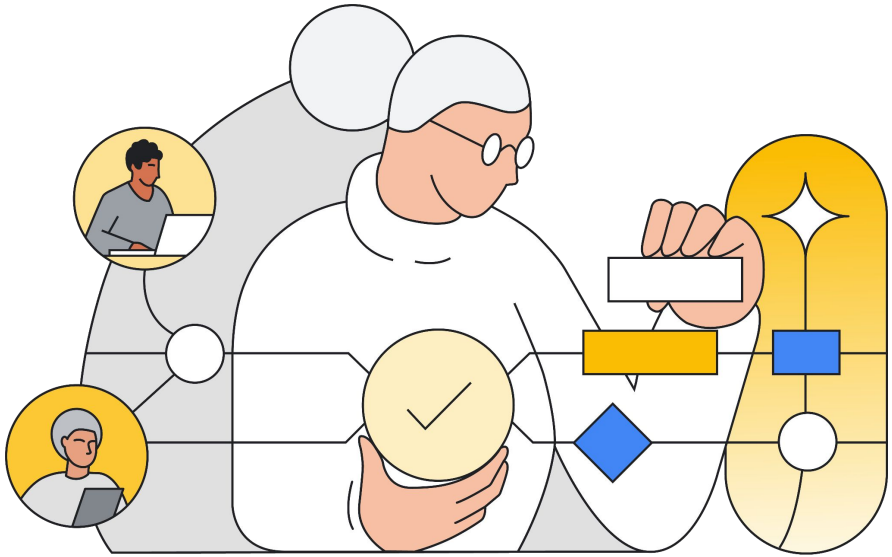
Google Cloud

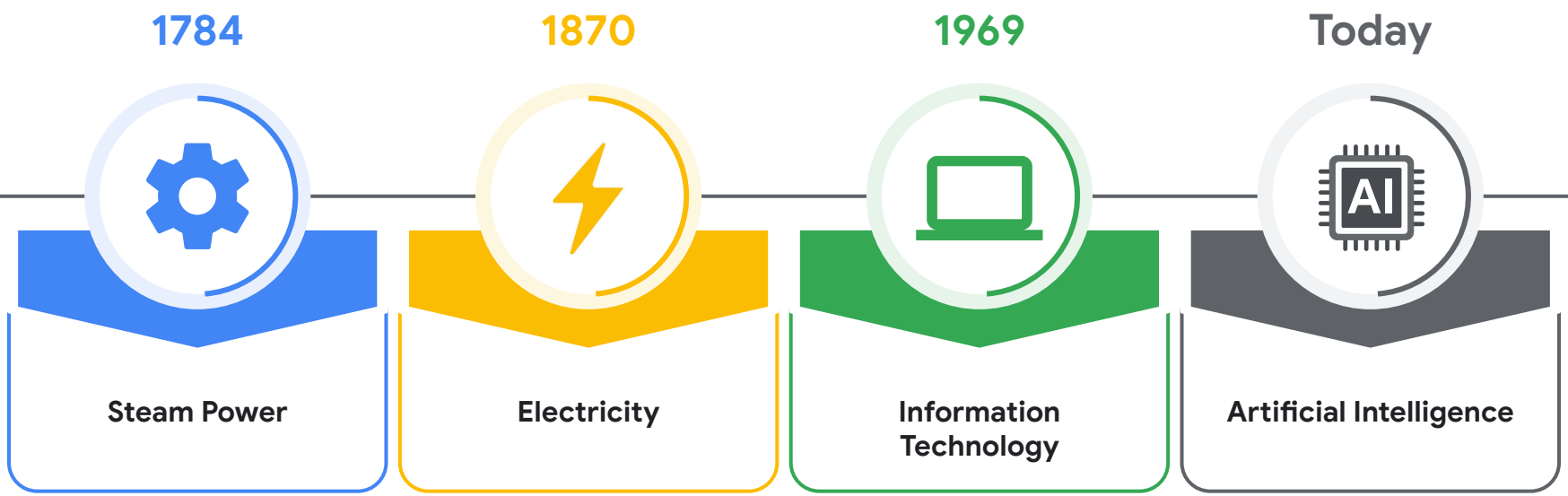
Supercharge your Applications with Google's Duet AI

Google Developer Groups, Cloud, Chennai
February 16, 2024

Kiran Vasudevan, Customer Engineer
Application Modernization



We're in an AI-driven revolution

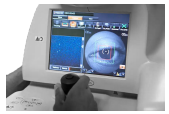


Source: AI: Recent Trends and Applications, Emerging Communication and Computing

AI Journey @ Google

2015

Google DeepMind AlphaGo defeats Go champion



2016

Google's DeepMind helps detect eye disease

2017

Google invents Transformer kickstarting LLM revolution

2018

Google's groundbreaking large language model, BERT

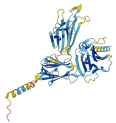
T5

2019

Text-to-Text Transfer Transformer LLM 10B P Model Open Sourced

2020

Google LaMDA Model Trained to converse



2022

AlphaFold predicts structures of all known proteins

Bard

2023

A conversational AI Service powered by LaMDA.

3,000 Researchers
7,000 Publications

Responsible AI

Built & Tested for Safety

Privacy in design

Upholds high scientific standards

Accountable to People

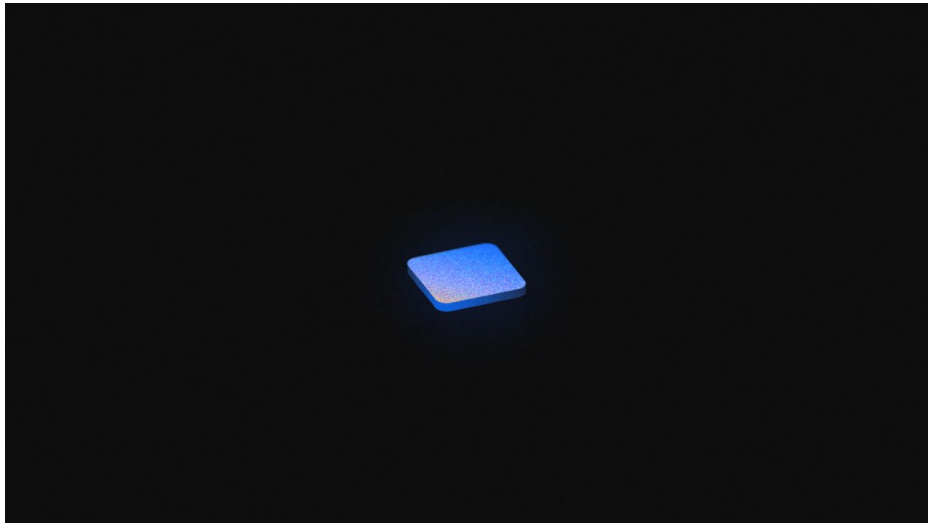
Socially Beneficial

Avoid creating unfair bias

The next chapter of Generative AI innovation

Gemini

Gemini is the most capable and general model we've ever built, and is the result of a large-scale collaborative effort by teams across Google, including Google DeepMind and Google Research.



