

Vertex AI

Accelerating generative AI adoption: From PoC to production at scale

October 8th, 2024



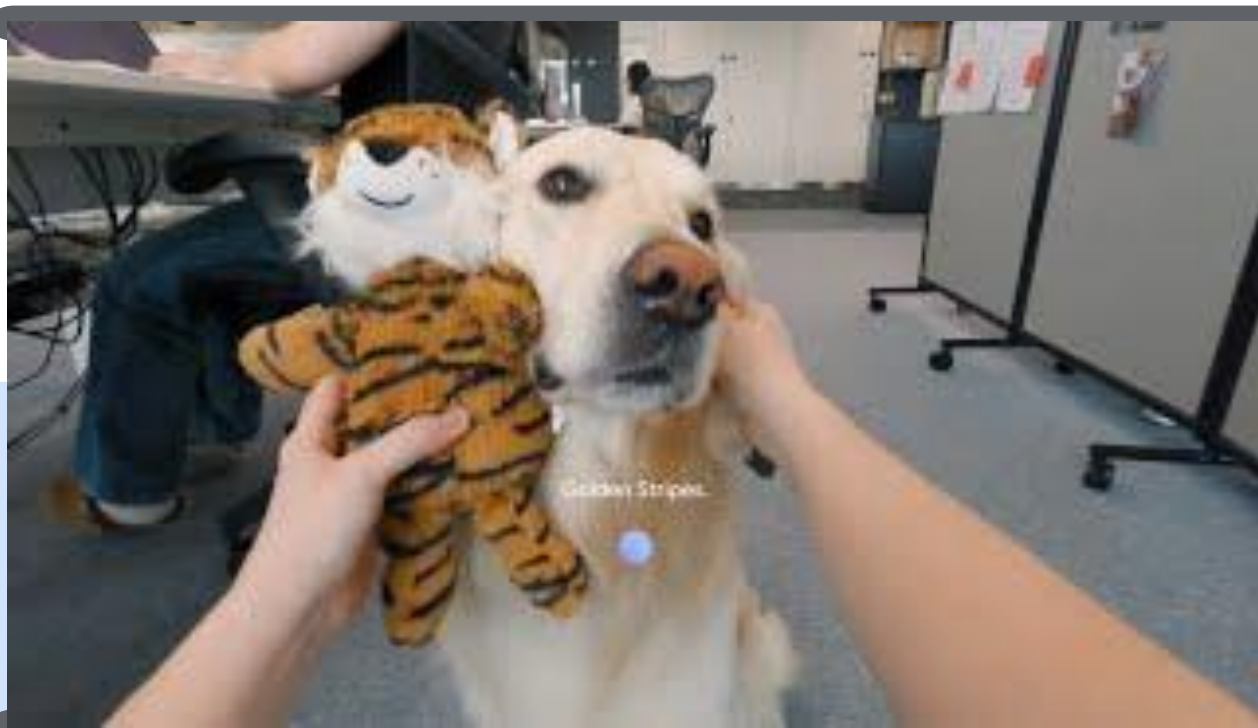
Miku Jha

Director, AI/ML and Generative AI

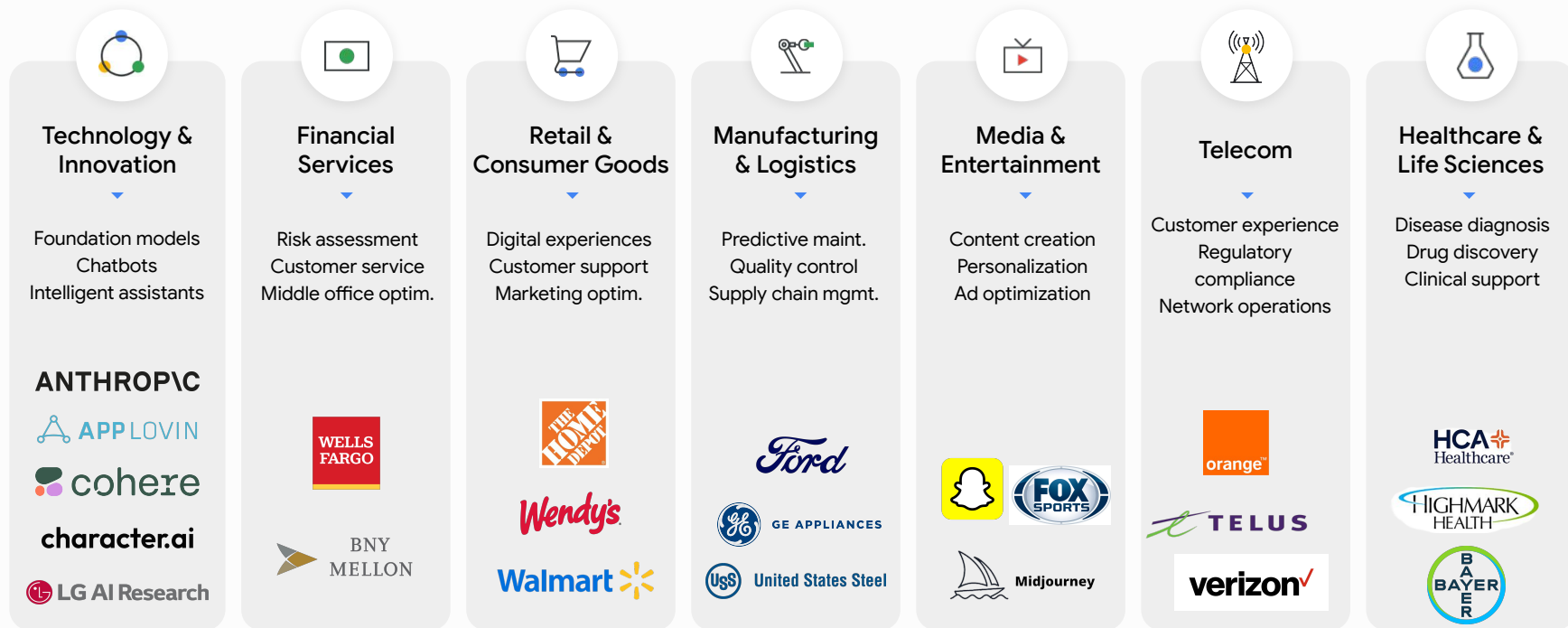


Project Astra: Our vision for the future of AI assistants

Introducing Project Astra. We created a demo in which a tester interacts with a prototype of AI agents supported by our multimodal foundation model, Gemini.



2024: The impact of gen AI has gone beyond a concept




We've spent the past 18 months expanding our enterprise AI platform to meet the needs of generative AI customers





70%

of Generative AI unicorns are
Google Cloud customers



Yahoo Mail Debuts AI Enhancements Built with Google Cloud for a Smarter Inbox

Yahoo Mail uses Google Cloud's Vertex AI platform and AI technology to build several AI features that improve the customer experience. This AI beta customers can opt into, including:

- ✦ **Shopping Saver:** Automatically finds forgotten gift cards, discounts, and store credits in your inbox and helps you use them after making a purchase
- ✦ **Search:** Instead of just keywords, you can now ask questions and use filters to search your emails
- ✦ **Writing Assistant:** Suggests email replies that match the tone of the conversation (e.g., urgent, grateful, apologetic)
- ✦ **Message Summary:** Summarizes key information in your emails, along with suggested tasks, calendar events, and follow-up topics

The image shows the classic Yahoo logo in a vibrant purple color, featuring the word "yahoo!" in a lowercase, sans-serif font with an exclamation point.

Samsung and Google Cloud Join Forces to Bring Generative AI to Samsung Devices

Samsung is deploying Google Cloud's generative AI technology to Samsung smartphone users around the globe. Leveraging various Google GenAI capabilities, Samsung users will gain access to numerous new product features:

- ◆ **Gemini Pro:** Generalize and seamlessly understand, operate across, and combine different types of information, including text, code, images, and video
- ◆ **Imagen 2:** Bring safe and intuitive photo-editing capabilities into the users' hands
- ◆ **Gemini Nano:** Enable on-device LLM delivered as part of the Android 14 operating system, the most efficient model of Gemini for on-device tasks

“

