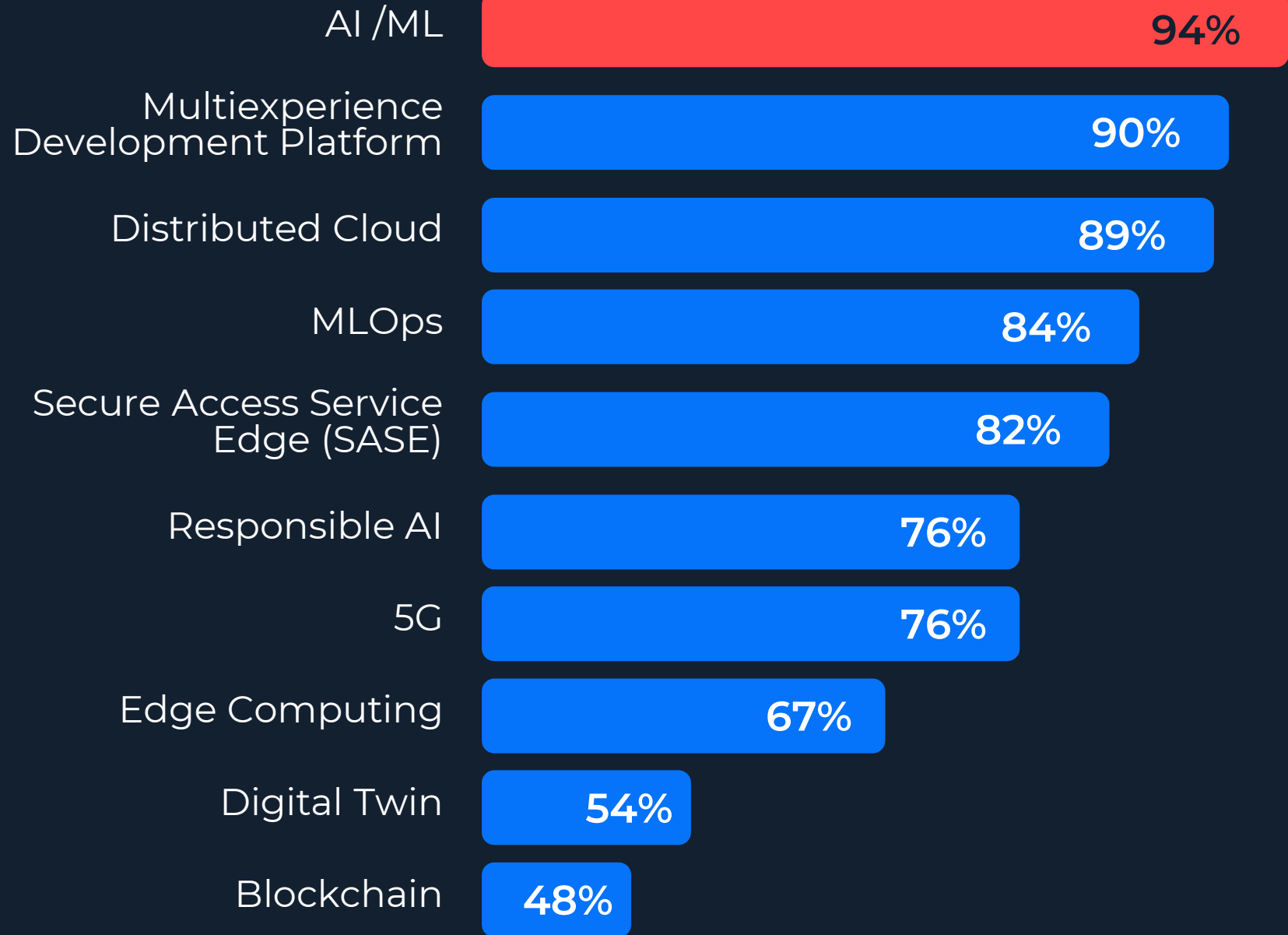
of cumulative economic benefits from AI through 2029 will be in increased sales



72%

of cumulative AI economic benefits will go to retailers over \$1b in size

Which technologies are most likely to be implemented by 2025



SOURCE: Gartner 2023 CIO Agenda
Insights for the Retail Industry

What are the challenges while developing a vision for digital change?

