Empisto architecture: advanced AI, entirely on your premises

15 April 2026

Modern AI is valuable, but most of that value sits behind a cloud API that your most sensitive data has to travel to. Empisto takes the opposite position: the platform, the models, and your data all run inside your own infrastructure, and nothing has to leave for the system to work. This paper sets out how that is put together.

One platform, inside your walls

Empisto installs as a single platform on hardware you control, on-premise or in a private data centre. Everything the product needs runs there: the language models, the retrieval index built from your documents, any applications you license, and the runtime that hosts your own. There is no separate cloud tenant, and no step in normal use where a request is sent to a third-party model.

The data boundary

The defining line in the architecture is the network boundary around your deployment. Inside it sit the models, the retrieval index, application data, and logs. Across it there is no outbound path for your content. The platform is built to run air-gapped, so it can operate with no internet access at all, which is what makes it usable for regulated records, client material under NDA, and data with residency constraints. Where the platform needs updates or licences, those are handled as explicit, reviewable steps, not as a background connection that also carries your data.

Retrieval that is useful on day one

The first thing most teams want is to ask questions of their own documents and get grounded answers. Empisto builds a retrieval index from the content you point it at

and serves search and assistance over it from the first install. Because the index and the models are local, the documents behind an answer never leave the boundary, and the same holds for the prompts your people type.

Applications and the SDK

On top of retrieval sit applications, licensed separately from the platform and released on their own schedule. Each one is a self-contained slice: a server module the platform loads, its own data, and its own place in the interface. When the available applications are not enough, your own developers can build new ones with the software development kit (SDK), on the same foundation and with the same guarantee about where data can go. The unit of extension is the application, so a custom build inherits the platform's boundaries rather than working around them.

Models on your own hardware

Empisto runs open models on hardware you own, through the same standard serving stack the industry already uses, so capability does not have to be traded away for control. The model weights and the inference run locally; there is no per-token call to an outside provider, which is what makes fixed-cost licensing possible and keeps usage from turning into a meter.

A commercial model that matches the promise

Empisto is licensed, with support and version upgrades, rather than billed per token or per unit of data processed. That is not only a pricing choice. Because the revenue comes from licences, there is no commercial reason to collect, retain, or train on your data, and the architecture reflects that. Trust is meant to be verifiable rather than asked for, so the data handling and the network behaviour are open to independent audit.

Deployment

A typical deployment places Empisto entirely within your existing network controls, behind your firewall and identity systems. Air-gapped installation is supported for the strictest environments. Because everything runs inside one boundary, there is a

single perimeter to reason about rather than a scattered set of cloud integrations to review.