SAMSUNG

Google and Samsung have long shared deeply-held values around the importance of making technology more helpful and accessible for everyone. We're thrilled that the Galaxy S24 series is the first smartphone equipped with Gemini Pro and Imagen 2 on Vertex AI. After months of rigorous testing and competitive evaluation, the Google Cloud and Samsung teams worked together to deliver the best Gemini-powered AI experience on Galaxy."

Janghyun Yoon

Corporate EVP and Head of Software Office of Mobile eXperience Business, Samsung Electronics

WPP and Google forge groundbreaking new collaboration for AI-driven marketing

WPP and Google Cloud announced a collaboration to redefine marketing through the integration of Google Gemini models with WPP Open.

This integration enables enhanced creativity with things like AI generated headlines, smarter content optimization, AI video narration, and hyper-realistic product representation. WPP intends to capitalize on its lead in the space by partnering with AI experts like Google.



Our integration of Gemini 1.5 Pro into WPP Open has significantly accelerated our in AI innovation and enables us to do things we could only dream of a few months ago. With Gemini models, we're not only able to enhance traditional marketing tasks but also able to integrate the end-to-end marketing process for continuous, adaptive optimisation. I believe this will be a game-changer for our clients and the marketing industry at large."

Stephan Pretorius

Chief Technology Officer, WPP

**What have we learned from our
customer success stories?**

Five key challenges that organizations overcome in successfully deploying AI

It's a journey,
not once-off



AI continues to evolve

Implementing AI is an iterative process, with new data sources and opportunities emerging. It's critical your skills grow and develop to meet future needs.

Ensuring security,
compliance & privacy



You must keep control of your data, and protect your IP from leaks

Some AI systems use the data that you input and incorporate it into their models, potentially giving your insights to competitors.

Reducing cost
& complexity



Some AI systems use a mixed stack of components, which can make time to value complex and costly

Integrating varied systems and components directs valuable time away from creativity.