Integrating the digital vision with existing enterprise-level strategies

Agreeing on a shared vision across different parts of the enterprise

Competing expectations from different stakeholders



The Generative AI Era

Level-Set on Generative AI

Foundation Models

Generative AI

Large Language Models



Text



Speech



Code



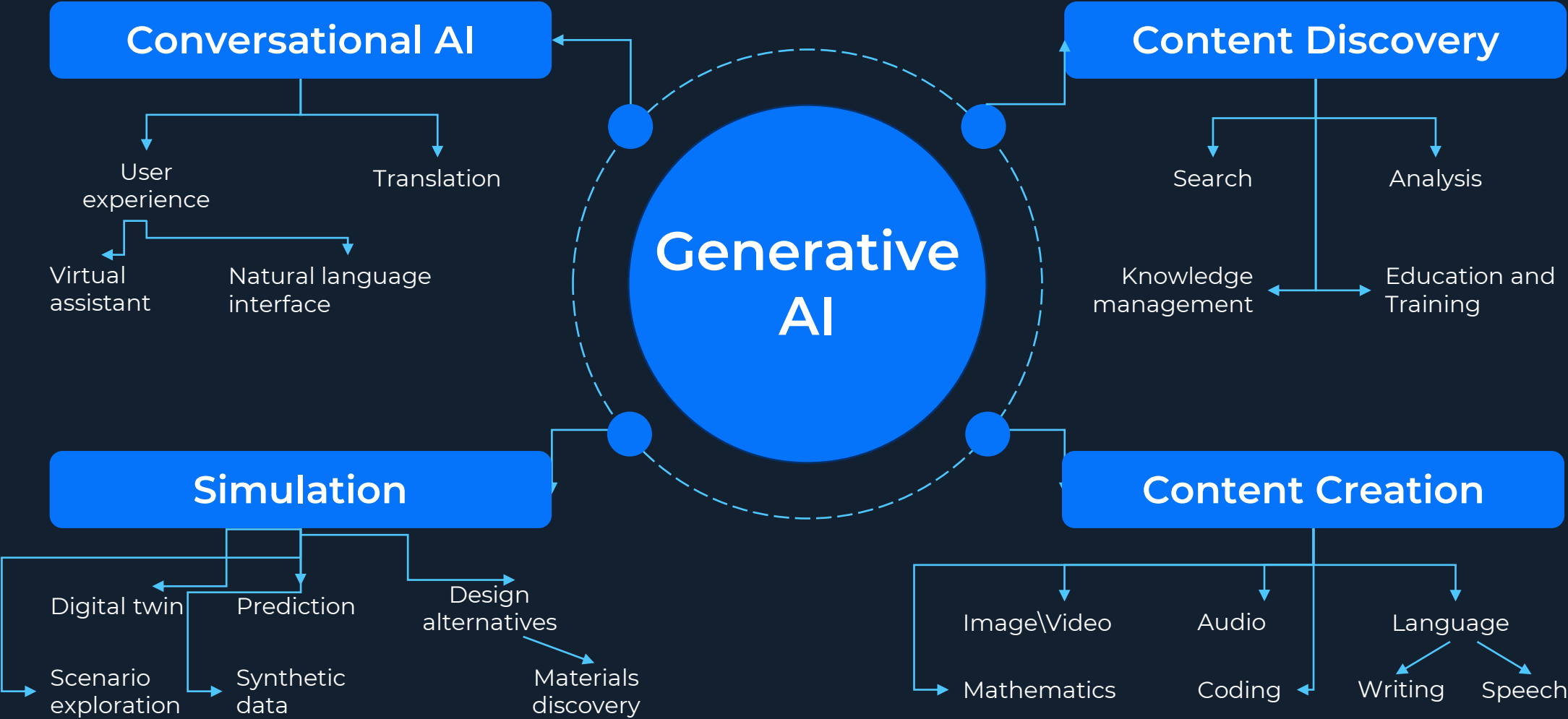
Other
Applications

Image

Video

3D

Generative AI Capabilities



You're probably using generative AI already ...

stripe

Products Solutions Developers Resources rates

Internet payment infrastructure

Increment Magazine
€14 per quarter

Pay

Or pay by card

Email

card information

Number

MM / YY

cvc

Country or region

USA

Postal Code

Introducing Virtual Volunteer™

AI powered Visual Assistant



Jasper

Quality Blog Articles

Get 10k Free Credits

Introducing Duolingo Max

The best plan for advanced learning. New AI-powered features plus all the benefits of Super.

Explain My Answer NEW
Get AI-powered personal feedback in your lessons

Roleplay NEW
Practice ordering coffee, planning a trip, and more with AI

START MY FREE 2 WEEKS

Correct answer: No me **gustan** esos vestidos.

Almost there! But when **gustar** is used with a plural item like **vestidos**, you'll use the plural form **gustan**.

Your answer, **no me gusta**, would be used for a single item, like **vestido**.

Does that help?

Yes, I'm all set

Can I see an example?

No, please elaborate

Khan Academy

For every student, every classroom. Real results.

We're a nonprofit with the mission to provide a free, world-class education for anyone, anywhere.

Get AI Guide Donate

Learners Teachers



Bard was just updated. [See update](#)

tell me about Dr Seth Dobrin, PhD

[View other drafts](#)

Dr. Seth Dobrin is a globally recognized and award-winning business leader in responsible AI. He is the former Chief AI Officer of IBM and a board member of several organizations, including the Partnership on AI, the AI for Good Foundation, and the IEEE Global Initiative on Ethics of Autonomous Systems.

Dr. Dobrin has a PhD in molecular biology from the University of California, San Francisco. He began his career in the life sciences, where he developed new technologies for genomic analysis. In 2005, he moved to IBM to lead the company's efforts in data science and analytics. He was appointed Chief AI Officer in 2017 and served in that role until 2022.

As Chief AI Officer, Dr. Dobrin was responsible for IBM's global AI strategy and execution. He led the development of IBM's Responsible AI framework, which is a set of principles for ensuring that AI is used for good. He also launched IBM's AI for Social Good program, which uses AI to solve some of the world's most pressing challenges.

Dr. Dobrin is a frequent speaker and writer on the topics of AI, ethics, and responsible leadership. He is a member of the World Economic Forum's Global Future Council on Artificial Intelligence and the Aspen Institute's Future of Artificial Intelligence Initiative.

Here are some of Dr. Dobrin's notable achievements:

- He was named one of the Top 100 Data Innovators of the Year by Analytics Insight in 2022.
- He was awarded the AI Innovator of the Year Award by the Data & Analytics Association in 2021.
- He was named one of the World's Most Influential People in AI by the MIT Technology Review in 2020.
- He was appointed to the U.S. Department of Commerce's Artificial Intelligence Advisory Committee in 2019.

Dr. Dobrin is a thought leader in the field of AI and a champion for responsible AI. He is committed to using AI to make the world a better place.

... searches
the web ...

... writes
code ...