AI Collective at Google

Industry-defining AI R&D

- Thousands of Researchers
- 7K+ Publications
- 2.6B+ Citations
- Transformer OCS Network TPU

Cutting-edge Foundation Models

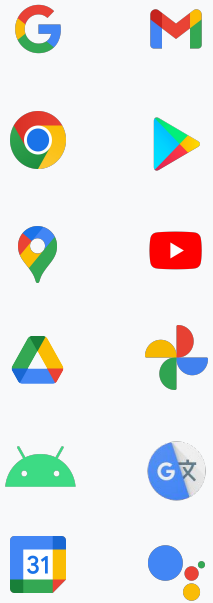
- Gemini
- PaLM2
- Imagen
- LaMDA
- GLaM
- BERT
- Chirp

Robust & Open SW Ecosystem

- TensorFlow
- PyTorch
- JAX
- XLA ACCELERATED LINEAR ALGEBRA
- kubernetes
- #1 OSS Contributor

Source: OSS Index

Planetary-scale App Expertise



World-class Infrastructure

- Purpose-built Compute
- Storage
- Network
- Container
- Sustainability

Supporting developers with Google's Generative AI

Developer Facing Applications

Duet AI

Workspace

Google Cloud Platform

3rd Party Integrations

replit

GitLab

In House API Integrations

Vertex Codey

Android Studio

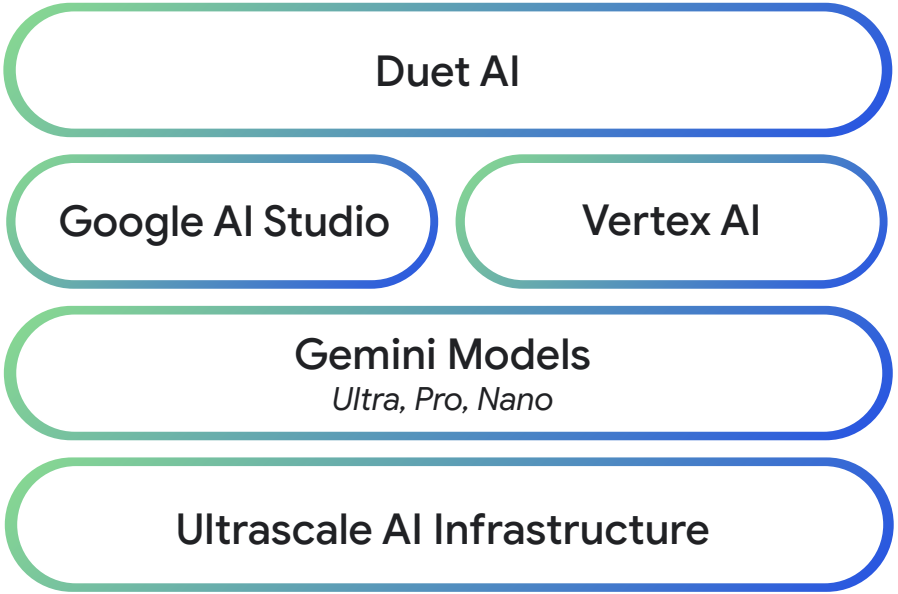
colab

Gemini

Code Intelligence Model research

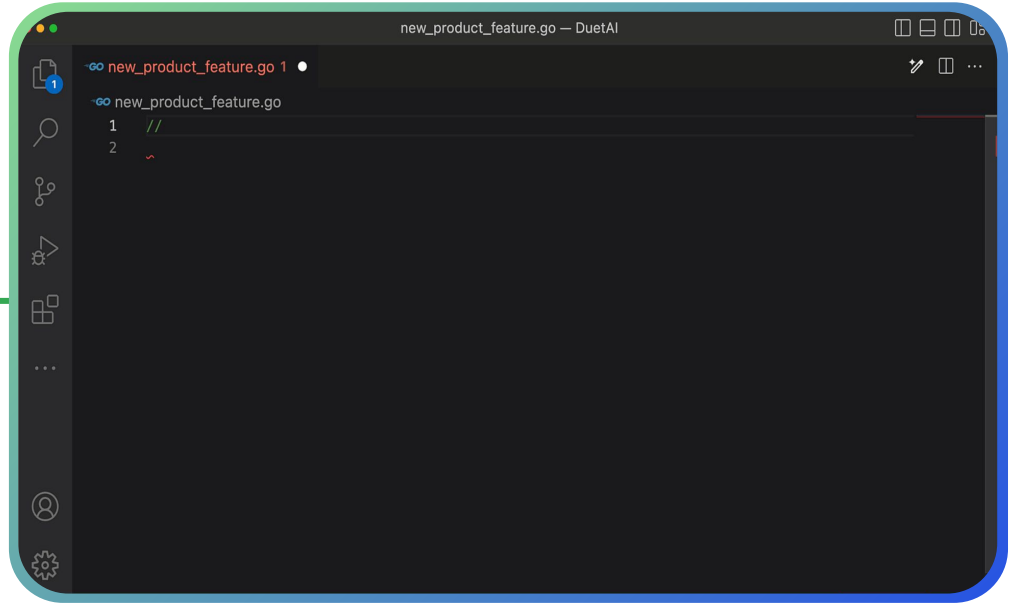
AI Research and Foundation Models (Palm, Gemini)

Generative AI is built on a vertically integrated technology stack

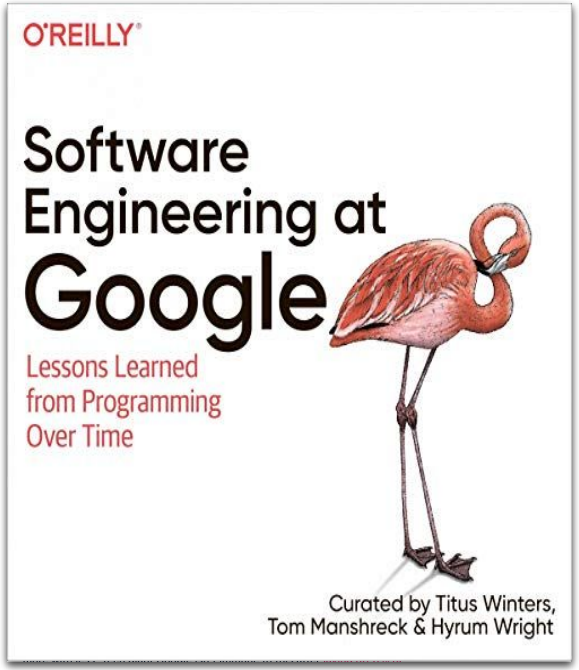


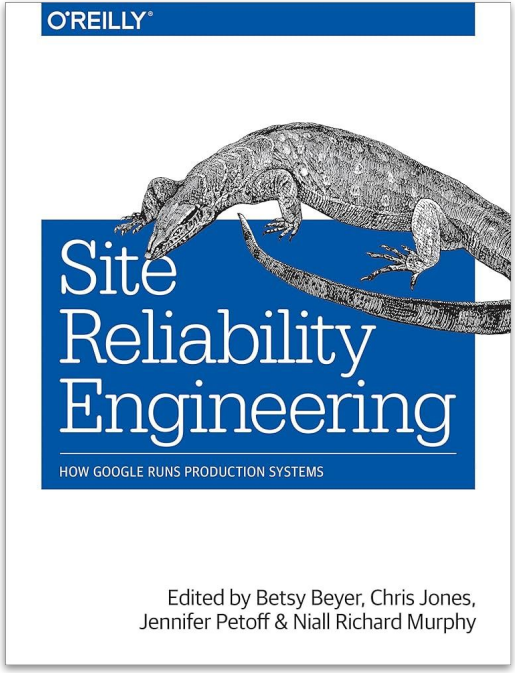
Duet AI for Developers provides AI-powered:

- Code completion
- Code generation
- Code refactoring
- Documentation generation
- Unit test generation
- Application deployment
- Operational diagnostics



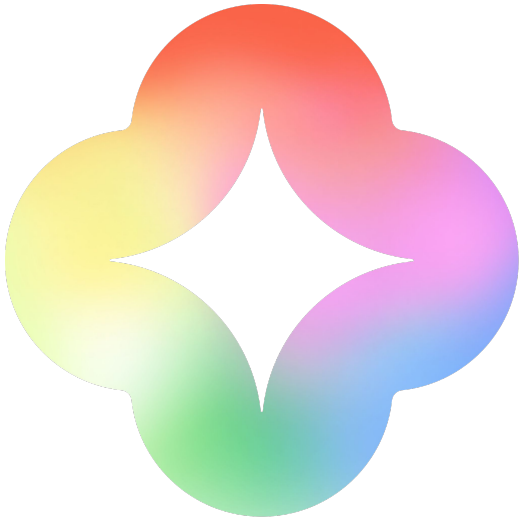
Software delivery is a key part of Google's DNA
Duet AI brings our internal learnings to everyone