Driving innovation
and efficiency



Often too much focus is placed on potential cost saving without also considering how to unleash creativity and future vision

AI is a strategic asset for your people to deliver the future potential of your organization.

Unlocking disparate
data sources



Data is often distributed across and locked within legacy systems

Accessing data needed to release value can be complex and means that it's hard to extract the strategic value and insights.

The 4 key success factors for enterprise AI



Do you have a single, **integrated platform** that provides your teams **optionality and choice**?



Can you **differentiate with your knowledge and data**?



Does your AI platform **future proof your AI investment** with innovation at every layer?



Is your AI **enterprise ready** so you can go to production with confidence?

Vertex AI is AI for your enterprise

An end-to-end platform that unlocks your data for every use case, expertise, or environment



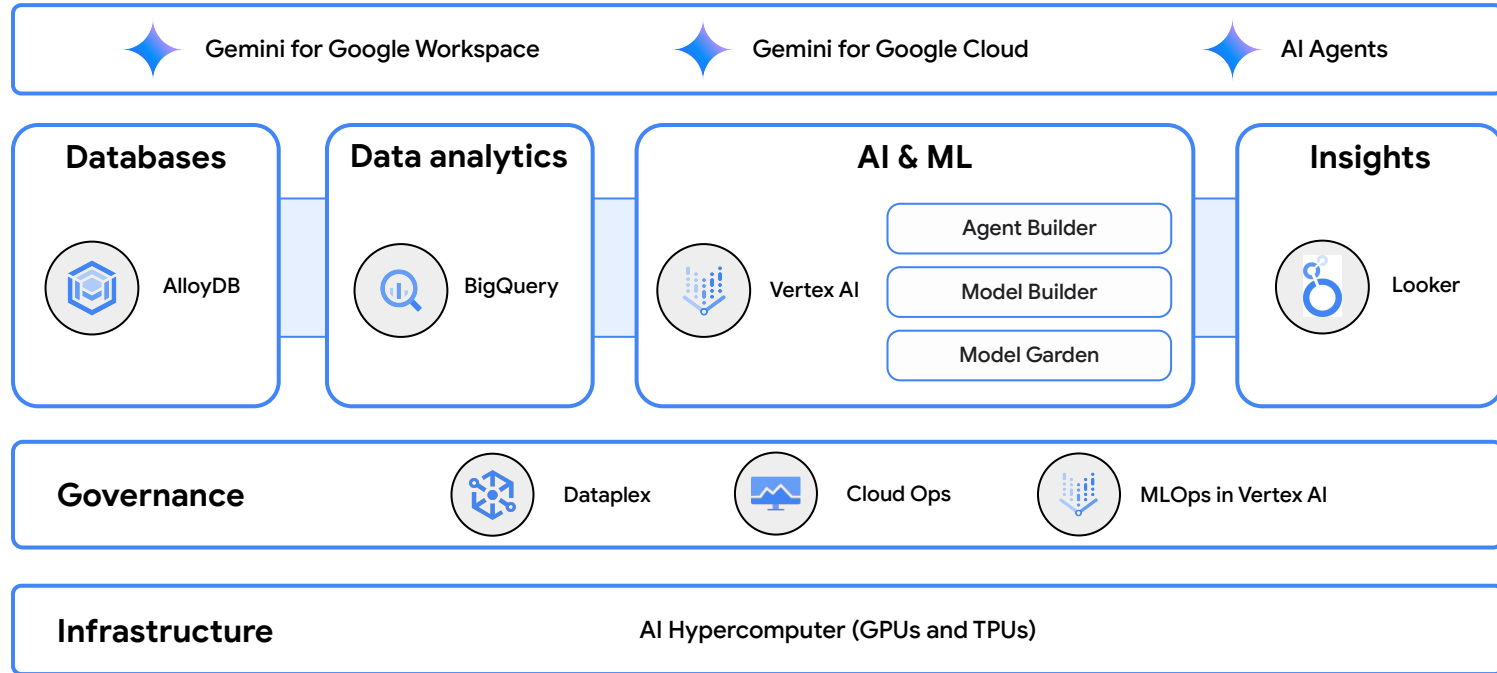
Vertex AI

Agent Builder

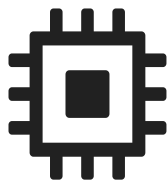
Model Builder

Model Garden

A unified platform from data to deployment and for all your predictive, generative, and agentic needs



Flexibility and curation at every layer of the stack to avoid lock-in



Data

Single unified access layer for all data: structured, unstructured, streaming



Omni for Multi-cloud
(AWS S3, Azure Storage)

Compute

Ultra performant AI hypercomputers for any workload



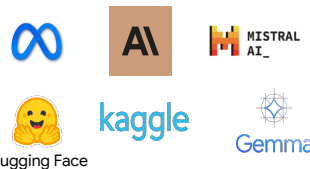
Frameworks

An open & comprehensive AI stack fueling the Gen AI revolution



Models

The best foundation models from Google, Partners, and the Open ecosystem in the Model Garden



Agents

Comprehensive tools from Google and partners to build and deploy agents.