3rd December 2022 3/3

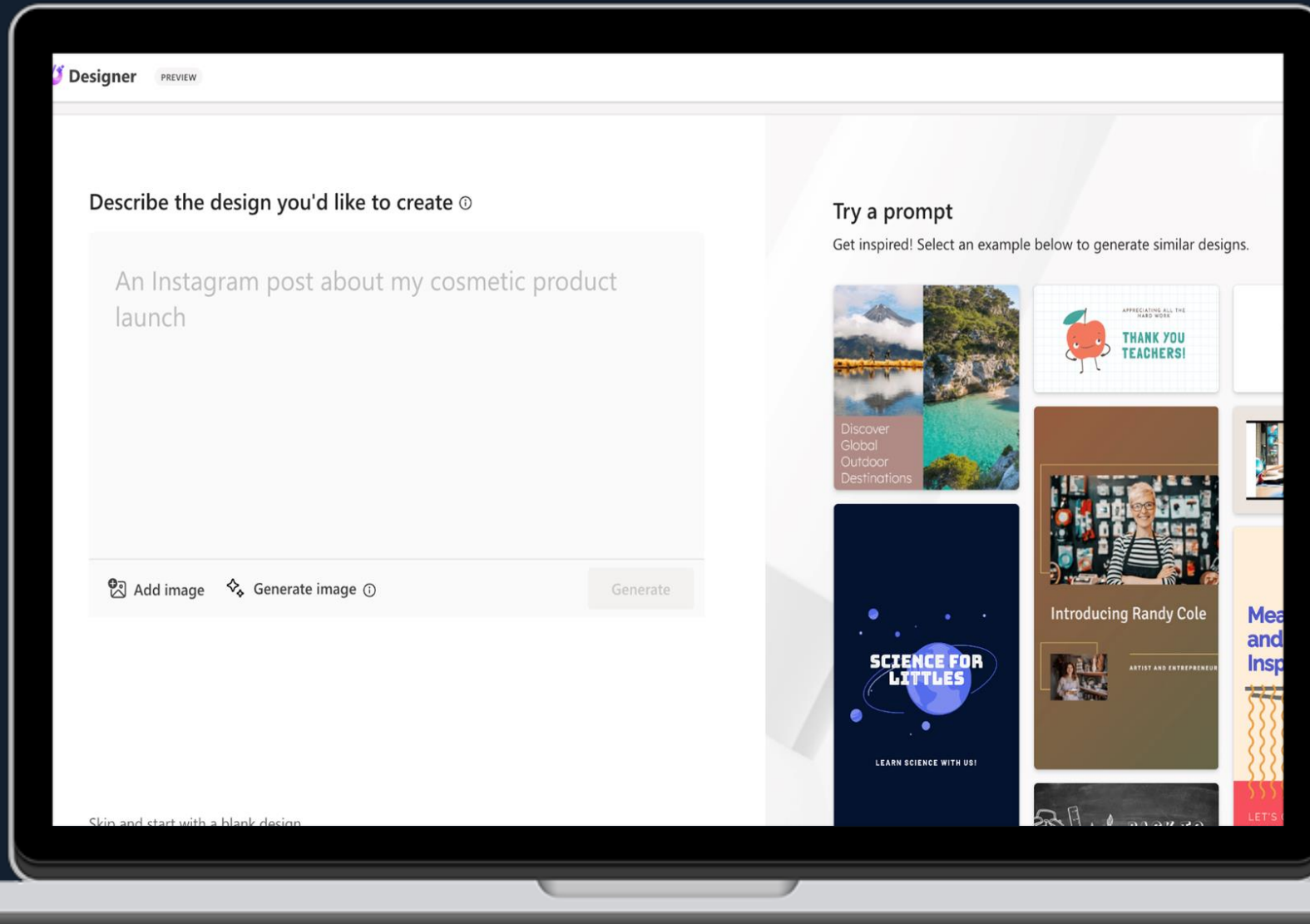
Building A Virtual Machine inside ChatGPT

JONAS DEGRAVE

411 Unless you have been living under a rock, you have heard of [this new ChatGPT assistant](#) made by OpenAI. You might be aware of its capabilities for [solving IQ tests](#), [tackling leetcode problems](#) or to [helping people write LaTeX](#). It is an amazing resource for people to retrieve all kinds of information and solve tedious tasks, like copy-writing!

Today, Frederic Besse told me that he managed to do something different. Did you know, that you can run a whole virtual machine inside of ChatGPT?

... creates designs ...





Is AI nirvana
here? ...

... small
data
models
from large
models ...

Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality

by the Team with members from UC Berkeley, CMU, Stanford, and UC San Diego

** According to a fun and non-scientific evaluation with GPT-4. Further rigorous evaluation is needed.*

We introduce Vicuna-13B, an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. Preliminary evaluation using GPT-4 as a judge shows Vicuna-13B achieves more than 90%* quality of OpenAI ChatGPT and Google Bard while outperforming other models like LLaMA and Stanford Alpaca in more than 90%* of cases. The cost of training Vicuna-13B is around \$300. The training and serving [code](#), along with an online [demo](#), are publicly available for non-commercial use.



