

GenAI in Action For Enterprises

Erwan Menard

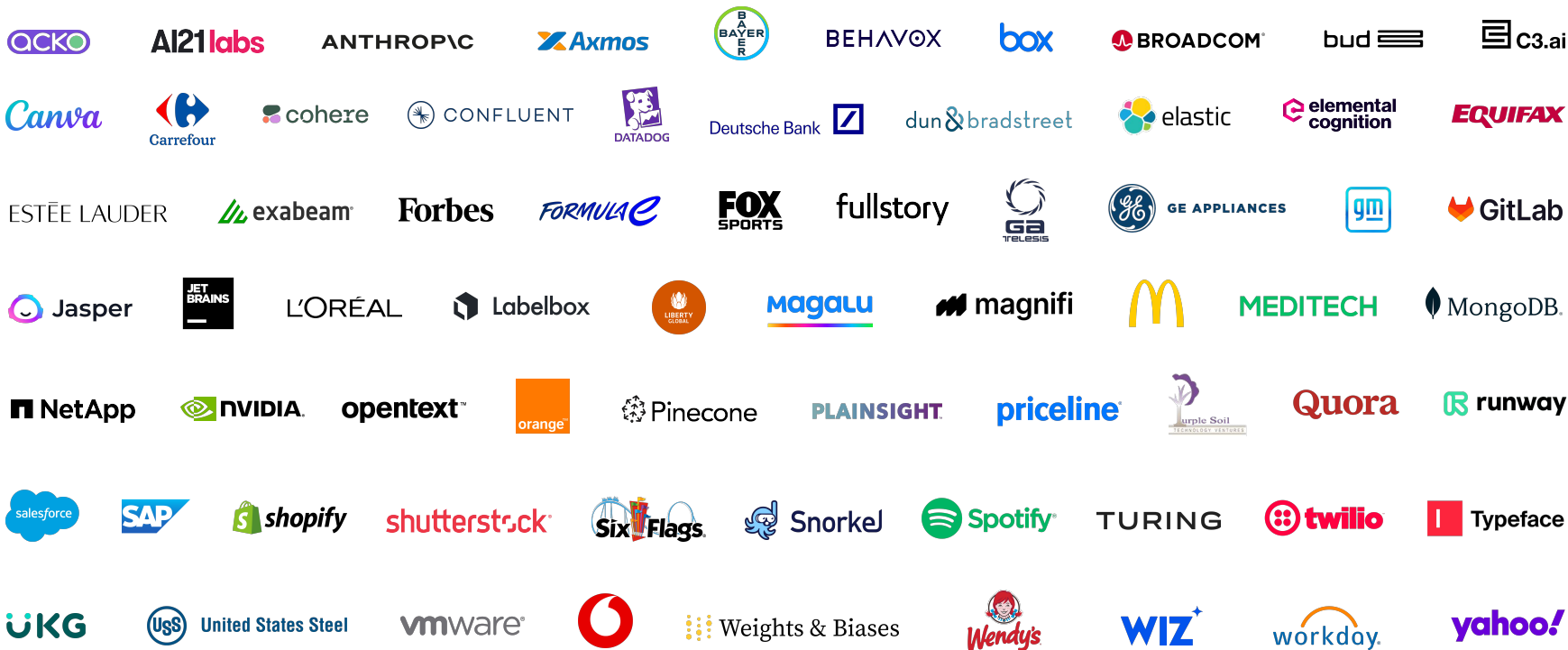
Director Product Management
Cloud AI

Reply Innovation Tour
March 21th, 2024



**Generative AI is
transforming how we
interact with technology**

Generative AI with Google Cloud Customers



Initial Patterns

Launching in Production

B2B2C GenAI

Powering consumer services with Vertex AI Search (and more)

B2B2B via ISV Assistants

Powering business services based on customer private corpus

Contact Center AI

Modernizing Customer Service

Search
(incl. Retail, Media, Healthcare)

Modernizing Website Search and Navigation

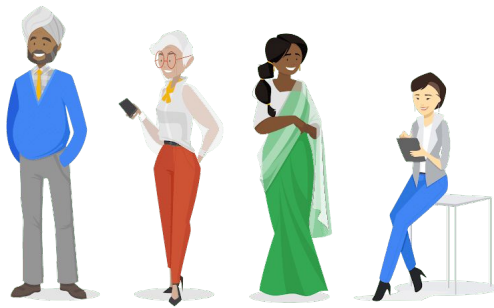
Knowledge Worker

Addressing internal use cases first to pace adoption

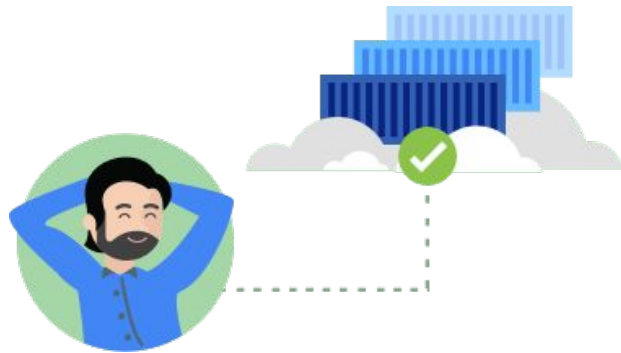
Establishing the Corporate AI Platform

Predictive + Generative

Over the last 18 months we've seen the nature of
ML development and **AI usage** change



Powerful managed models in
the hands of developers



The ability to tune foundation
models with small amounts of data

The latest of **Generative AI** innovation Gemini 1.5 Pro with large context window



Revolutionizing interactions with LLMs

Up to 1hr video or 500 pages of text in one prompt



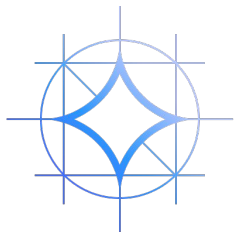
Google Model Ecosystem



Gemini: Google's largest and most capable AI models.