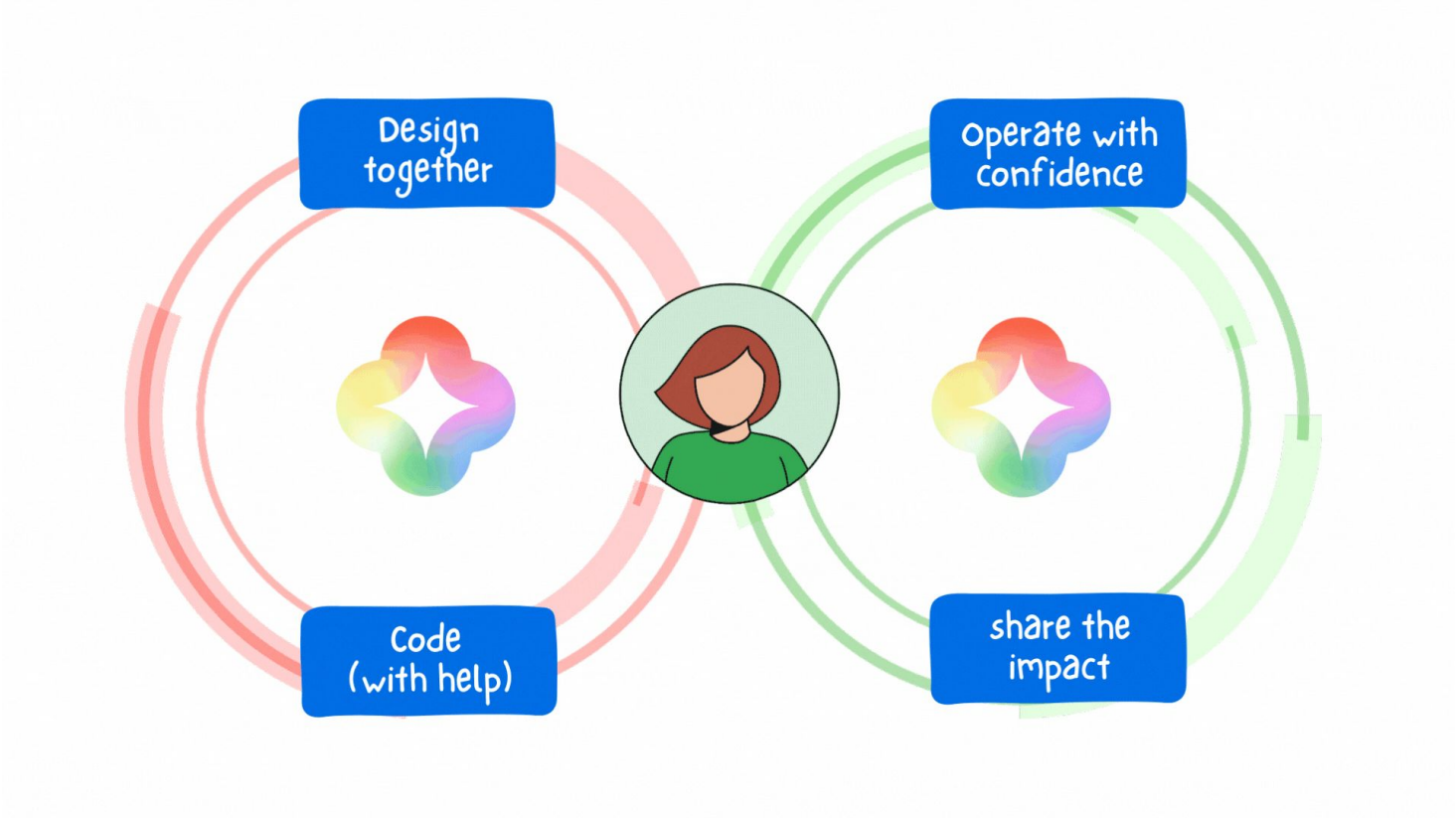


Google uses Duet AI

The first customer to enable Duet AI for day-to-day development. Google internal usage **improved** the capabilities making Duet enterprise-ready:



Hands-On with Duet AI

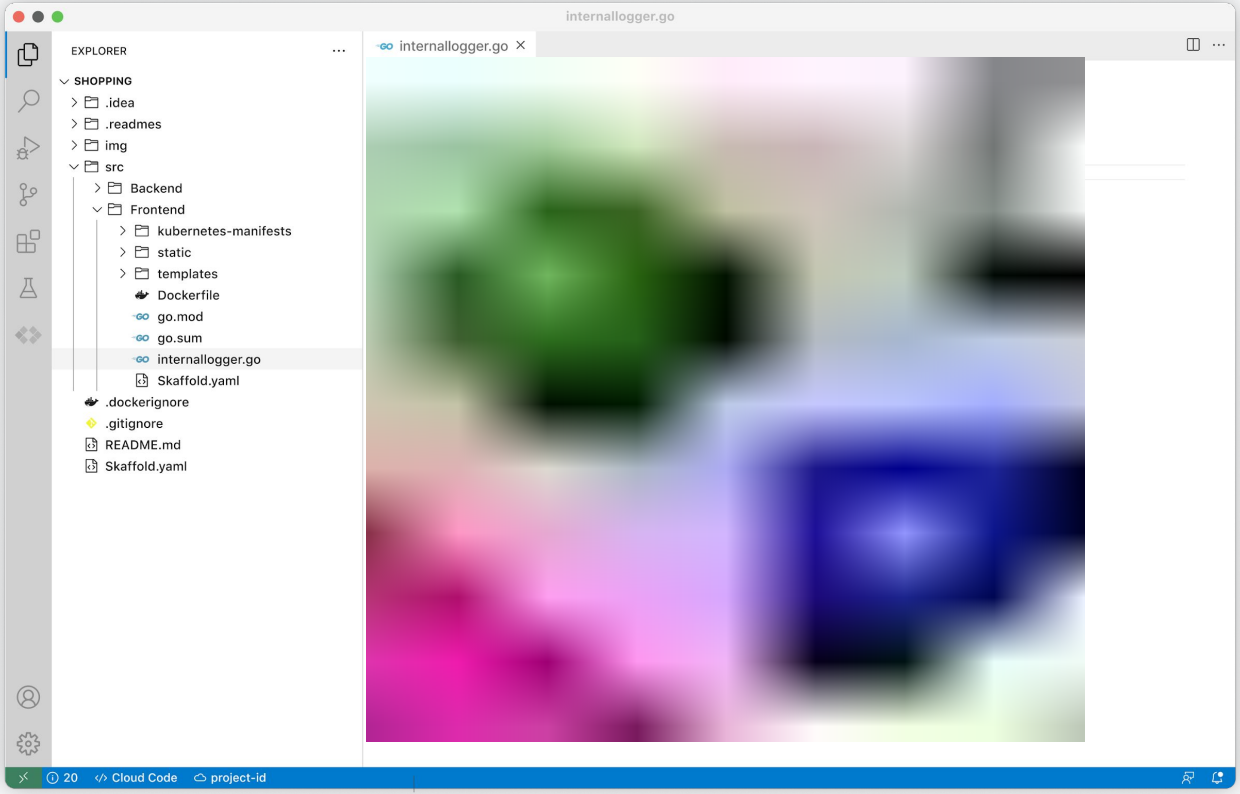


Journey 1: AI Code completion

Assistive Development

Devs can **write high quality code faster** and stay in a **flow** state longer so to can be productive translating their idea into code.

Real time code completion/generation and autofix suggestions within IDE.



