# A bank runs its AI with no line out

## Models, knowledge and sign-in inside the bank's own network, and nothing leaves it.

### THE CHALLENGE

A top-tier bank in Asia wanted to give its staff an AI assistant: ask a question in their own language, and get the answer from the bank's own policies and procedures. A cloud AI service was not an option. The bank's production network has no connection to the internet, and its documents were not going to leave it. It ran a tender to find a platform that could work that way.

### WHY IT WAS HARD

Most AI software assumes a line out. It downloads its models, checks its licence with the vendor and pulls its updates from the internet. Inside this bank, none of that is possible.

| | THREE CONDITIONS THE BANK SET |
|---|---|
| 01 | Nothing leaves the bank's network, and nothing is fetched into it |
| 02 | Staff sign in through the bank's own directory |
| 03 | A wrong answer to staff is a risk to the bank, not a glitch |

**On sign-in.** The tender made it mandatory. Staff use the credentials they already have, and the bank's own directory decides who gets in.

**On accuracy.** The bank's executives were clear that an incorrect answer given to staff creates operational and compliance risk. They wanted a plan to drive it down, not a promise.

The bank did not take the system on trust either. It ran it inside its own network for about five months before it formally accepted it.

## WHAT KATONIC DELIVERED

| 0 | 2 | 4 | 5 |
|---|---|---|---|
| connections out of the bank's network | open models, served inside the bank | GPUs in the bank's data centre | months in the bank before formal sign-off |

**A platform that installs with no line out.** Every container image was carried into the bank's own registry. The models were loaded in advance. The licence is applied inside the cluster and never calls home.

**Two open models, served locally.** Both run on four GPUs in the bank's data centre, beside a smaller model that indexes the documents. The platform is not tied to either one.

**An assistant staff can rely on.** Staff sign in with their bank credentials, ask in their own language, and get an answer drawn from the bank's own documents, with its sources shown. Documents are tagged by business area, so a question is answered from the right part of the bank. The partner's tests showed a clear gain in accuracy.

**Upgrades without opening the network.** The upgrade made before sign-off was run first on a test environment outside the bank. It was then carried in and performed on site by the bank's partner, with Katonic engineers in support.

**The result.** The bank formally accepted the system, and it is in production. Its staff have an AI assistant that works in their language and answers from the bank's own documents. The bank kept what it set out to keep: its documents, its models and its sign-in never left its network. Its next step, now being planned, is agents.

> A bank does not have to choose between AI and a closed network. The platform has to work with nothing coming in and nothing going out.

### About Katonic AI

Katonic AI is the enterprise platform for serving models and building, running and governing AI agents. It runs on your machines, with your models, under your administrators. Katonic is listed in the Red Hat Ecosystem Catalog and runs on Red Hat OpenShift. It is in production at a top-tier bank in Asia, air-gapped, inside the bank's own perimeter. It was founded in Sydney in 2020 by Prem Naraindas, with its global hub in Dubai and teams in Sydney, Mumbai and Dubai.

KATONIC.AI