The powerhouse of the family, offering unmatched capabilities for cloud-based AI projects.

- **Gemini Ultra:** Most capable model for large-scale highly complex text and image reasoning.
- **Gemini Pro:** The best performing model with features for a wide variety of use cases.
- **Gemini Nano:** Google's Android model. The pioneer of on-device AI, empowering Android developers to push the boundaries of mobile applications.

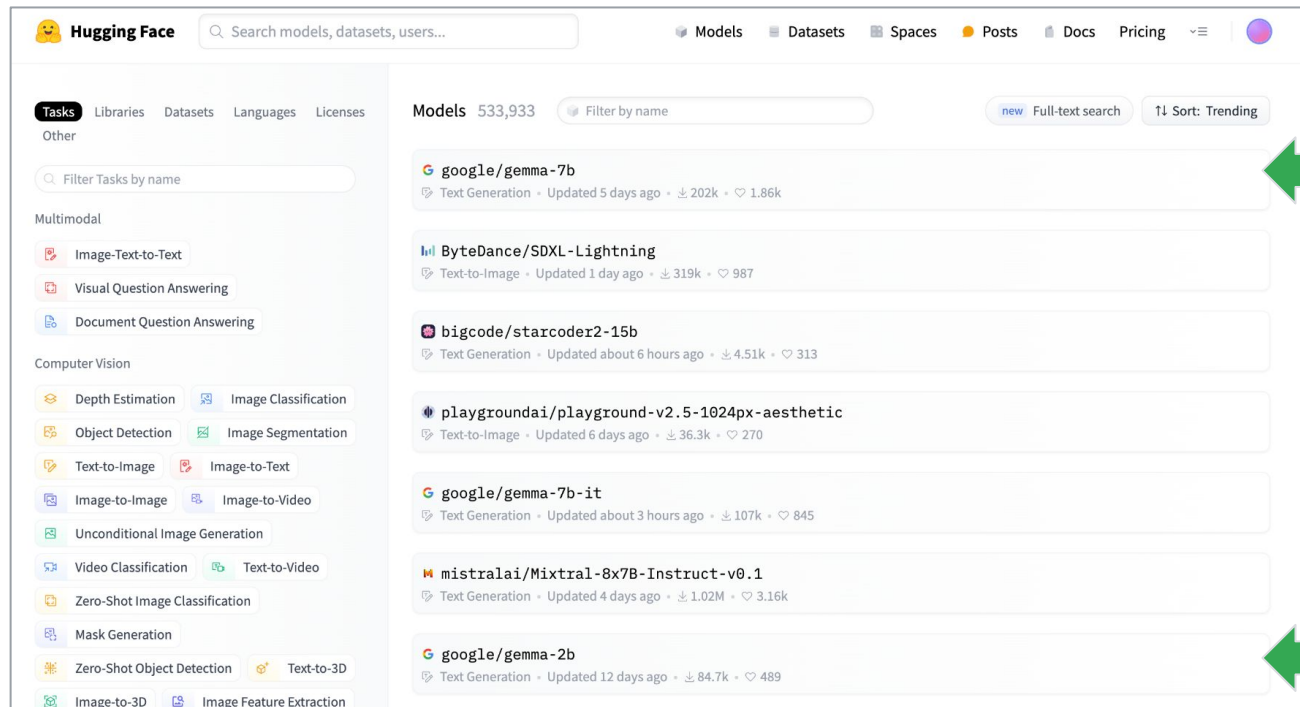


Gemma

Gemma: Google's open models. Gemma is built for the open community of developers and researchers powering AI innovation.

- **Gemma 2B and 7B:** Small & efficient models for narrowly defined, text-based tasks

Strong community interest for Gemma



The screenshot shows the Hugging Face website's Models page. The left sidebar contains navigation links for Tasks, Libraries, Datasets, Languages, and Licenses. The main content area displays a list of models, with 'google/gemma-7b' at the top, followed by 'ByteDance/SDXL-Lightning', 'bigcode/starcoder2-15b', 'playgroundai/playground-v2.5-1024px-aesthetic', 'google/gemma-7b-it', 'mistralai/Mixtral-8x7B-Instruct-v0.1', and 'google/gemma-2b'. Two green arrows point to the 'google/gemma-7b' and 'google/gemma-2b' model entries, highlighting their high ranking and popularity.

Hugging Face Search models, datasets, users...

Models 533,933 Filter by name new Full-text search Sort: Trending

google/gemma-7b
Text Generation • Updated 5 days ago • 202k • 1.86k

ByteDance/SDXL-Lightning
Text-to-Image • Updated 1 day ago • 319k • 987

bigcode/starcoder2-15b
Text Generation • Updated about 6 hours ago • 4.51k • 313

playgroundai/playground-v2.5-1024px-aesthetic
Text-to-Image • Updated 6 days ago • 36.3k • 270

google/gemma-7b-it
Text Generation • Updated about 3 hours ago • 107k • 845


mistralai/Mixtral-8x7B-Instruct-v0.1
Text Generation • Updated 4 days ago • 1.02M • 3.16k

google/gemma-2b
Text Generation • Updated 12 days ago • 84.7k • 489



Gemma

Google AI Experiences, powered by Gemini and Gemma



Meet business users, developers and AI practitioners
where they are, with all Google innovation

Consumer assistant: Gemini

AI Hobbyist and Developers: Google AI Studio, ai.google.dev, Hugging Face

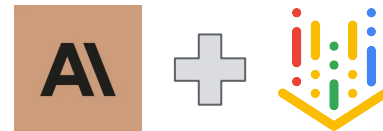
Enterprise assistant: Gemini for Business as part of Google Workspace

Enterprise AI developer and builder platform: Vertex AI on Google Cloud

Anthropic's Claude 3 models in Google Cloud Vertex AI

GA March 19, 2024

- **Claude 3 Opus:** Anthropic's most capable and intelligent model yet.
- **Claude 3 Sonnet:** Anthropic's best combination of skills and speed.
- **Claude 3 Haiku:** Anthropic's fastest, most compact model.