IDE Coverage

2023 Target

Stretch goal

Journey 2: License attribution assistance

Assistive Development

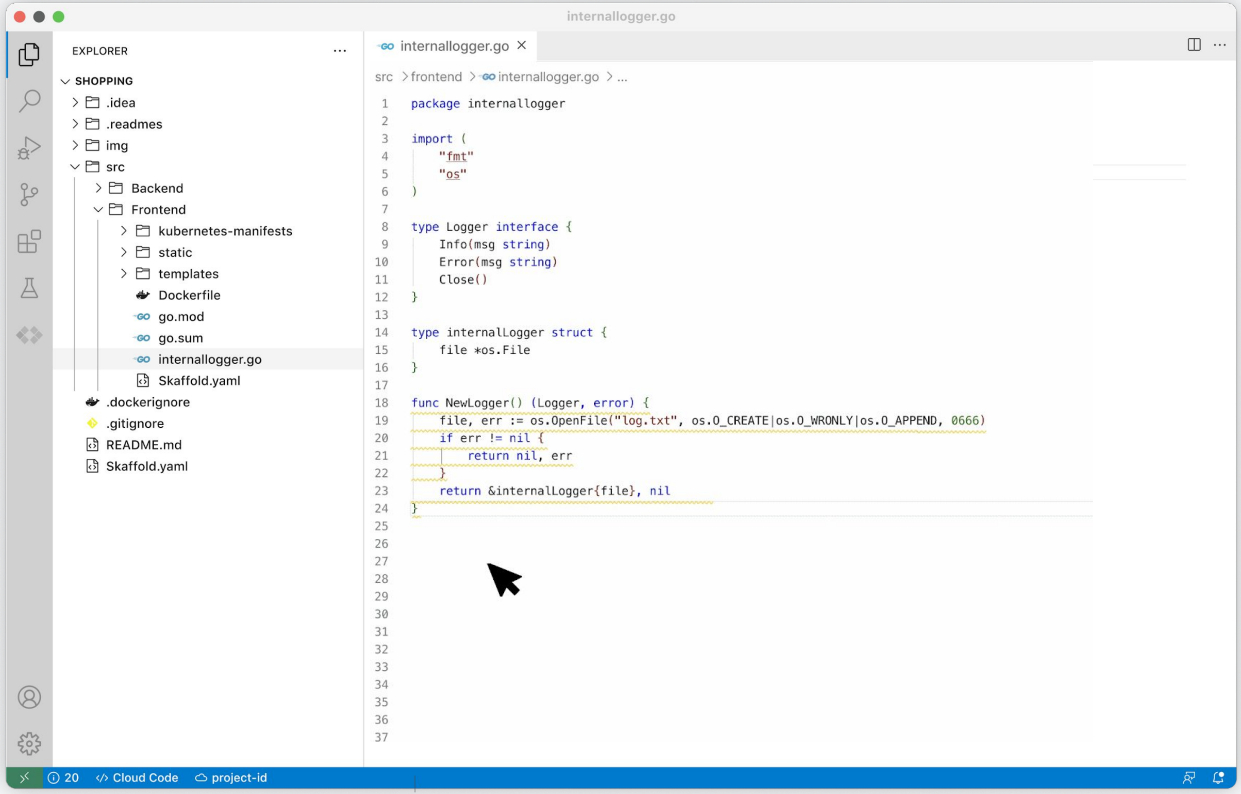
Devs are made aware when a suggestion falls **under a permissive license** and receive the information to provide license attribution, so their organization can **stay compliant with license terms**

Real time code license flagging via recitation checking.

IDE Coverage

2023 Target

Stretch goal



Journey 3: Explain code (Chat)

Assistive Development

To improve test coverage for a code base a dev is not familiar with, they can I ask Duet AI to **explain the code** so to quickly and effectively understand how to start testing it.

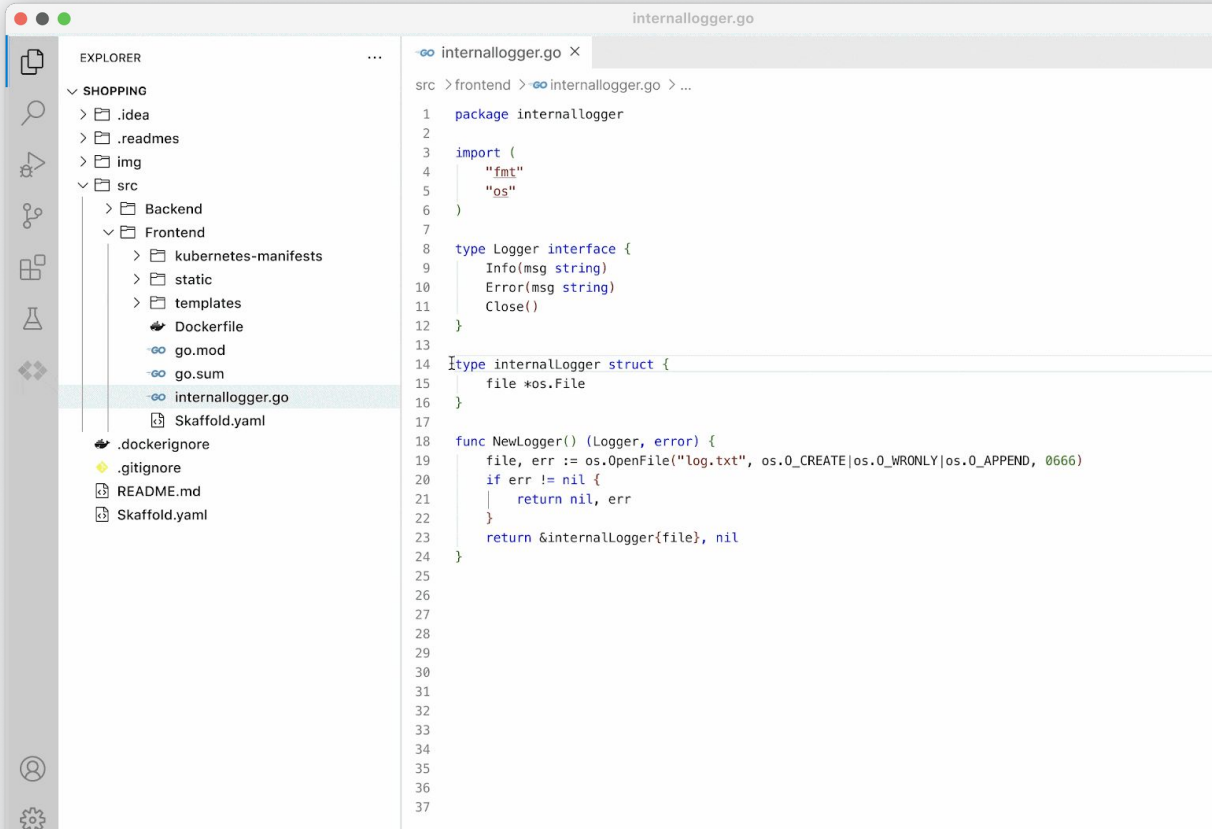
Key moment

Duet AI explains selected code in natural language and allowing users to ask follow up questions.

IDE Coverage

2023 Target

Stretch goal



Journey 4.a: Test generation - Suggest a test plan (Chat)