BloombergGPT: A Large Language Model for Finance

Shijie Wu^{1,*}, Ozan İrsoy^{1,*}, Steven Lu^{1,*}, Vadim Dabravolski¹, Mark Dredze^{1,2}, Sebastian Gehrmann¹, Prabhanjan Kambadur¹, David Rosenberg¹, Gideon Mann¹

¹ Bloomberg, New York, NY USA

² Computer Science, Johns Hopkins University, Baltimore, MD USA

gmann16@bloomberg.net

Abstract

The use of NLP in the realm of financial technology is broad and complex, with applications ranging from sentiment analysis and named entity recognition to question answering. Large Language Models (LLMs) have been shown to be effective on a variety of tasks; however, no LLM specialized for the financial domain has been reported in literature. In this work, we present BLOOMBERGGPT, a 50 billion parameter language model that is trained on a wide range of financial data. We construct a 363 billion token dataset based on Bloomberg's extensive data sources, perhaps the largest domain-specific dataset yet, augmented with 345 billion tokens from general purpose datasets. We validate BLOOMBERGGPT on standard LLM benchmarks, open financial benchmarks, and a suite of internal benchmarks that most accurately reflect our intended usage. Our mixed dataset training leads to a model that outperforms existing models on financial tasks by significant margins without sacrificing performance on general LLM benchmarks. Additionally, we explain our modeling choices, training process, and evaluation methodology. As a next step, we plan to release training logs (Chronicles) detailing our experience in training BLOOMBERGGPT.

... proprietary
industry-
specific large
models ...

... open source
industry-
specific large
models ...

A Large Language Model for the NHS



Prof James Teo

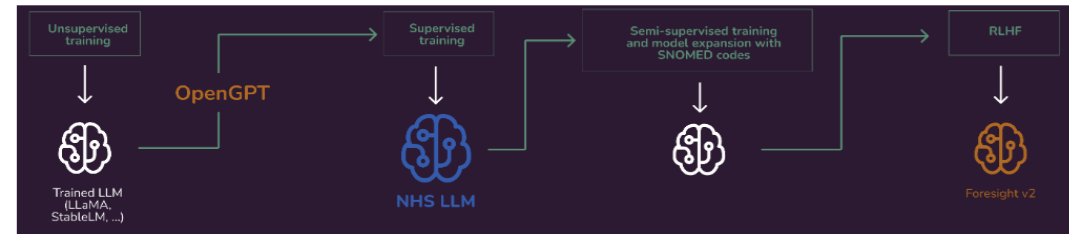
Clinical Director of AI and Data; Professor of Neurology

6 articles

+ Follow

May 10, 2023

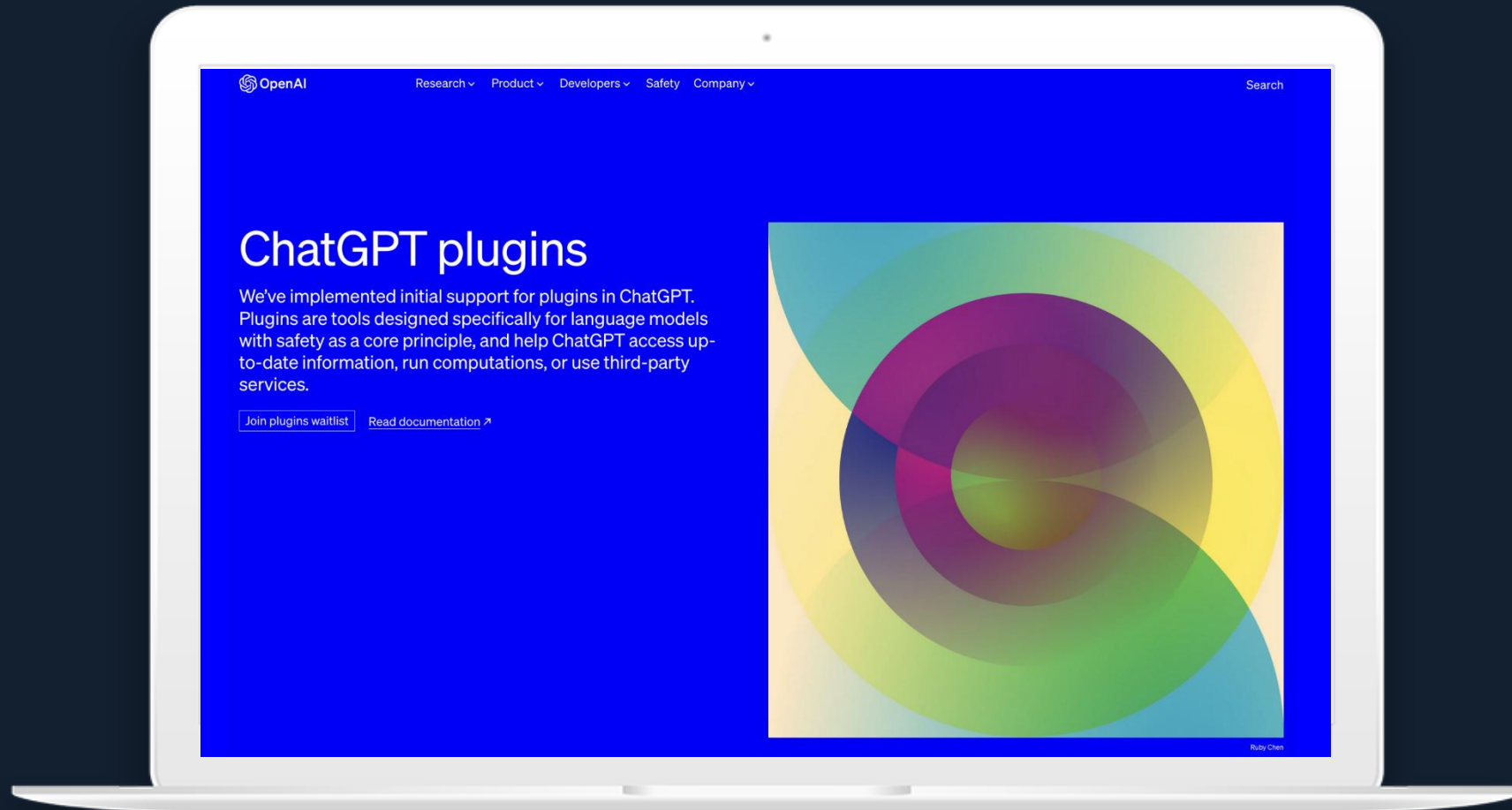
In 2022, [CogStack](#) introduced Foresight version 1 GPT. The team at [King's College London](#), [King's College Hospital NHS Foundation Trust](#), [Institute of Psychiatry, Psychology & Neuroscience](#) and [Guy's and St Thomas' NHS Foundation Trust](#) now present a preview of the next phase of evolution with a Large Language Model for the [#nhs](#)