130+ Enterprise-ready Foundation Models

Vertex AI Model Garden

Gemini Foundation Models	<div>Gemini 1.0 Pro</div> <div>Gemini 1.5 Pro</div> <div>Gemini 1.0 Ultra</div>				
Google Foundation Models	<div>PaLM 2</div>	<div>Imagen 2</div>	<div>Chirp</div>	<div>Codey</div>	<div>Embeddings API</div> <div>Embeddings</div>
Google Task Specific Models	<div>Speech-to-Text</div>	<div>Text-to-Speech</div>	<div>Natural Language</div>	<div>Translation</div>	<div>Doc AI OCR</div> <div>Occupancy analytics</div> <div>Vision</div> <div>Video Intelligence</div>
Google Domain Specific Models	<div>MedLM</div> <div>Life Science and Healthcare</div>	<div>Sec-PaLM</div> <div>Cybersecurity</div>			
Partner & Open Ecosystem	<div>Llama 2</div> <div>Code Llama</div>	<div>Falcon</div>	<div>Claude 2</div> <div>Pre-announce</div>	<div>MISTRAL AI</div>	<div>Gemma</div>



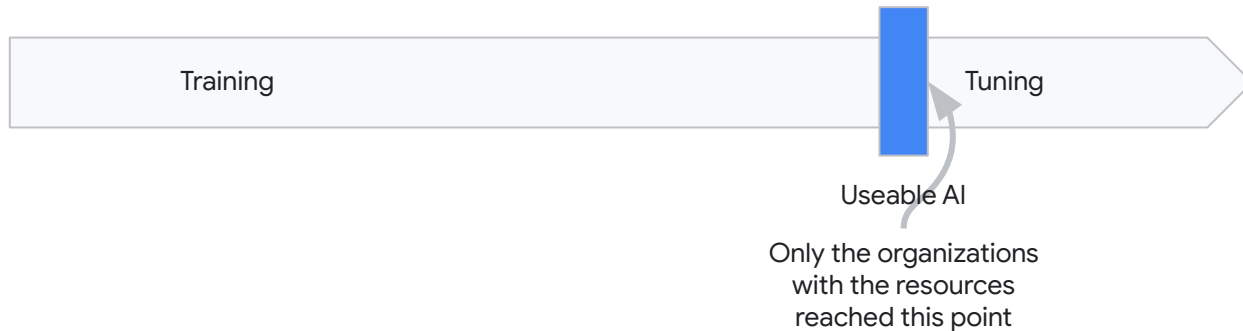
Vertex AI Model Garden

- **Choice and flexibility** with Google, open source, and third-party foundation models
- **Multiple modalities** to match every use case
- **Multiple model sizes** to match cost and efficacy needs
- **Domain-specific models** for specialized industries
- Enterprise ready with **safety, security, and responsibility**
- Decrease time to value with **fully integrated platform**

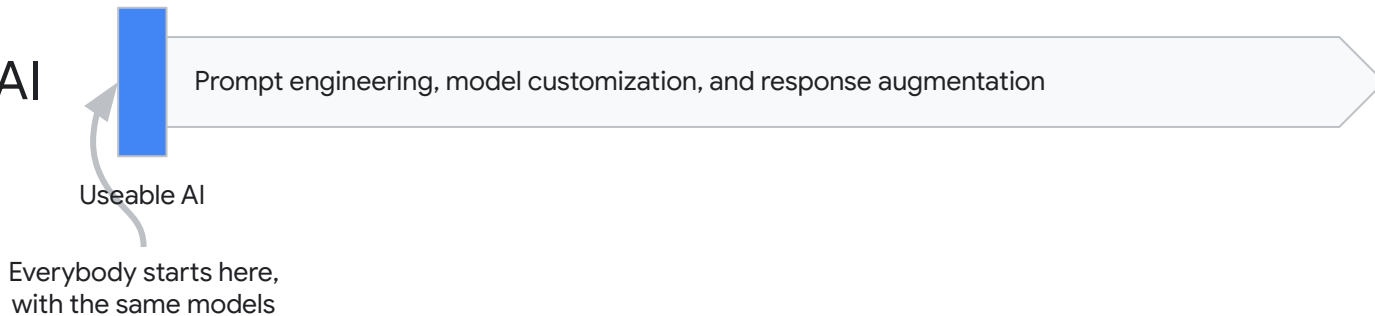
**Generative AI is not only
about model innovation**