Assistive Development

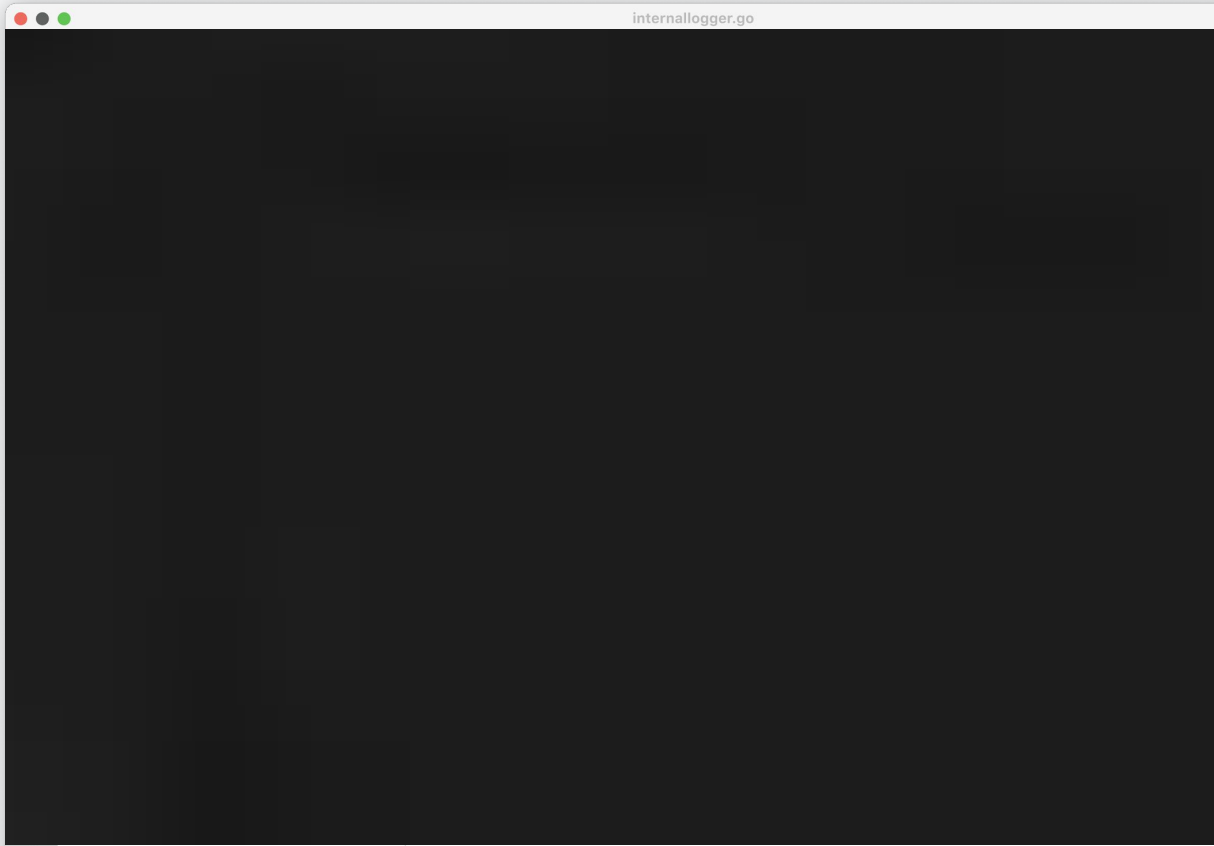
To define a test plan to ensure the code works as intended. Devs can ask Duet AI to **suggest a unit test plan** for the code, and iterate on it to ensure good coverage of edge cases.

Key moment
Duet AI suggest the test plan and elaborate on the edge cases

IDE Coverage

2023 Target

Stretch goal



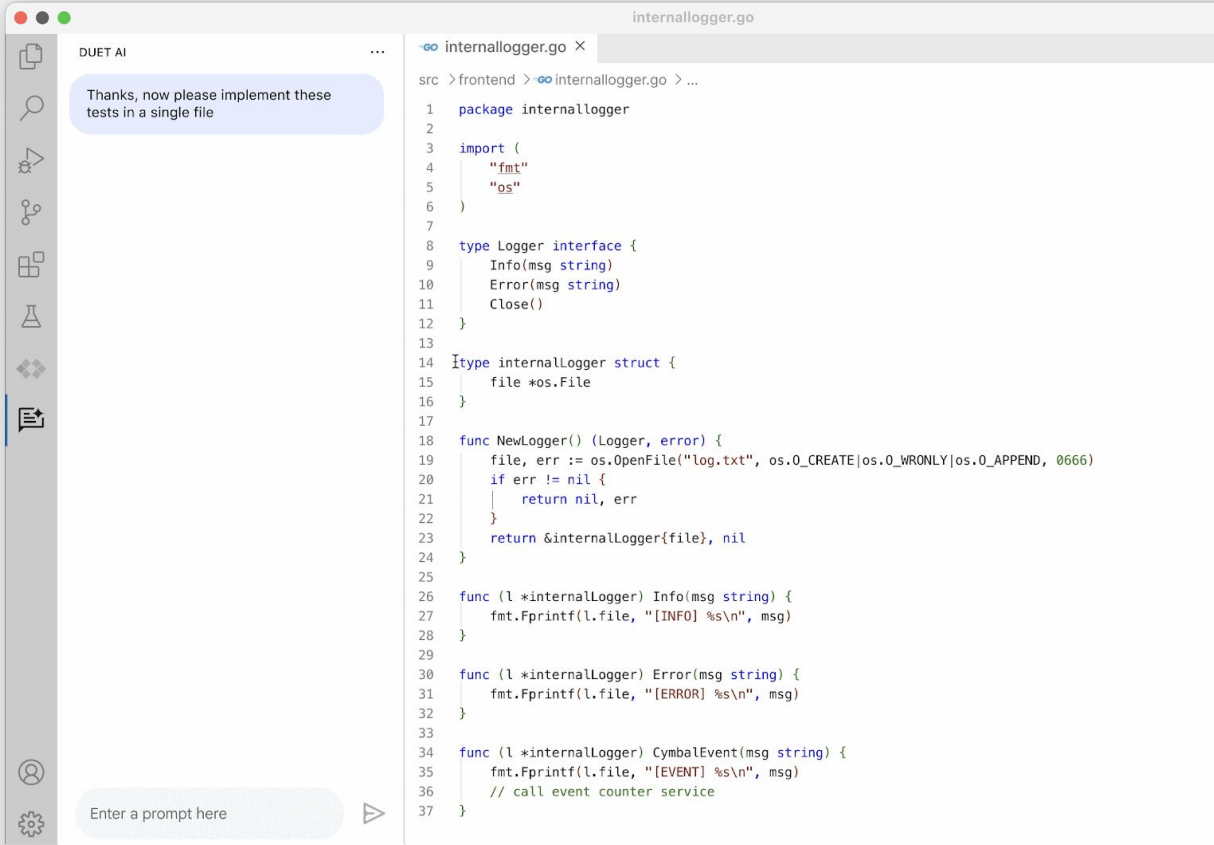
Journey 4.b: Test generation (Chat)

Assistive Development

Once the test plan looks good, Devs can ask Duet AI to **implement the test plan defined**

Key moment
Duet AI generate application specific unit tests that is ready for use

IDE Coverage
2023 Target
Stretch goal

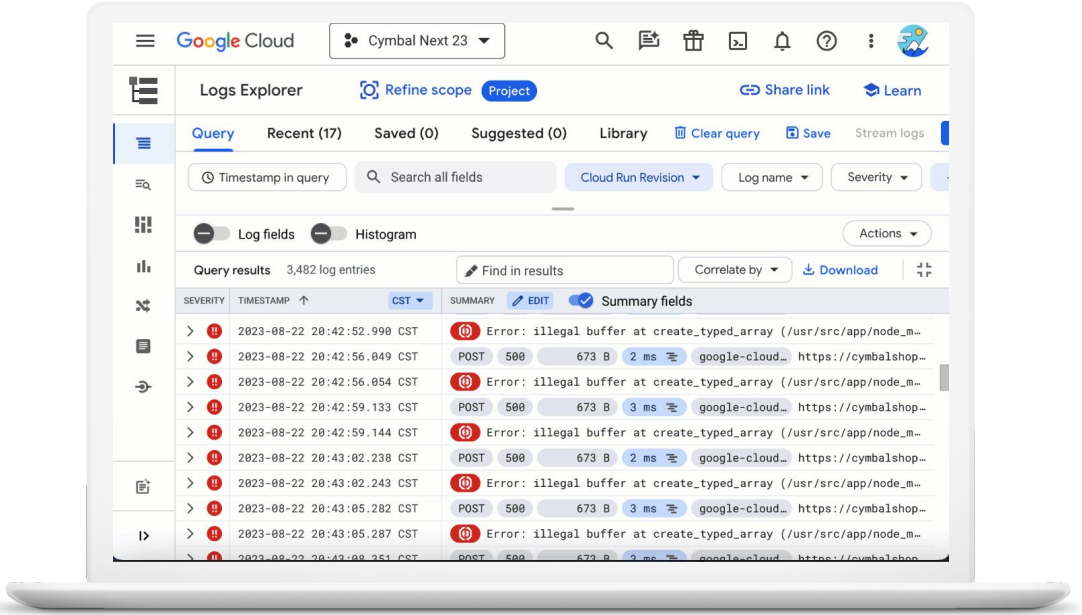


Log analysis using Duet AI

Duet AI helps get from **error to root cause analysis faster**

Duet helps understand the errors in more detail

With Duet AI you can use natural language to describe how to modify the search query