For the [#NHS](#) healthcare community [#ai4nhs](#) [#aiforhealth](#) , we introduce:

1. OpenGPT - an open-source framework for building LLMs with supervised training and instruction-based datasets ([github.com](#))
2. NHS-LLM - a 13B large language model trained for healthcare

... We are seeing the birth of the next “App Store”...





Maybe not ...

**Problem:
LLMs are not
safe
for business
applications**

Generative AI Models



... and growing weekly

The Technology Gap



No concept of organizational truths



Presents a regulatory risk



No or incomplete evidence



Hallucinations and incorrect responses



Not auditable



No protection against prompt injection

AI that
Lies ...



47%
fabricated

High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content

Mehul Bhattacharyya ¹, Valerie M. Miller ², Debjani Bhattacharyya ³, Larry E. Miller ¹

1. Clinical Research, Miller Scientific, Johnson City, USA 2. Leadership, University of the Cumberlands, Williamsburg, USA 3. Education, University of Massachusetts Lowell, Lowell, USA

Corresponding author: Larry E. Miller, larry@millerscientific.com

Abstract

Background

The availability of large language models such as Chat Generative Pre-trained Transformer (ChatGPT, OpenAI) has enabled individuals from diverse backgrounds to access medical information. However, concerns exist about the accuracy of ChatGPT responses and the references used to generate medical content.

Methods

This observational study investigated the authenticity and accuracy of references in medical articles generated by ChatGPT. ChatGPT-3.5 generated 30 short medical papers, each with at least three references, based on standardized prompts encompassing various topics and therapeutic areas. Reference authenticity and accuracy were verified by searching Medline, Google Scholar, and the Directory of Open Access Journals. The authenticity and accuracy of individual ChatGPT-generated reference elements were also determined.

Results

Overall, 115 references were generated by ChatGPT, with a mean of 3.8±1.1 per paper. Among these references, 47% were fabricated, 46% were authentic but inaccurate, and only 7% were authentic and accurate. The likelihood of fabricated references significantly differed based on prompt variations; yet the

46%
inaccurate
...

... makes up
information
and
references

Lawyer's AI Blunder Shows Perils of ChatGPT



Justin Wise
Reporter

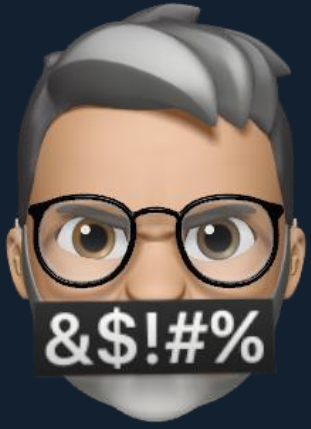


- NY lawyers face possible sanctions over fabricated case citations
- The case comes as AI technology becomes more widely used

A lawyer's citation of court decisions fabricated by ChatGPT shows the peril of relying on the artificial intelligence chatbot without proper safeguards.

New York lawyers Steven Schwartz and Peter LoDuca face a June 8 [hearing](#) on potential sanctions after a court brief they submitted cited six nonexistent cases. Schwartz acknowledged that ChatGPT invented the cases, even though he initially believed the tool had surfaced authentic citations, according to an affidavit he filed May 25 in Manhattan federal court.

"Maybe this is an extreme example" of lawyer over-reliance on ChatGPT, said



AI that has

regulatory issues ...



The legal and regulatory issues aren't yet understood ...

MIT Technology Review

ARTIFICIAL INTELLIGENCE

OpenAI's hunger for data is coming back to bite it

The company's AI services may be breaking data protection laws, and there is no resolution in sight.

April 19, 2023

By Melissa Heikkilä

Five Key Legal Issues to Consider When It Comes to Generative AI

PRACTICAL LAW THE JOURNAL

Issue Sections Contributors About Us



ChatGPT and Generative AI: Transactions June 2023

The Verge / Tech / Reviews / Science / Entertainment



OpenAI's regulatory troubles are only just beginning

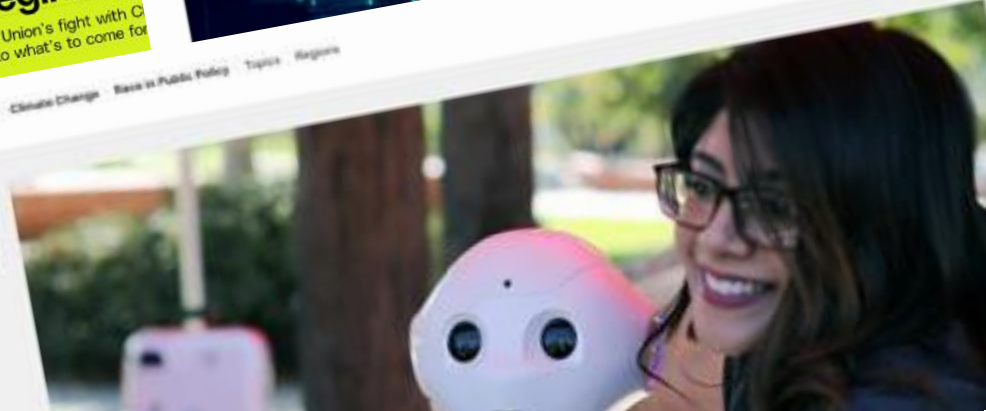
The European Union's fight with OpenAI is a glance into what's to come for AI services.