AI Paradigm Shift

Predictive AI

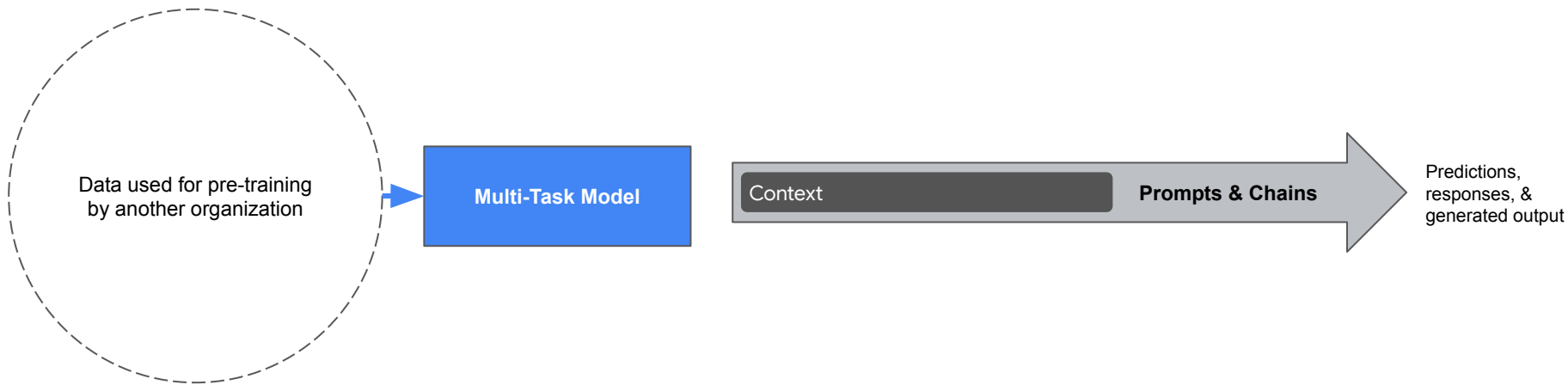


Generative AI



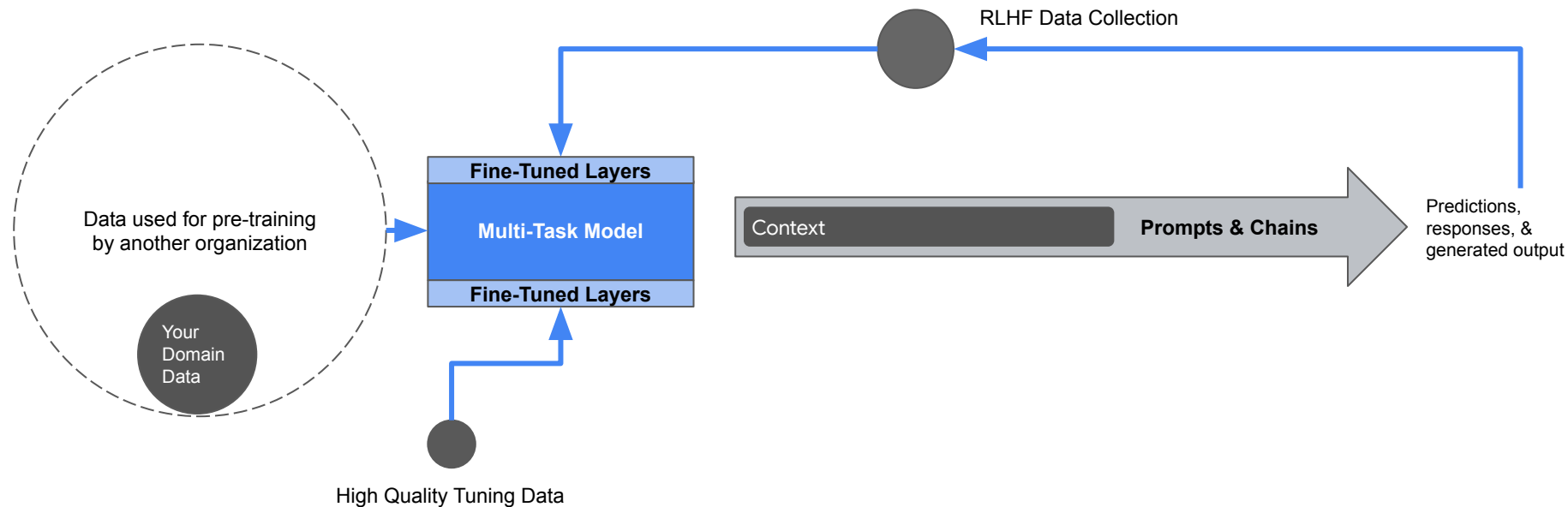
From AI Models to Enterprise Data & AI Systems