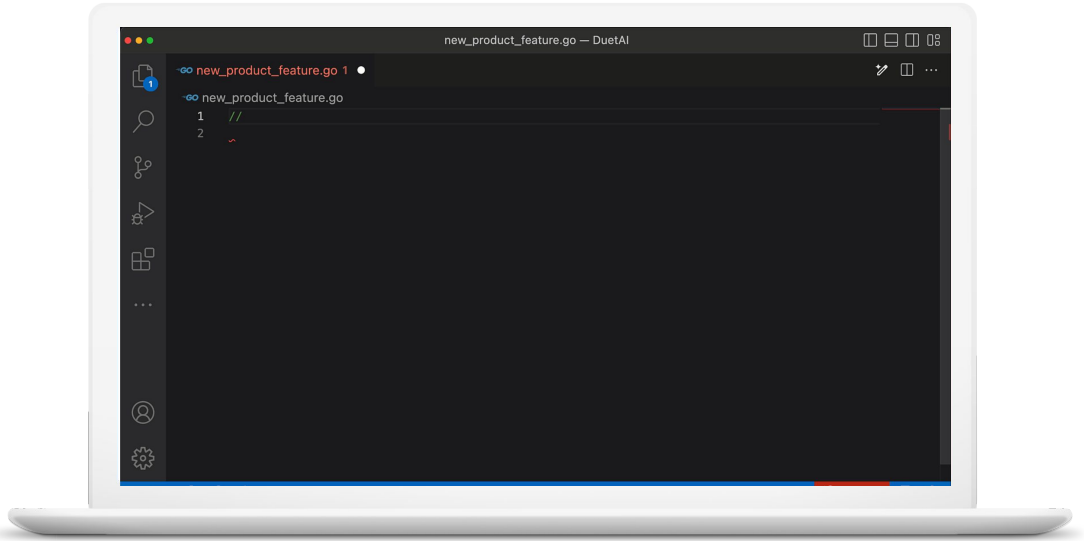


Context Aware Code Generation

Duet AI can be **customized with enterprise knowledge** based on specific company SDKs and code base to generate context-aware code.

Typically, writing such code would require very specialized knowledge.

Duet AI can save time, resourcing, training hours–all without compromising quality.

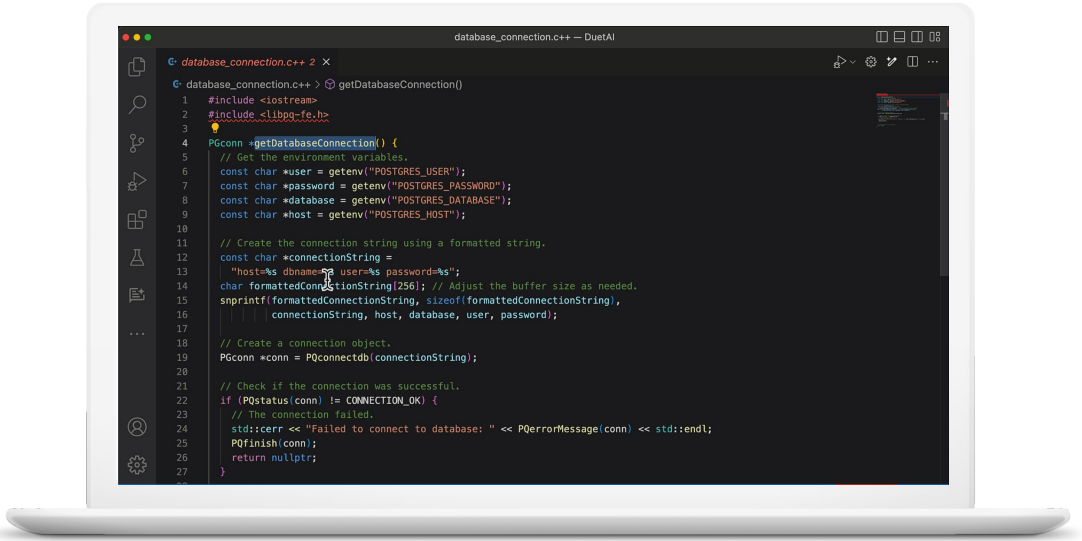


Code Translation and Application Migration

Powerful developer features into Duet AI to push the boundary of the AI code assistance

Modernize applications faster by assisting with code refactoring

Migrate database connections to google cloud managed databases



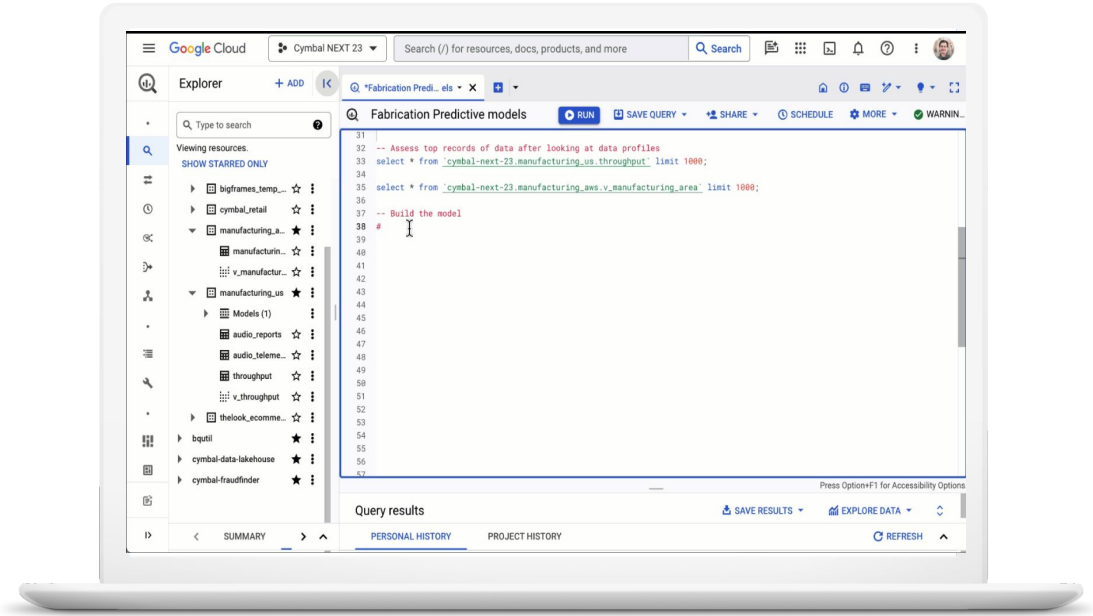
Duet AI responds to a natural language prompt through chat interface and converts a function to Go use SQL

Duet AI in BigQuery

Assisted coding within the data suite of products with a chat experience to increase efficiency and works with open source formats like PySpark, Beam, and Airflow

Duet AI to help write a query to connect data across clouds in SQL via natural language.