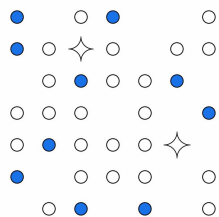
Differentiate with your knowledge and data

✦ Use RAG and other information retrieval capabilities to bring your enterprise knowledge to LLMs

✦ The best AI Search tooling for RAG and Grounding with your 1P enterprise data + 3P data & world knowledge

✦ Training, tuning, and augmentation to customize your data-driven use cases

Training, tuning, and augmentation to **customize your data-driven use cases**



Training

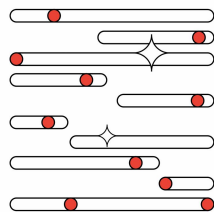
Build your own predictive and generative models from scratch with your own proprietary data

Colab Enterprise Notebooks

Vertex AI Training

Vertex AI Experiments

Vertex AI Prediction



Tuning

Customize foundation models for your specific use cases

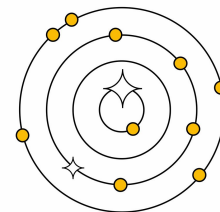
Prompt Design

Supervised Tuning

Reinforcement Learning with Human Feedback

Distilling Step-by-Step

Model Evaluation



Augmentation

Connect and take action on your data and applications

Grounding

Function Calling

Extensions

Connectors

Customize your models from Google, Partners, and Open Ecosystem

How to tune foundation models



Simple, cost efficient

Complex, more expensive

Prompt design

Adapter tuning

Reinforcement
learning with human
feedback

Full fine tuning

Use RAG and other information retrieval capabilities to bring your enterprise knowledge to LLMs



Connect to your data

Web Crawler, Files,
DBs, Connectors

How do I get my data,
from wherever it is,
into the pipeline?



Parsing / Understanding

DocAI

How do I process my data
to extract information
(tables, images etc) and to
make it easier to be found
later



Chunking

How should I segment
while preserving
meaning?



Embedding

Gecko or similar

Which vector
dimensions?
How do I encode
Multi-modal?



Storage

For fast, accurate retrieval

How do I store my data
to be able to retrieve it
later

preparation



Query Expansion & Understanding

Need to spell check?
What about rewording?
Extracting filters & rules,
conversational search



Search & Relevance

Vector Search

Find the most relevant list
based on semantic
search, keywords, filters



Ranking & Optimization

Boost fresh results, rank
based on user behavior,
maximize business KPIs
e.g. clicks / revenue