Off-the-Shelf Models



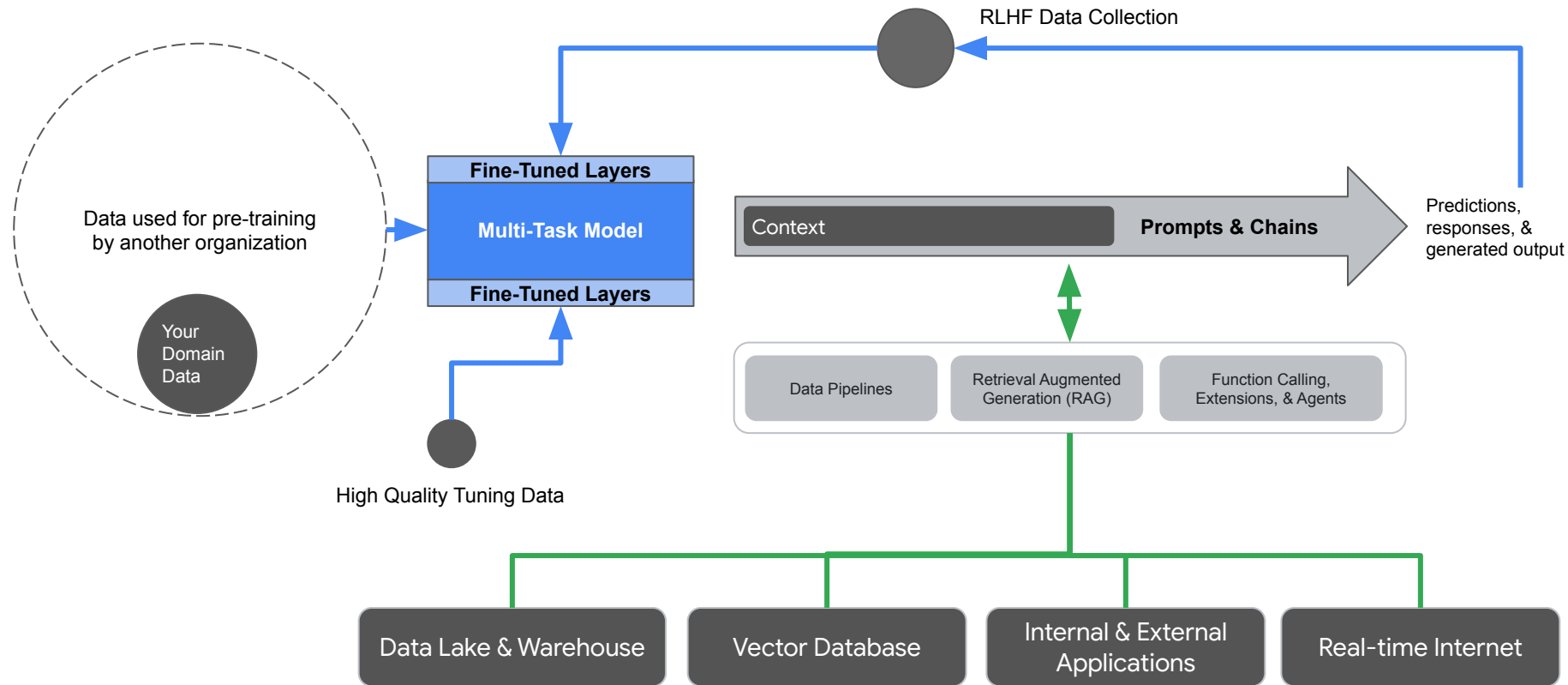
From AI Models to Enterprise Data & AI Systems

Customizing Models



From AI Models to Enterprise Data & AI Systems

Augmenting Models



Strategic takeaways

- 2024 = going to production
- Expect continuous pace of innovation
- Adopt a platform, not a model
- A new era for developers

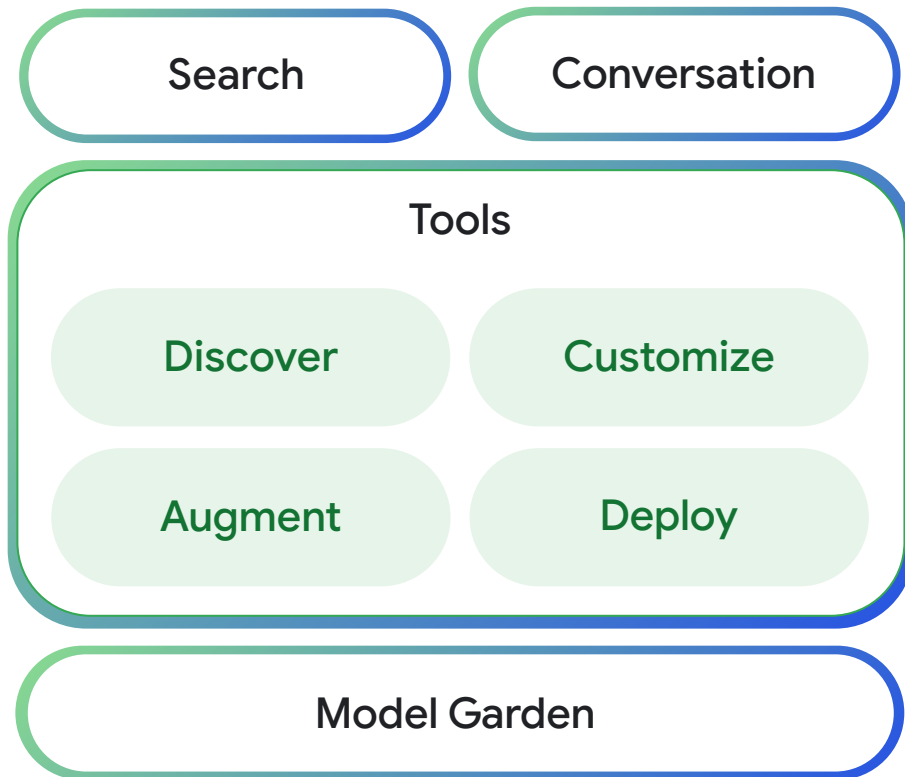
Generative AI in Action



Vertex AI

Beyond the models,
It's about the **platform**

Google Vertex AI, the
**predictive and
generative** AI platform



Platform and Tools for Builders



Tune & customize
to meet your
needs



Have confidence in
quality & safety



Augment with
real-world
information



Add new
capabilities



Evaluate & deploy
with ease



Keep your data
secure, private,
sovereign

Built on a foundation of **enterprise readiness**

Whether 1st-party, 3rd-party, or open-source, Google Cloud gives you the tools, services, and infrastructure to make every deployment enterprise ready



