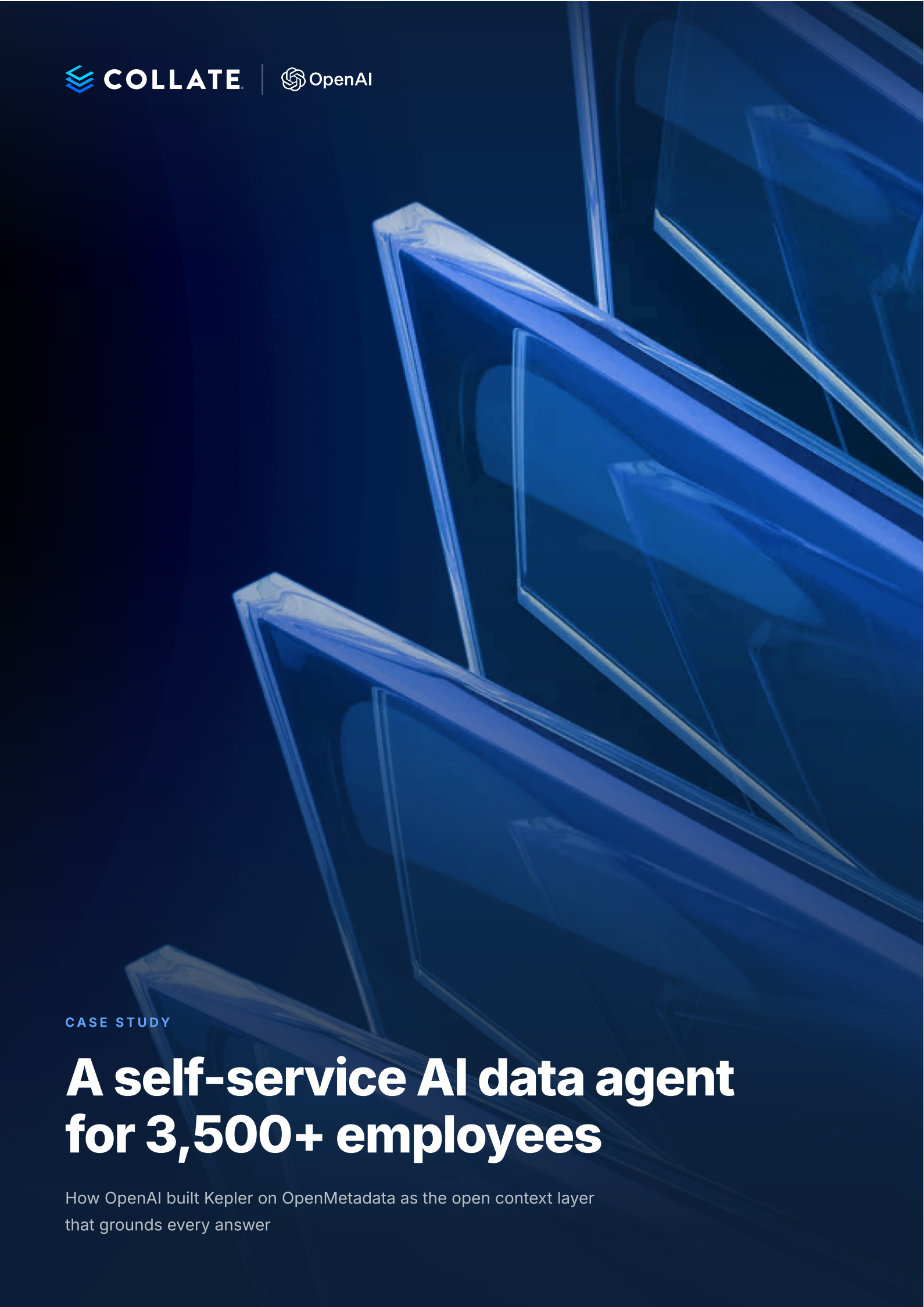
COLLATE | OpenAI



CASE STUDY

A self-service AI data agent for 3,500+ employees

How OpenAI built Kepler on OpenMetadata as the open context layer that grounds every answer

3,500+
Internal users served by Kepler, OpenAI's AI data agent on OpenMetadata

580+ PB
Of data processed daily across 70,000 datasets

<90 sec
Time-to-answer on repeat queries, down from 22+ minutes

OpenAI's data platform serves 3,500 internal users on 15 tools, with 580+ petabytes of data processed daily across 70,000 datasets. At that scale, simple questions like "How many ChatGPT Pro users do we have in Italy?" routinely cost analysts days of table discovery, multiple code deep-dives, and several cross-team Slack threads — and could still return answers off by orders of magnitude. OpenAI's Data Productivity team built **Kepler**, an internal AI data agent, on top of OpenMetadata as the open context layer that grounds every response. The result: self-service analytics across the company, with repeat-query runtime down from 22 minutes to under 90 seconds.

> *"It's about getting the right context — knowing what data exists, what it means, and how it relates to the business. This is where OpenMetadata has been so foundational."*
>
> Bonnie Xu, Tech Lead, Data Productivity at OpenAI

INDUSTRY
AI Research & Deployment

HEADQUARTERS
San Francisco, California

FOUNDATION
OpenMetadata — open context layer for Kepler

TECHNOLOGIES
Spark, Airflow, Notion, Slack, OpenAI APIs, Codex

Days of Searching, Multi-Team Threads, and Wildly Wrong Answers

Eighty percent of OpenAI's employees rely on the data platform every day. As the product scaled, so did the complexity of the questions — and the cost of answering them. Tables proliferated, semantics drifted, and the institutional knowledge needed to use them stayed locked in Slack threads and engineers' heads.