Answer & Conversation

Gemini or similar

Prompt Engineering,
Tuning,
Citations, Agentic
iterations, Actions



Serving

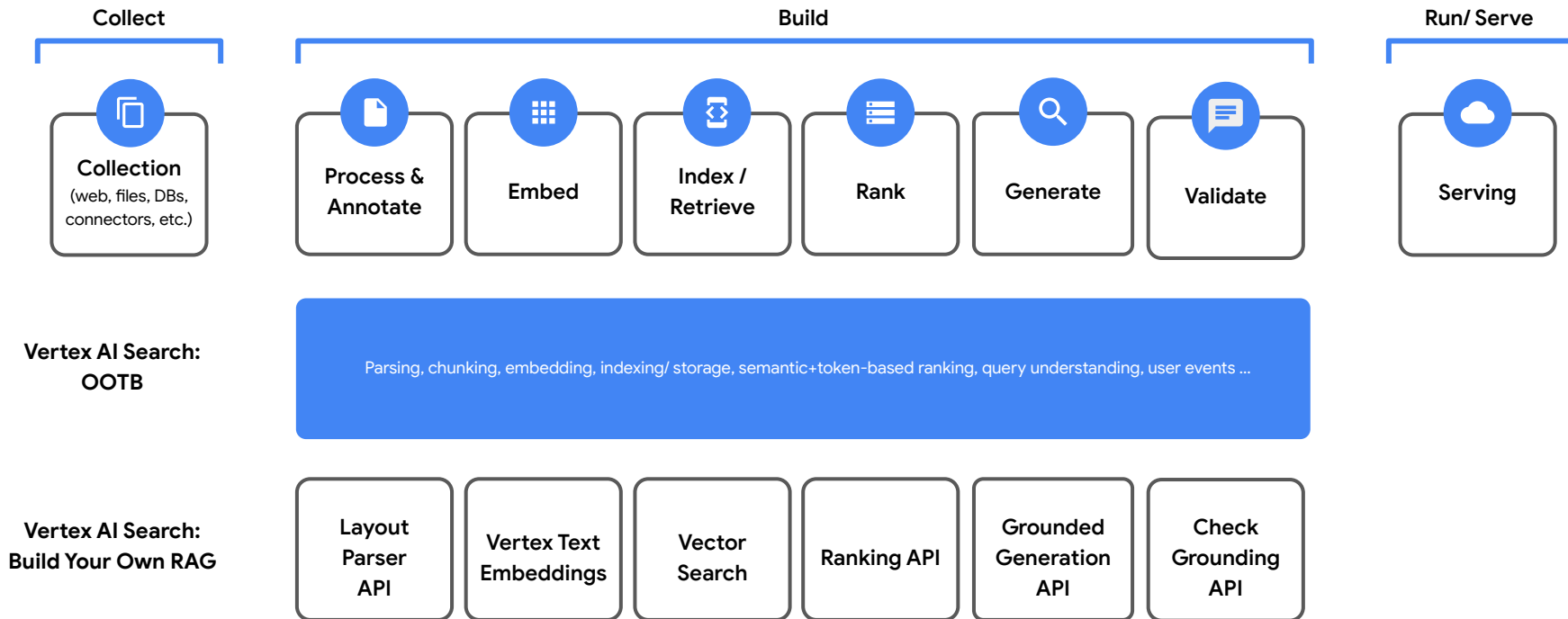
Cloud host

Will my serving API scale
to demand? Is my infra
secure?

runtime

Vertex AI Search - Retrieval Augmented Generation

Prebuilt Managed OOTB + Build Your Own RAG Platform that **bring your enterprise knowledge to LLMs**



Grounding with 1P enterprise, 3P, and world knowledge to find the relevant data

<p>Experimental</p> <p>Grounding with high-fidelity</p> <p>New</p>	<p>GENERALLY AVAILABLE</p> <p>Grounding on your data</p>	<p>GENERALLY AVAILABLE</p> <p>Grounding with Google Search</p>	<p>Coming Next Quarter</p> <p>Grounding with 3P data</p> <p>New</p>	<p>Coming Next Quarter</p> <p>Dynamic retrieval</p> <p>New</p>
<p>Ensures high levels of factuality in response</p>	<p>Ground on private documents and data in Vertex AI Search</p> <p>Provide context to Grounding API directly</p>	<p>Only provider to offer grounding with Google Search (with Gemini)</p>	<p>Currently working with premier providers such as</p> <p>MOODY'S THOMSON REUTERS MSCI</p>	<p>Smartly decide if retrieval is needed</p> <p>Optimizes cost while ensuring factuality</p>



Improve factuality



Access fresh information



Citation answer to its source

What is the 401k contribution limit?

Model without grounding

Output

For 2023, the 401k contribution limit depends on your age:

- Under 50 years old: \$22,500
- 50 years old and older: \$30,000 (includes a \$7,500 catch-up contribution)

It's important to note that these limits apply to employee contributions only. Employer contributions, such as matching contributions, are not included in these limits. However, there is also a combined employee and employer contribution limit of \$66,000 for 2023 (\$73,500 for individuals 50 and older).

Gemini + Google Search

Output

For 2024, the 401(k) contribution limit for employees under 50 years old is \$23,000. If you are 50 or older, you can contribute an additional \$7,500 in catch-up contributions [1], bringing your total potential contribution to \$30,500.

It's important to note that these limits apply only to employee contributions. Employer contributions, such as matching contributions, are not included in these limits. However, there is a combined employee and employer contribution limit of \$69,000 for 2024 (\$76,500 with catch-up contributions).

Remember that contribution limits can change periodically, so it's always a good idea to check with the IRS or a financial advisor for the most up-to-date information.

Grounding Sources

1. [401\(k\) Contribution Limits for 2023 vs. 2024 - Investopedia](#)



Future proof your AI investment

✦ Rich history of open-source and proprietary innovation across search, hardware, data, algorithms, and models

✦ Best infrastructure for running large scale workloads

✦ Continued improvement in features, models, and tooling to optimize performance and cost

✦ Best platform TCO for your AI use cases

What makes Google Cloud AI the right choice?