Data governance,
indemnity, and
privacy



Security and
compliance
support



Infrastructure
reliability and
sustainability



Responsible AI

Vertex AI Serves the Full Spectrum of Developers

Low/No Code Development Journey

Google-quality out-of-the-box experience
for implementers

Fast and easy setup

API-First Development Journey

Build Your Own ML/AI to leverage your
proprietary institutional knowledge

Model training/ tuning



General Developers

No ML skill required

AI Practitioners

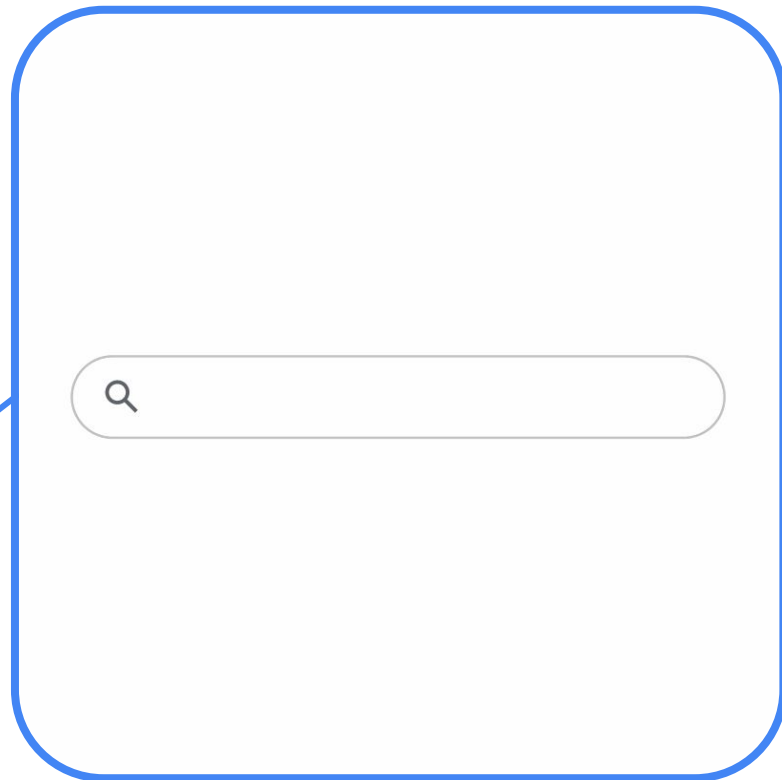
Advanced ML skill



Search with Vertex AI

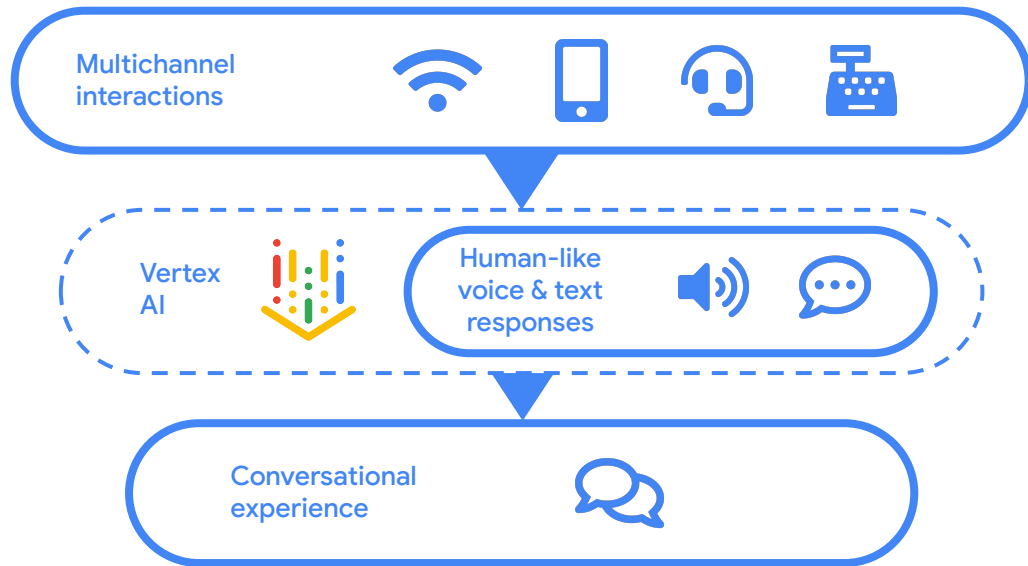
Gemini will power multimodal,
advanced search

Answer Generation
Summarization
Blended Search
Advanced Search



Conversation with Vertex AI

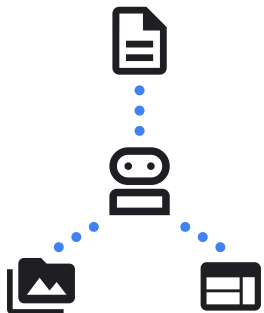
Gemini will enable multichannel,
multimodal conversations



With Vertex AI Conversation developers can easily advance bots from informational to transactional

01

Connect to your data



02

Config/build extensions



Code Interpreter



Vertex AI Search



BigQuery



AlloyDB



Build your own extension

03

Add prebuilt tasks



Authenticate



Explain bill



Check order status



Make payment

04

More control with Playbooks

Goal

Help customers complete a task (book a trip, make a reservation make payment,, complete an order etc.)