Extracts features which help create ML model. Duet AI helps **write a query to build ML model**

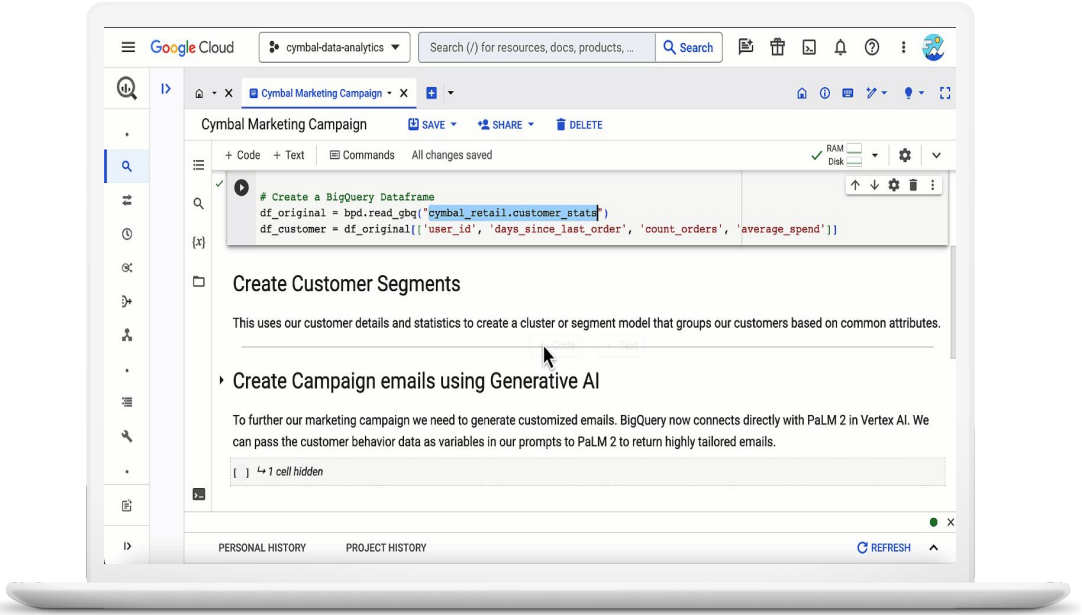


BigQuery Studio

Keeps data in place–which saves time and lets you work with the enterprise scale and security of BigQuery

Duet AI inside the notebook helps **generate code** - segment customers

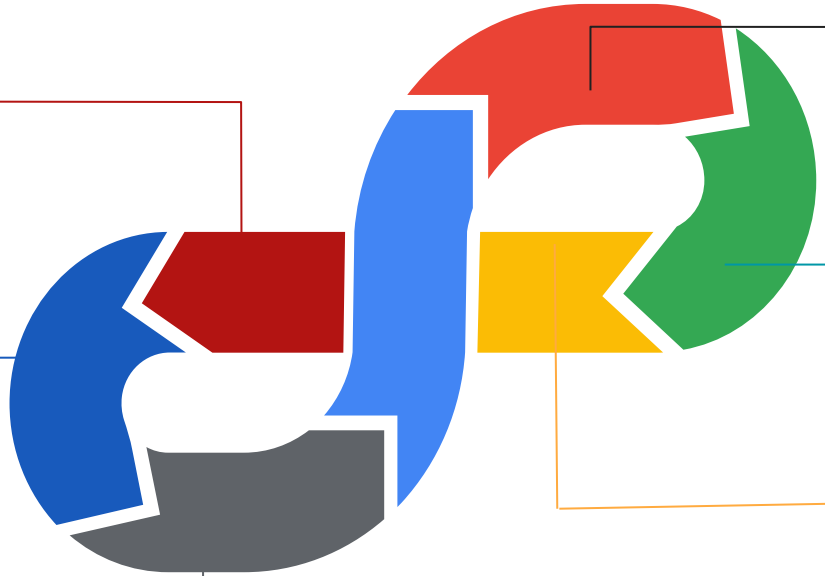
BigQuery integration with Vertex AI to **connect enterprise data directly to LLMs** - Eg: PaLM 2



Duet AI improves each stage of the SDLC

Assisted operations

Assisted development



01. Design

Design application from product requirements document to an API specification and documentation.

02. Build & Test

AI generated code and test cases via smart actions in the IDE. Build/Refactor code in the IDE with **code suggestions** leveraging the same models Google uses in-house

03. Deploy

Contextual suggestions to setup optimal production environments. Duet AI **autonomous agents** automate repetitive tasks in the code review process, based on human comments.

04. Secure

AI-driven **threat analysis** to find security issues. Code generation to **produce infrastructure-as-code** and get guidance on other operational tasks

05. Operate

Log summarization and **natural language log queries**, for faster troubleshooting. Analysing application data to make better business decisions

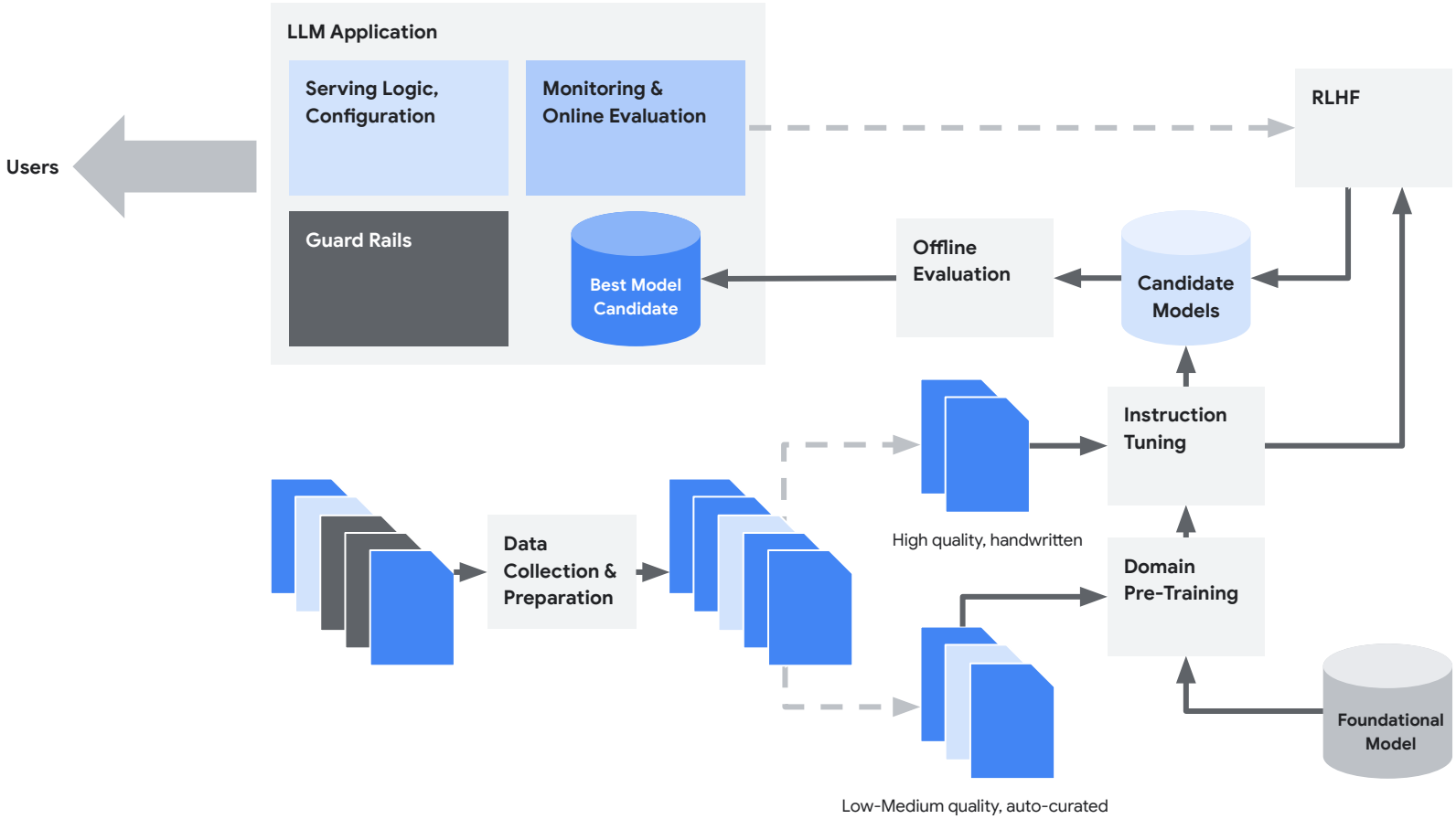
06. Scale

Assistance in scaling the application components

Gen AI assistance across the most popular IDEs via the Codey API

Duet AI - Enterprise-focused

Codey enabled ecosystem


LLM Application
Serving Logic, Configuration
Monitoring & Online Evaluation
Guard Rails
Best Model Candidate
Users
RLHF
Offline Evaluation
Candidate Models
Instruction Tuning
High quality, handwritten
Data Collection & Preparation
Domain Pre-Training
Foundational Model
Low-Medium quality, auto-curated