BROOKINGS

Around the halls: What should the regulation of generative AI look like?

By Aaron Vargot, Mark MacCarthy, Tom Wheeler

U.S. Economy Basic Technology & Information Climate Change Basic in Public Policy Topics Regions



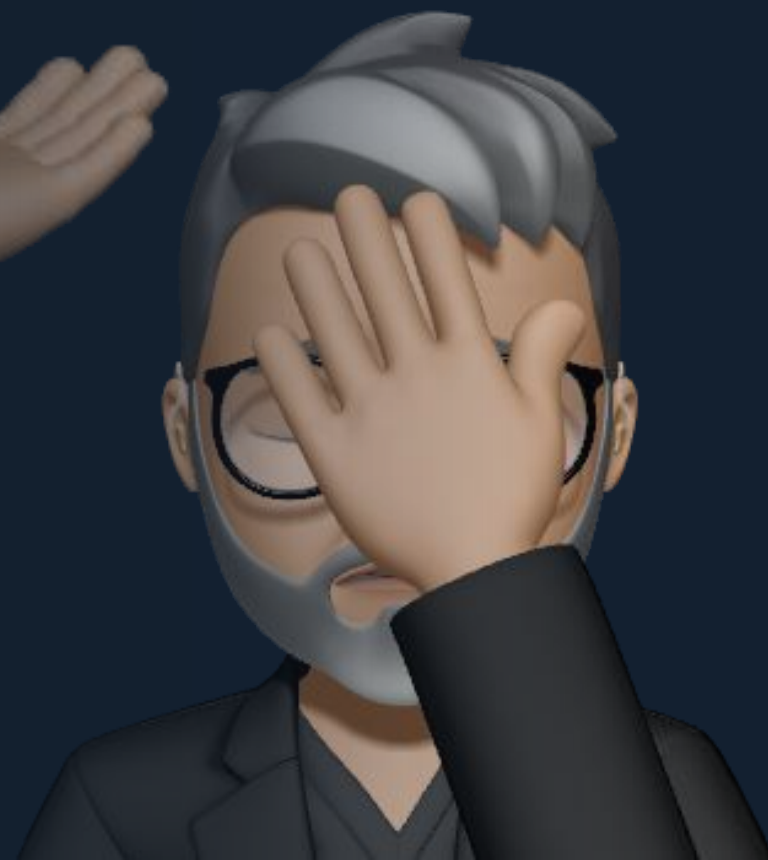
Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21 labs	ALEPH ALPHA	ELEutherAI	Totals
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ● ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ● ○	27
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ●	● ● ● ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	16
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	15
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	10
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ● ○	21
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	9
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

... is not compliant with global regulations

What
about ...



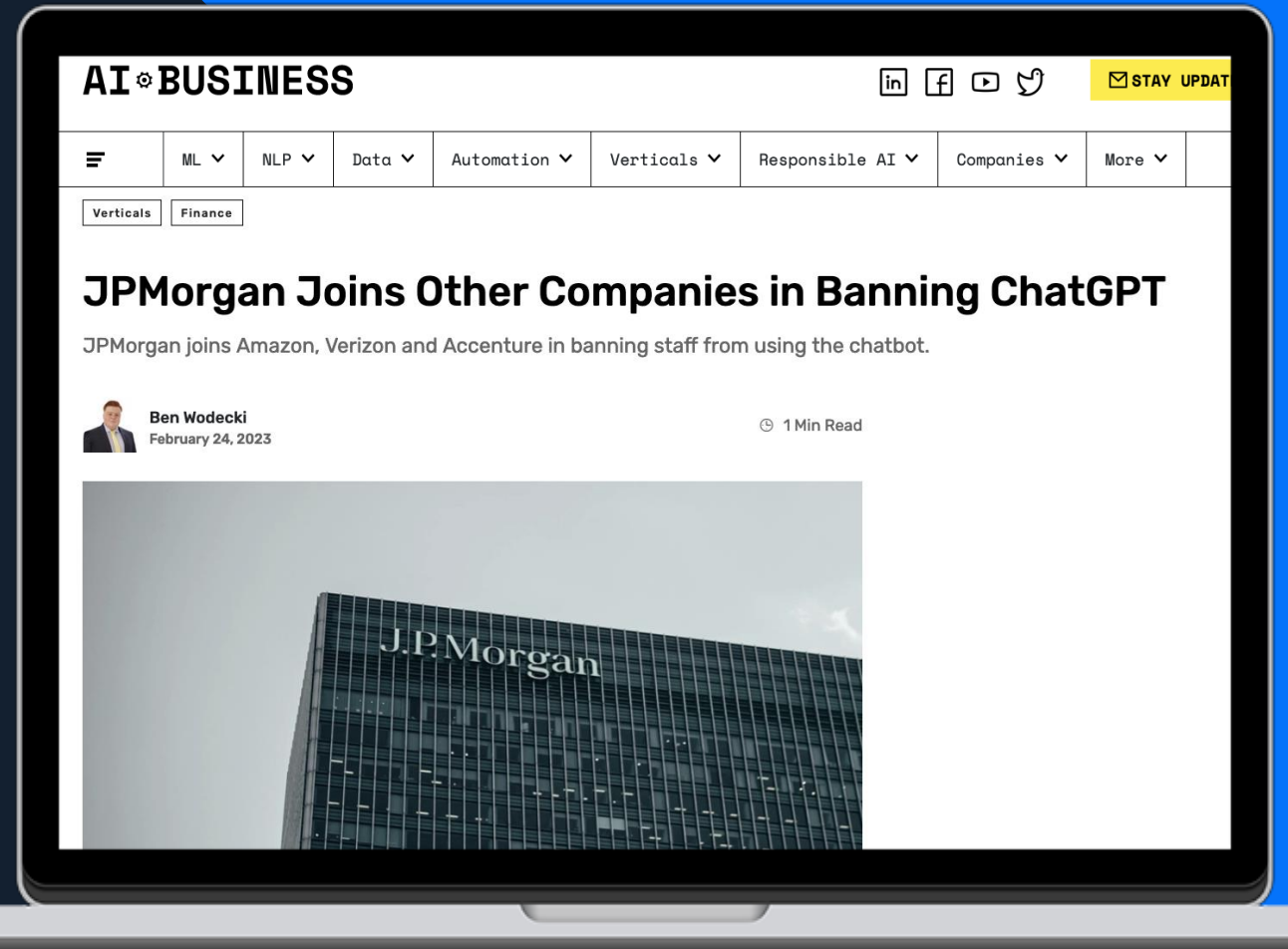
ChatGPT banned from New York City public schools' devices and networks