Instructions

Greet the customer in a friendly way

Ask what they like to do

If you can't tell from the previous response, ask

Vertex AI platform is GA

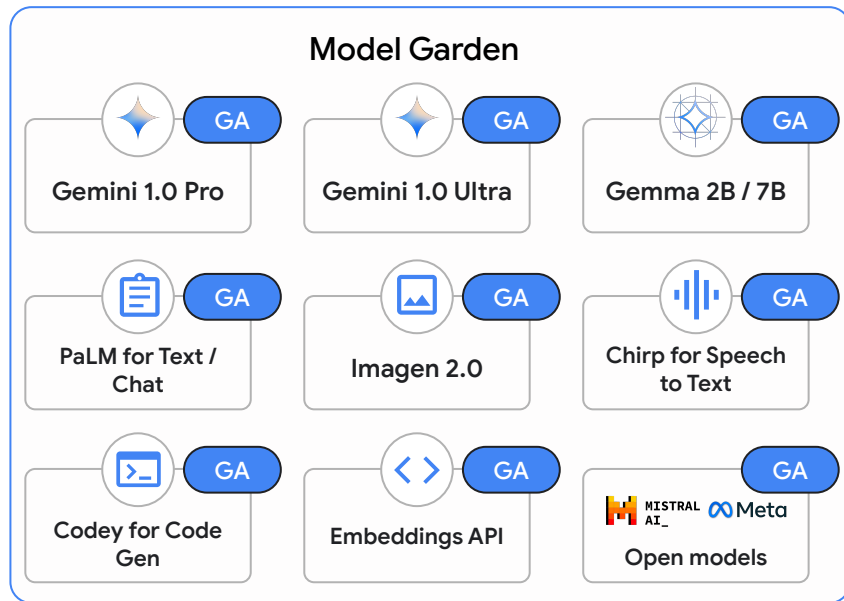
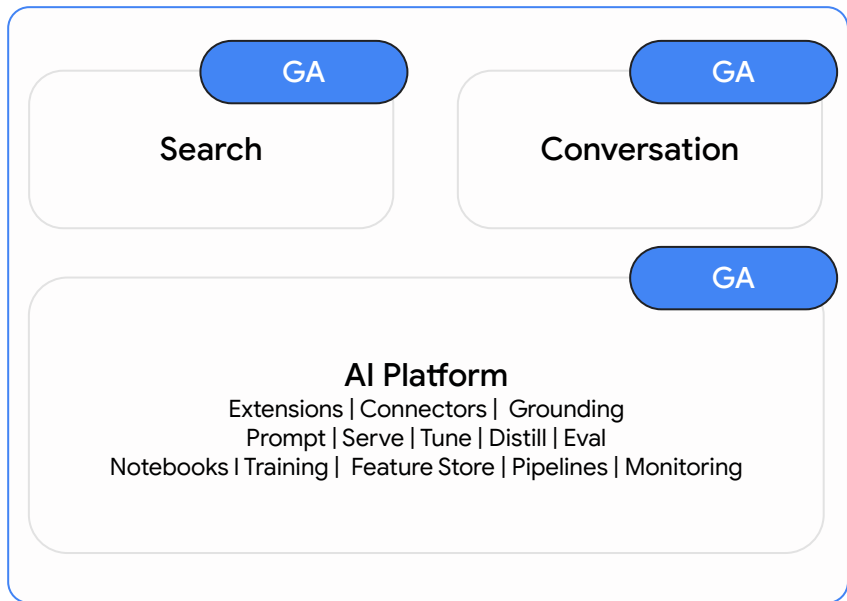
Ready for production



Vertex AI

Momentum towards production: >500%
#API calls growth in H2 2023 vs. H1

Multimodal matters: 3x unique customers
accessing Gemini models vs comparable
PaLM models after Public Preview



Generative AI in Action



Contact Center AI Platform

Google's AI answering Verizon support calls



- Verizon runs the largest contact center in the world
- Bid to deliver a more natural and streamlined Digital Experience (DX)
- CCAI is creating shorter call times and more satisfied customers
- Verizon is able to deal with more customers calling each day

- ~ 85% **containment** achieved with voicebots
- ~ 30% **reduction of time agents on call** with Agent Assist
- 100%+ increase** in the number of live chats agents one can handle with customers at any given time (from three to six)
- Increased accuracy of live chats
- Greater consistency in agent responses

“Verizon's commitment to innovation extends to all aspects of the customer experience. These customer service enhancements, powered by the Verizon collaboration with Google Cloud's CCAI, offer a faster and more personalized digital experience for our customers while empowering our customer support agents to provide a higher level of service.”

Shankar Arumugavelu
Global CIO & SVP, Verizon



Generative AI in Action



Gemini (Duet) AI
for Google Cloud

Assistance powered by Gemini

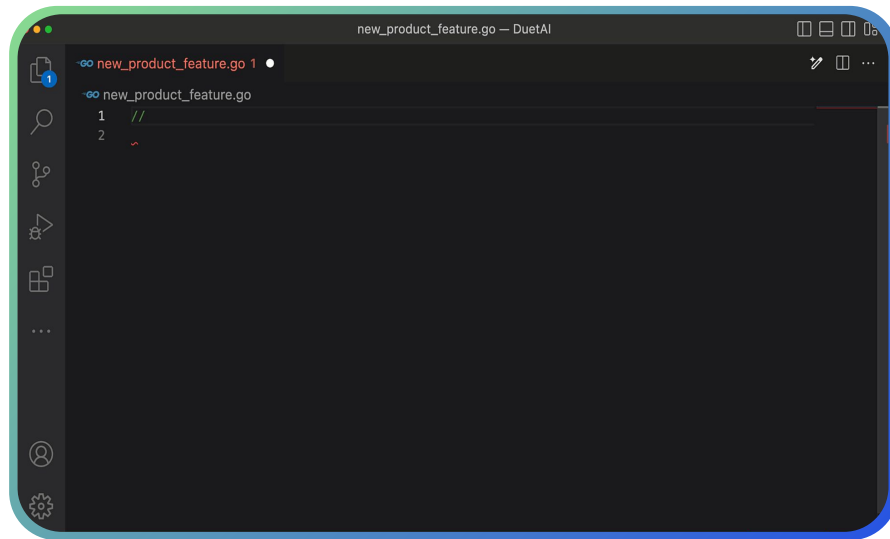
brings intelligence
and efficiency to
Engineering