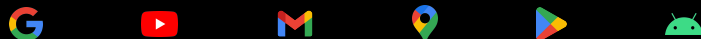
World leading models

Gemini Gemma

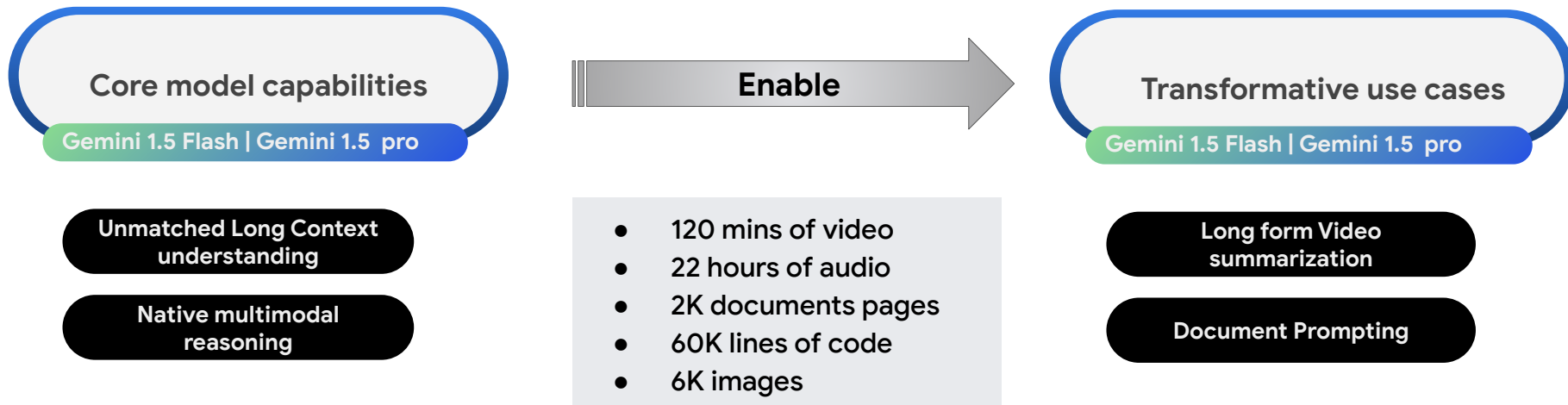
World leading infrastructure

TPU GPU

The knowledge and experience to integrate ML into services delighting billions of users across the world



Industry leading gen AI



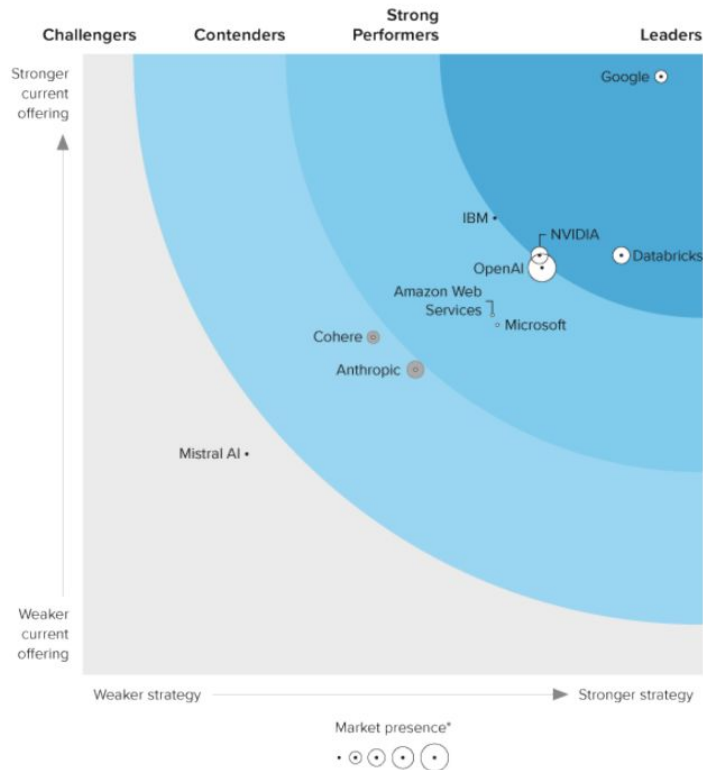
World leading models
Gemini Gemma

World leading infrastructure
TPU GPU



Google is a Leader in The Forrester Wave™: AI Foundation Models for Language, Q2 2024

The Forrester Wave™ is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave™ are trademarks of Forrester Research, Inc. The Forrester Wave™ is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings, and comments. Forrester does not endorse any vendor, product, or service depicted in the Forrester Wave™. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.



*A gray bubble or open dot indicates a nonparticipating vendor.

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.



Vertex AI

Enterprise ready for production



Data governance, security, and privacy to protect your data



Indemnity for generative AI training data and generated outputs



Data residency and ML processing for your global regulatory needs



RAI tooling to make every model safe for your use cases

Continued serving improvements to optimize performance and cost

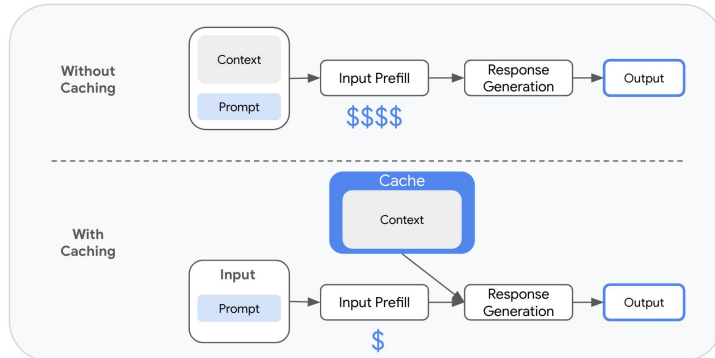
Context Caching (Preview)

Only provider to offer context caching API

75% Lower input price
with context caching*

Take advantage of millions-of-tokens context windows,
Available across both 1.5 Pro (June 27th) and 1.5 Flash (July 2nd)