A spokesperson for OpenAI, which developed ChatGPT, said it is "already developing mitigations to help anyone identify text generated by that system."



Technology
being
banned from
schools ...

... or the
corporate
world?



TECHNOLOGY

ChatGPT is temporarily banned in Italy amid an investigation into data collection

March 31, 2023 · 5:06 PM ET



Juliana Kim



... or entire countries ...

... or the AI
industry
itself ...



[Our mission](#)

[Cause areas](#) ▾

[Our work](#) ▾

[About us](#) ▾

[Home](#) » [Pause Giant AI Experiments: An Open Letter](#)

[← All Open Letters](#)

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

2390

[Add your signature](#)



Now
what ...



FORMULA

for safe and compliant
Generative AI

PIXAR



PIXAR

ANIMATION STUDIOS

Pixar's secret to success?

A formula

Once upon a time there was a fish, Marlin, and his son Nemo

Every day, Marlin warns Nemo about the dangers of the ocean

One day, Nemo ignores his father

Because of that, Nemo winds up in a fish tank

Because of that, Marlin sets off on a journey to find Nemo

Until finally, Marlin finds Nemo and brings him home safely

Implement generative AI in a safe and compliant manner

A formula

Create top-down AI strategy aligned to business objectives.

Update governance and accountability to reflect new challenges.

Prioritize use cases to create real value quickly.

Provide education from the mailroom to the board room.

Implement technology stacks designed for generative AI.

Where
to start ...



Start here ...



Top-down AI strategy aligned to business objectives



Update governance and accountability to reflect new challenges



Prioritization of use cases dictate the right model for the job



Provide role-based education from the mailroom to the board room



Implement architectures designed for generative AI

Select Generative AI Use Cases by Industry	Industries							
	Automotive and Vehicle Manufacturing	Media	Architecture and Engineering	Energy and Utilities	Healthcare Providers	Electronic Product Manufacturing	Manufacturing	Pharmaceutical
Drug Design								✓
Material Science	✓			✓		✓		
Chip Design						✓		
Synthetic Data	✓		✓	✓	✓	✓	✓	✓
Generative Design (Parts)	✓		✓				✓	

Cross-Industry Use Cases



Customer
Service



Sustainability
and ESG



Sourcing
and Supply
Chain



Cybersecurity

Retail



Generative AI Use Cases for Retail



Marketing
and Digital Commerce



Customer
Experience



Operations
and Supply Chain

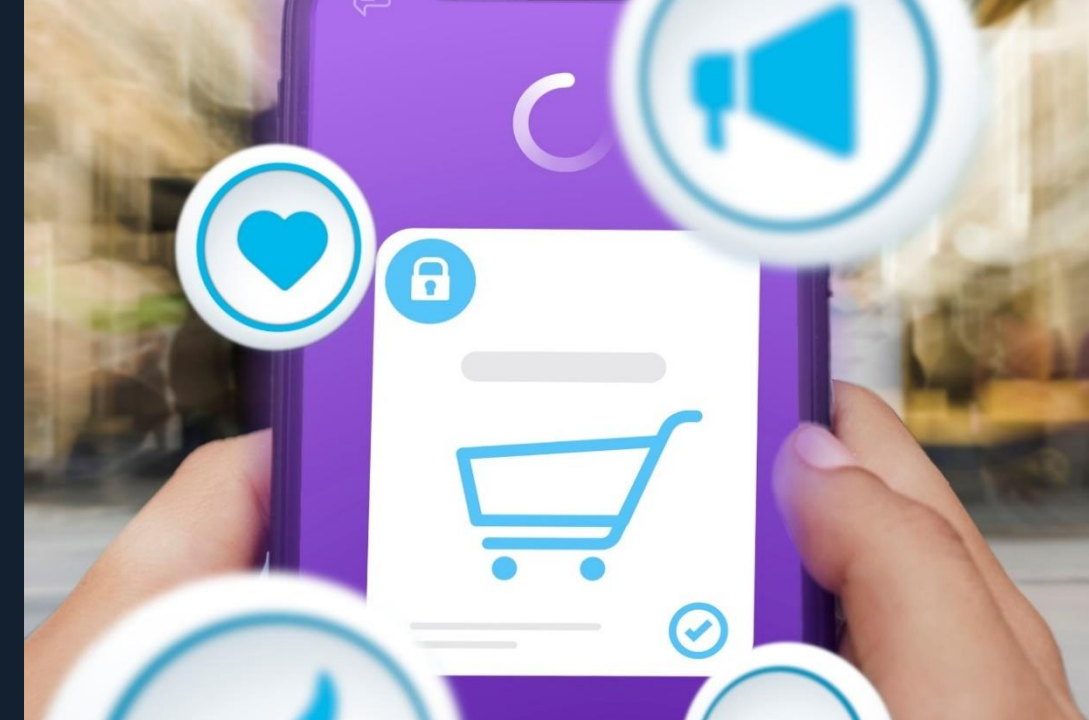


Merchandising
and Planning

Marketing and Digital Commerce

Focus Areas

- › Personalized messaging
- › Social media
- › Product attributes
- › Descriptions of products
- › Images of products/services