Gemini (Duet) AI for Developers

Customizable,
complete, secure,
measurable assistance

AI-powered code
completion, generation,
refactoring, and more

Works in popular
languages and tools



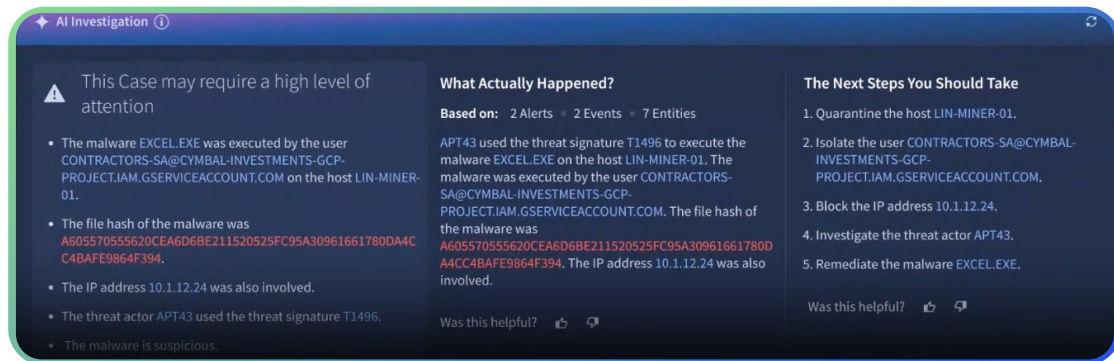
Assistance powered by Gemini

brings intelligence
and efficiency to
Cybersecurity

Gemini (Duet) AI for Security Operations

Combines threat intelligence and security operations assisted by generative AI

Summarizes & prioritizes threat information and automates security workflows



Generative AI in Action



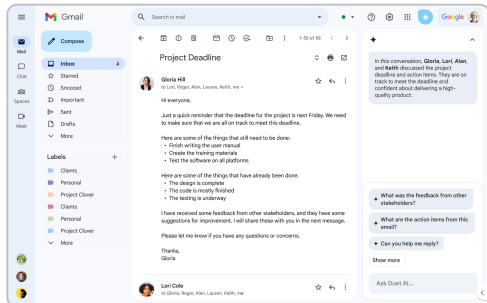
Gemini (Duet) AI
for Workspace

Gemini (Duet) AI for Workspace



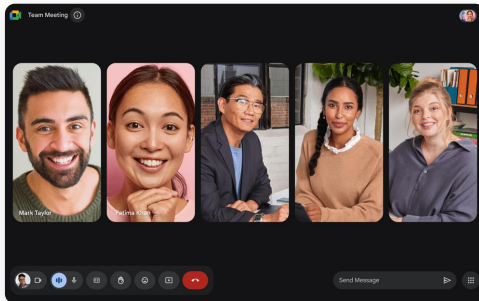
Gmail

Communicate faster with the just the right words to express yourself



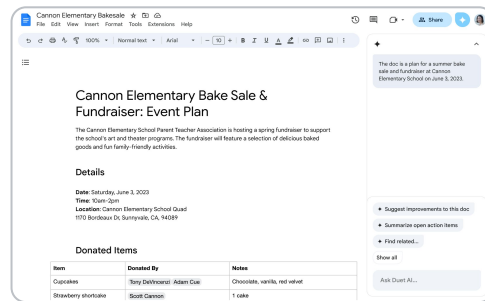
Meet

Expressive and productive conversations for more meaningful connections



Docs

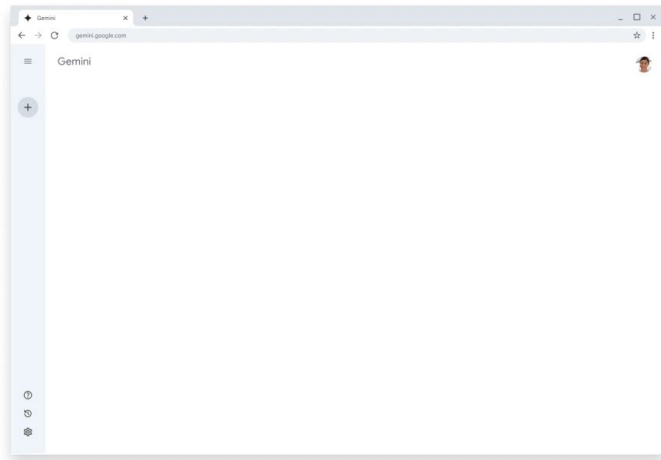
Think bigger, work faster, and supercharge your imagination



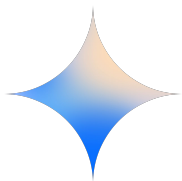


Spark Breakthrough Ideas with Gemini, your AI Thought Partner

- Conversational Exploration for Deeper Understanding
- Fuel Knowledge & Decision-Making
- Easy Workflow Integration
- Enterprise-grade data protections
- One of Google's Most Powerful AI Models

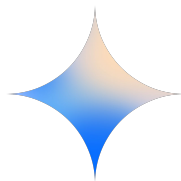


What makes Google Cloud AI the right choice?



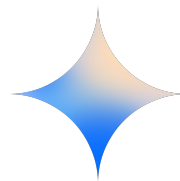
A deep history of research and innovation

Google's generative AI builds on over 20 years of internal R&D across search, databases, hardware, and AI



A deeply integrated, vertically differentiated portfolio

From super scalable AI infra, to world-class models from a variety of sources, to an intuitive AI development and deployment platform, to AI-powered assistants – you will find the right AI for your skills and needs.



Built on a foundation of enterprise readiness

Go to production with the peace of mind provided by data governance and indemnity, security and compliance, and safety and responsible AI.