Writing code: ~8-9% decrease in Edit-Feedback Loop

Single-line

```
def test_tasks_registered(self):
  tasks = ["product_salient_terms", "query_salient_terms", "vanilla_query_salient_terms"]

def test_product_task_registered(self):
  self.assertIsNotNone(_TaskRegistry.get("product_salient_terms"))
```

Multi-line

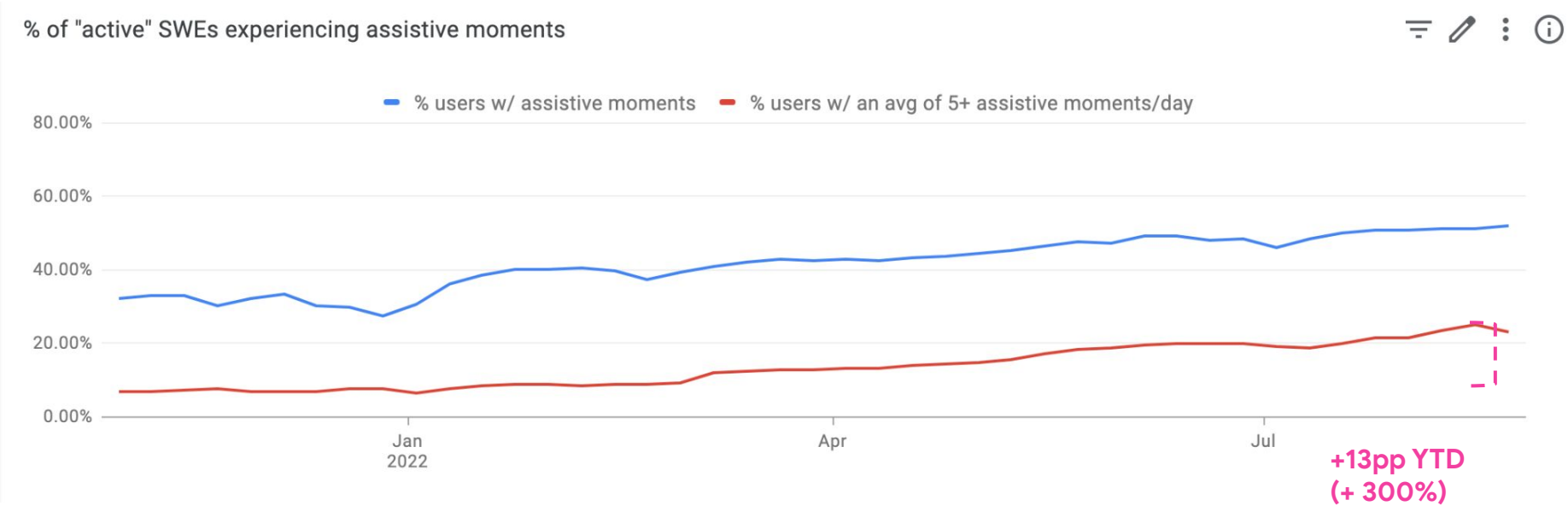
```
345     case strings.HasPrefix(setting, "num_multi_token_predictions="):
346       i, err := strconv.Atoi(strings.TrimPrefix(setting, "num_multi_token_predictions="))
          if err != nil || i < 0 {
            log.Errorf("Unexpected int setting %v with error %v", setting, err)
            continue
          }
347       ret.query.numMultiTok = i
348       ret.mixing.modeMultiTok = mlMultiTokAppend
```

Idiomatic templates

```
@parameterized.parameters((name, args) for (name, args) in GetTests())
def testSuccess(self, name, args):
  inception_response = text_format.Parse(
      self.ReadFile(name + '.textproto'), service_config_pb2.Config())
  original = self.ReadFile(name + '_input.gfe.part.cfg')
  expectation = self.ReadFile(name + '_output.gfe.part.cfg')
  with mock.patch.multiple(gfile, Open=self.mocked_gfile.Open,
                           MakeDirs=self.mocked_gfile.MakeDirs), \
       mock.patch.object(inception_cli, 'GetConfig',
                         return_value=inception_response):
```

Impact so far

Objective: Assistance is pervasive for SWEs: **24%** of SWEs experience assistive moments 5+ times/day



Additive value of genAI in developer environments

- Transformer network architecture invented at Google set the LLM AI revolution in motion
- Palm2 now powering over 25 Google products and features in production
- Palm2 as a basis for many domain specialized derivative models such as health and medicine
- Codey is our our text-to-code foundation model to help improve developer velocity with code generation and code completion, and to improve code quality.

Codey model

LLM R&D (PaLM 2)

Additive value of genAI in developer environments

- Vertex platform provides highly reliable and low latency model serving infrastructure with access to specialized and proprietary TPU hardware.
- Provides GCP-grade API security and opportunity for tenant isolation
- Duet for developers brings the power of Google's generative AI seamlessly into the developer environments and workflows in delightful and low-friction IDE integrations.

Duet (IDE Integration)

Vertex Serving infra

Codey model

LLM R&D (PaLM 2)

Additive value of genAI in developer environments

- Barrier of entry: faster and simpler path to become a developer
- Developer productivity and delight: reduced developer friction and context switching
- Complete security model from workstation to custom models

BeyondCorp (zero trust)

Cloud Workstations

Duet (IDE Integration)

Vertex Serving infra

Codey model

LLM R&D (PaLM 2)

AI Data Governance

Google Cloud's approach to governance of customer data for Cloud AI Large Models and Generative AI

By default, we do not use customer data to train our models, in accordance with *GCP Terms* and *Cloud Data Processing Addendum*.

We will only use customer data as part of model training when given explicit permission to do so.

We are committed to transparency, compliance with regulations like the GDPR, and privacy best practices.

Assistive AI Principles

These principles ***provide context*** for how Google will deliver differentiation and value across all AI assistive experiences:

1. **Uniquely Google:** Deliver experiences that leverage the breadth of Google's expertise across Core, Labs, Cloud, Search, (etc.) that provides differentiated and durable value. Example: AI LLM is informed by data from beyond the IDE and Console to include data within Google Workspace (docs, sheets, email, etc.) and/or Google Search to make more informed AI guidance

2. **Respect customer data:** We will adhere to the GCP AI/ML Data Use Commitment to respect the sovereignty of customer data, even for LLMs that are fine-tuned for specific customers. (LLMs can be trained on publicly available source code (GitHub) with permissive licenses) Won't share model with anyone else

3. **Responsible AI:** We will follow Google Cloud's Responsible AI and AI Design pattern guidance that we have shared with customers. Recitation checking and toxicity filtering will be adopted for all source code completions

4. **Transparency Builds Trust:** While adhering to GCP AI/ML Data Use Commitment, we will provide transparency on citations and what influenced the AI driven decisions and guidance wherever possible

5. **Promotes Best Practices:** Recommended secure and performant code suggestions, ops configurations, most cost-effective solutions (right size) to unlock the potential for every developer to be as effective as a Google developer

6. **Cohesive and Contextual user experience:** AI Assistance will be presented in-context within the user's workflow across "inner and outer loop" GCP workflows and therefore the UI appears in multiple ways. HOWEVER, these experiences will NOT be siloed or feel disjointed, but build upon each other and understand their context and interaction with the user

7. **Responsive user experience:** We will design user interaction models that are tailored to the latency involved in leveraging different AI models which range from instantaneous (milliseconds) to slow (seconds) to long-running (minutes+)

8. **Provide Enterprise Value:** We will design for more stringent Enterprise needs, which will benefit all users.
Example: Support local (within the enterprise) AI "fine tuning" of user data, segregated, and local, to within the enterprise, with opt-in engagement

Duet AI Customization: Your Data, Your Terms