Risks

- › Bias
- › Intellectual property
- › Consumer privacy
- › Copyright
- › Deepfakes

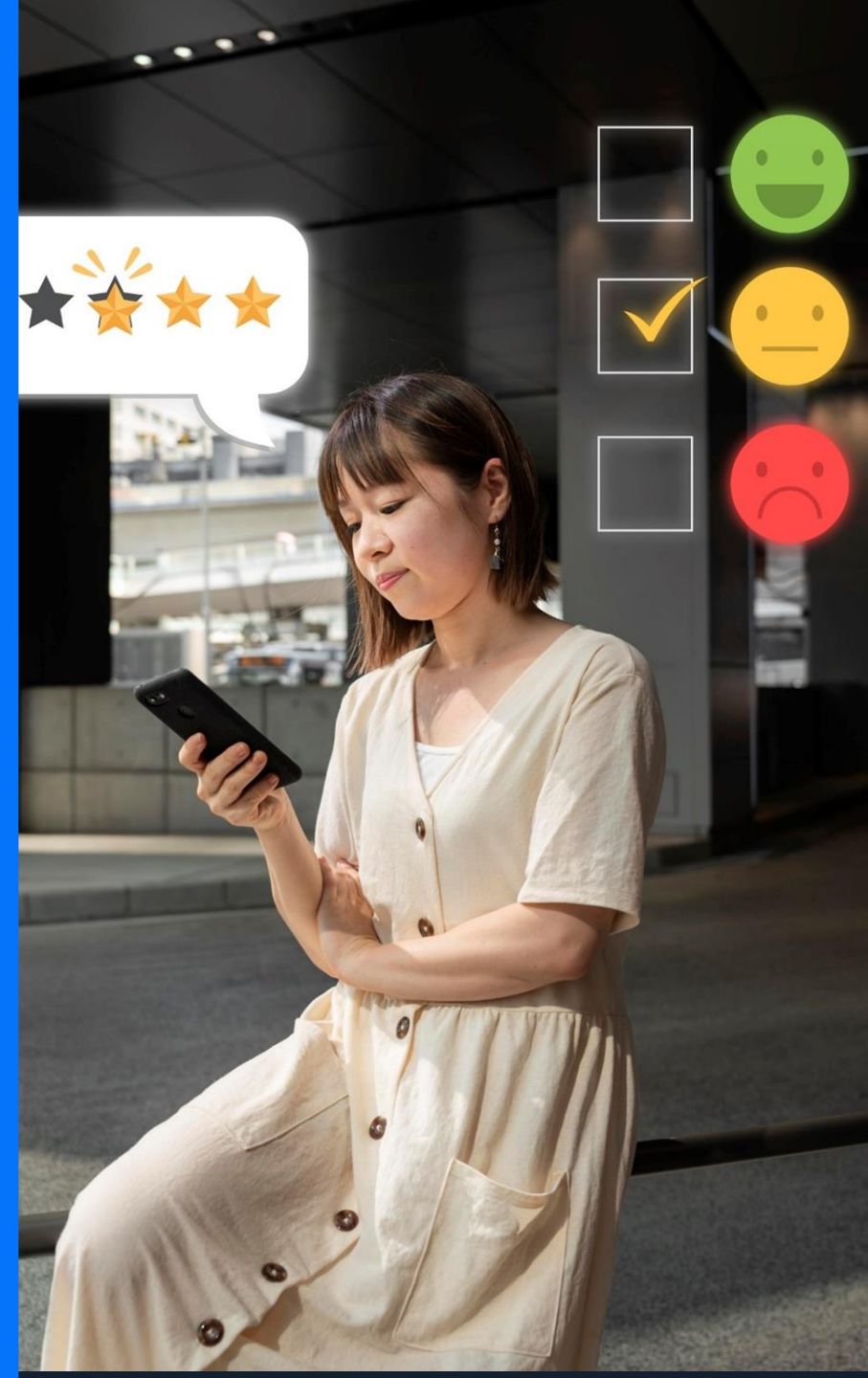
Customer Experience

Focus Areas

- › Conversational commerce
- › Transaction facilitation
- › Associate assistance
- › Customer service
- › Customer support

Risks

- › Bias
- › Poor recommendation
- › Process change management
- › Lack of consumer privacy
- › Lack of consumer protections



Operations and Supply Chain

Focus Areas

Risks

- Bias
- Incorrect service recommendation
- Process change management
- Lack of transparency
- Lack of processing capacity

- Content discovery
- Sourcing
- Procurement
- Logistics
- Supplier interactions



Merchandising and Planning

Focus Areas

- › Product development
- › Trend analysis
- › Social media analysis
- › Image creation
- › Synthetic data creation

Risks

- › Bias
- › Failure to identify change
- › Pricing changes too frequently
- › Lack of real-time information
- › Lack of processing capacity





Elevate your AI IQ

Dr. Seth Dobrin

CEO, Qantm AI

Seth@Qantm.AI