CONFIDENTLY WRONG BY ORDERS OF MAGNITUDE

With 70,000 datasets across 15 tools, analysts faced a haystack of similar-sounding tables with subtle, consequential differences. Missing one nuance could skew answers by orders of magnitude — asked how many ChatGPT users there are, an agent answered 5,062,338. The real number was 800 million.

SIMPLE QUESTIONS, EXPENSIVE ANSWERS

A question that should take minutes cost days of work across teams. A lead asks how many ChatGPT Pro users are in Italy; the data scientist consults engineer after engineer. By the time the answer lands, it has taken three deep-dives, two meetings, and five Slack threads.

CONTEXT TRAPPED IN TRIBAL KNOWLEDGE

A table's name and schema only tell part of the story. Why was it created? What filter was applied at ingestion? That context lived in Notion docs, Slack channels, and individuals' memories — invisible to any agent looking only at metadata.

NO MEMORY ACROSS CONVERSATIONS

Without a memory layer, every correction had to be relearned. Team-specific definitions, edge-case filters, and learned corrections disappeared the moment a session ended — so repeat questions cost as much time as first-time ones.

PERMISSIONS BEYOND THE TABLE

OpenAI's data includes information not every user should see. An agent ranging across all of it had to respect who could access what — extending permission checks to retrieval, sanitized queries, and chain-of-thought, not just row-level access.

An AI Data Agent on the Open Context Layer

OpenAI's Data Productivity team built **Kepler** — an internal AI data agent surfaced wherever work happens: Slack, the IDE, web agents, and workflows. Instead of building a metadata layer from scratch, they put OpenMetadata on top of their data warehouse and let it carry the context the agent needs to reason, organized into six layers it draws on per query.

OPENMETADATA AS THE CONTEXT FOUNDATION

Kepler's first layer is table metadata from OpenMetadata — schemas, query history, lineage, and co-derived tables — transformed into embeddings for semantic search. The agent finds the right table by meaning, not keywords, grounding every answer in governed structural metadata.

CODEX ENRICHMENT & RUNTIME CONTEXT

Codex (OpenAI's coding agent) reads the pipeline code behind each table to explain why it exists — purpose, usage, primary keys. Institutional knowledge from Slack and Notion adds business meaning, and live calls to Spark, Airflow, and OpenMetadata fill in real-time detail.

EVALS & PERMISSION ENFORCEMENT

An evals system using OpenAI's grader — SQL AST normalization, result-set comparison, chain-of-thought debugging — prevents regressions. Permission checks at every layer ensure the agent only surfaces data the user is authorized to see, with query history sanitized end-to-end.

HUMAN ANNOTATIONS, EDITED IN PLACE

Layer two is human annotations — descriptions, tags, usage, ownership — maintained by the data team in the OpenMetadata UI and read via its APIs. Where descriptions are stale, the agent auto-generates candidates the team confirms, so context improves as it runs.

A MEMORY LAYER THAT COMPOUNDS

The final layer is memory. Kepler suggests memories from conversations that users confirm — scoped global or per-user. The agent gets smarter every time someone corrects it: a filter learned today becomes inherited context for every query that follows.

Self-Service Analytics at Enterprise Scale

Kepler now runs in production across OpenAI's data platform, wherever employees work. Users ask in plain language and get answers grounded in the same context a senior analyst carries — and the open foundation, OpenMetadata, is one any data team can adopt.

FROM MULTI-DAY SEARCHES TO SELF-SERVICE

Every data user can now ask questions and get reliable answers independently — no routing through a data scientist who routes to an engineer who hunts through tables. The result is significant time savings and productivity gains across the organization.

TRUSTED ANSWERS GROUNDED IN CONTEXT

Investigating a March 2025 growth spike, Kepler pulled weekly-active-user tables, cross-checked a Notion doc and a dashboard, ruled out a logging-duplication bug, and used web search to land on the real cause: a viral image-generation trend.

VERIFIABLE ANSWERS, EVERY TIME

Every response streams the agent's chain of thought, links each query it ran, and cites every assumption. Analysts can sanity-check the agent's picks and click through to raw results — auditable end-to-end, even when a chat is shared.

REPEAT QUERIES: 22 MIN → UNDER 90 SEC

With memory compounding across conversations, a query that first took 22 minutes and 41 seconds later resolved in 1 minute and 22 seconds — because the agent saved the corrections as reusable context for everyone downstream.

> *"At that scale, the hard part is not just about finding the right table or writing SQL, it's about getting the right context, knowing what data exists, what it means, and how it relates to the business. This is where OpenMetadata has been so foundational. It's an important part of our six-layer context model so that the agent can reason more like a real analyst."*

Bonnie Xu
Tech Lead, Data Productivity at OpenAI

COLLATE

open-metadata.org

THE OPEN STANDARD FOR DATA CONTEXT

Quote verbatim from Bonnie Xu's public Summit '26 keynote and OpenMetadata award acceptance (June 2026). DRAFT — OpenAI logo and likeness use pending PR sign-off before external distribution.