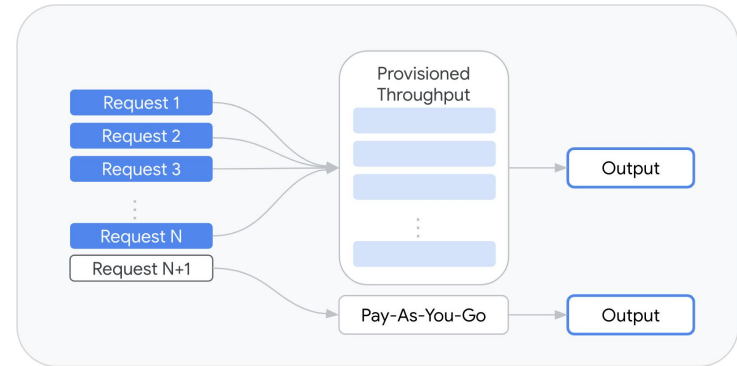
*with >=32K context window



Provisioned Throughput (GA)

Bring predictability and reliability to customer production workloads

Giving customers the **assurance required to scale** gen AI workloads aggressively



Indemnity for generative AI training data and generated outputs

Indemnity is a complex topic, but to put it plainly: if customers are challenged on copyright grounds, we will [assume responsibility](#) for the potential legal risks involved when our services are used in a [responsible way](#)

Industry-first, two-pronged indemnity approach

Training data indemnity

- Covers Google's use of training data to create Google models utilized by all our Generative AI services
- Has always been implied; we're just providing an explicit, public clarification

Generated output indemnity

- Covers the generated output created by our customers
- Applies to Gemini in Google Workspace and a [selection](#) of Google Cloud services

The best infrastructure for running large scale workloads

Tensor Processing Unit: Designed by Google for AI at Scale



Cloud TPU v5e
up to 2.7x Inference Perf/\$ vs v4

Cloud TPU v5p
up to 2.8x LLM training vs v4

Cost Efficient & Versatile
(Training & Inference)

Powerful & Flexible
(full range of AI models)

NVIDIA GPUs: Latest NVIDIA GPUs on Google Cloud



G2 GPU VM
2-4x performance improvement vs T4

Powered by
NVIDIA L4 GPU

A3 Mega GPU VM
3x training improvement vs A2 VM

Powered by
NVIDIA H100 GPU

- Provide a **wide variety of hardware** options
- **Speed up training & inference** time with high-performance computing
- Improve **price-performance** & cost
- **Scale** AI models exponentially
- Leverage our fully-managed AI platform optimized for **efficiency**
- Build with an **open source software ecosystem**

How Pendo Leverages AI



In-app Guides

Deliver contextual communication to web & mobile users

Drive Awareness & Action

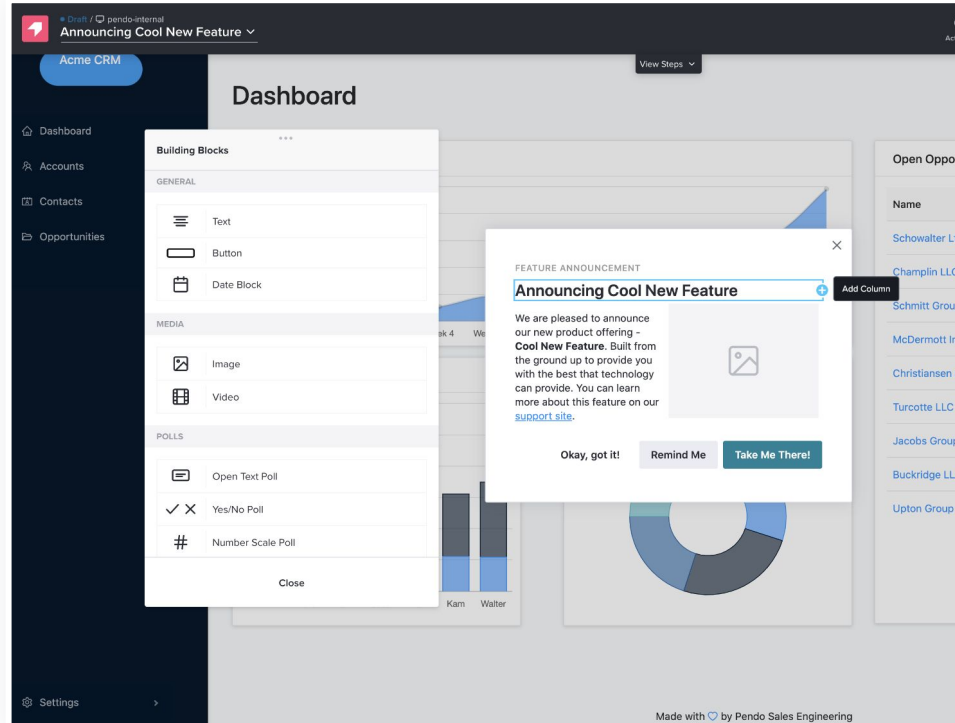
Onboarding, adoption, in-app support, growth, & more.

No-Code Visual Design Studio

Build beautiful guides with layouts library & visual interface.

Data Driven & Personalized

Create helpful, personalized guides with Segments, Conversions, & Experiments.



Why leverage AI?

**Time to first guide
publish**
Baseline: 10 - 20 days

**Q2 = 20.6
days**

**Q3 = 14.3
days**

**Q4 = 18.7
days**

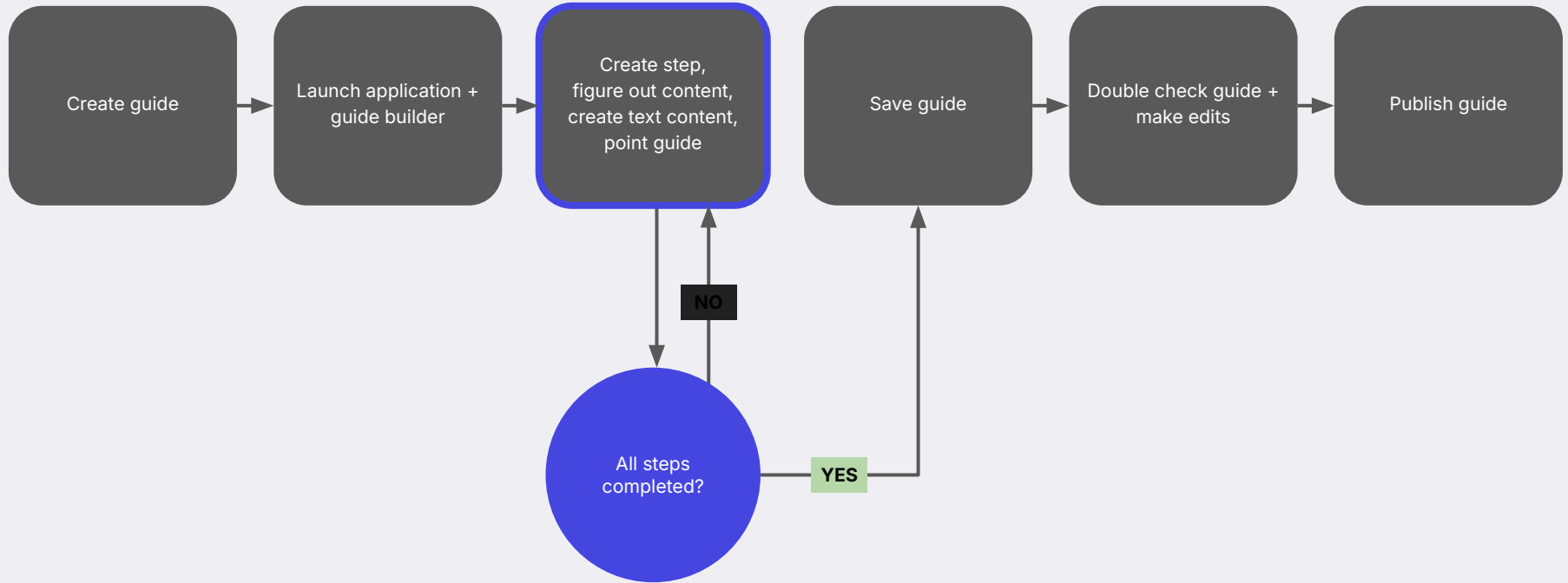
- On average, it takes ~18 days from install of Pendo, to publishing first Guide
- Customers who use Guides show high retention, and consistently provide higher NPS scores

The screenshot shows a financial dashboard with a 'New Premium Feature' overlay. The dashboard includes sections for 'CREDIT CARDS', 'LOANS', and 'CREDIT SCORE'. The 'CREDIT SCORE' section shows a score of 776 for June 30, 2020. The 'LOANS' section lists a Mortgage Loan for \$2,286.80 and an Auto Loan for \$649. The 'CREDIT CARDS' section lists a Personal Credit Card for \$5. The 'CREDIT SCORE' section shows a table of payments:

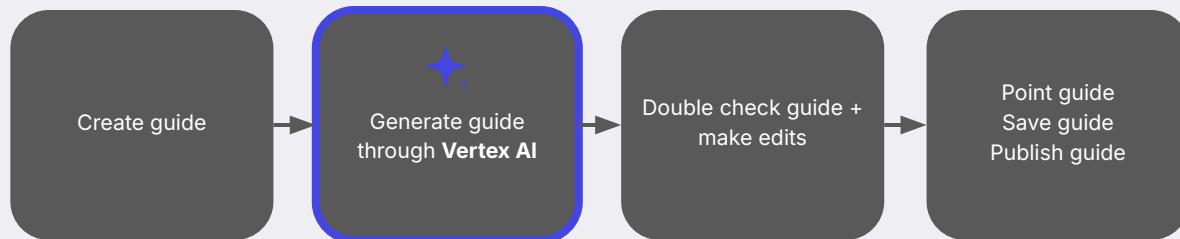
Amount	Balance
\$2,286.80	\$544,846.38
\$2,286.80	\$545,375.37
\$2,286.80	\$546,953.93
\$2,286.80	\$547,292.33
\$2,286.80	\$548,288.25

The overlay features three stars, the text 'New Premium Feature', and 'Upgrade your account for unlimited access'. It has two buttons: 'No thanks' and 'Upgrade now'.

The process today



With Pendo + Google AI



Google Cloud

Thank You

