[intradiegetic

Building the next communication OS

Local and Cloud AI Systems for Corporate Communications

LOCAL CLOUD HYBRID ARCHITECTURE

A practical architecture guide for deciding what should run locally, what should run in the cloud, and how communications departments change when AI becomes infrastructure.

Prepared by Intradiegetic · May 2026

What this guide is for

This report is a companion to the Human × AI Toolkit. It moves from work allocation to infrastructure: where intelligence should live, how local and cloud systems differ, and why the communications department changes shape when AI becomes part of its operating architecture.

The purpose is not to argue that local AI is better than cloud AI, or the reverse. The useful question is architectural: which cognitive work should remain close to organisational memory, which work can be extended into cloud systems, and where human judgment must remain the authority layer.

How to use this document

Read it once as a strategic argument. Then use the diagrams as operating models: cloud for reach, local for depth, hybrid for control, and the Intradiegetic example as a concrete pattern for communications architecture.

SECTION 01

Where intelligence lives.

> The first AI question was about access. The next one is architectural: where should intelligence live inside the communications department?

The wrong question

Most organisations begin with a vendor question because vendor questions are easier to ask. A department can ask whether it should use Microsoft Copilot, ChatGPT Enterprise, Claude, Gemini, Perplexity, or a specialist communications platform without having to reconsider its own shape.

But the vendor question hides the architectural one. Cloud services and local systems do not merely differ in where computation happens. They differ in what kind of department they make possible.

CLOUD AI

Immediate access to powerful external intelligence: elastic, constantly updated, and usually easy to deploy. NIST defines cloud computing around on-demand network access, pooled configurable resources, rapid elasticity, and measured service.[1]

LOCAL AI

An internal intelligence layer: models, retrieval, automation, and memory placed close to organisational materials, often on a workstation, server, or private network. Ollama, for example, runs local models through a local API endpoint.[2]

> The strategic question is not cloud or local in the abstract. It is which kinds of cognition should be external, which should be internal, and which should be mediated by a governed handoff between the two.

1. NIST: https://www.nist.gov/document/nist-sp-500-322-evaluation-cloud-computing-services-...
2. Ollama documentation: https://docs.ollama.com/quickstart

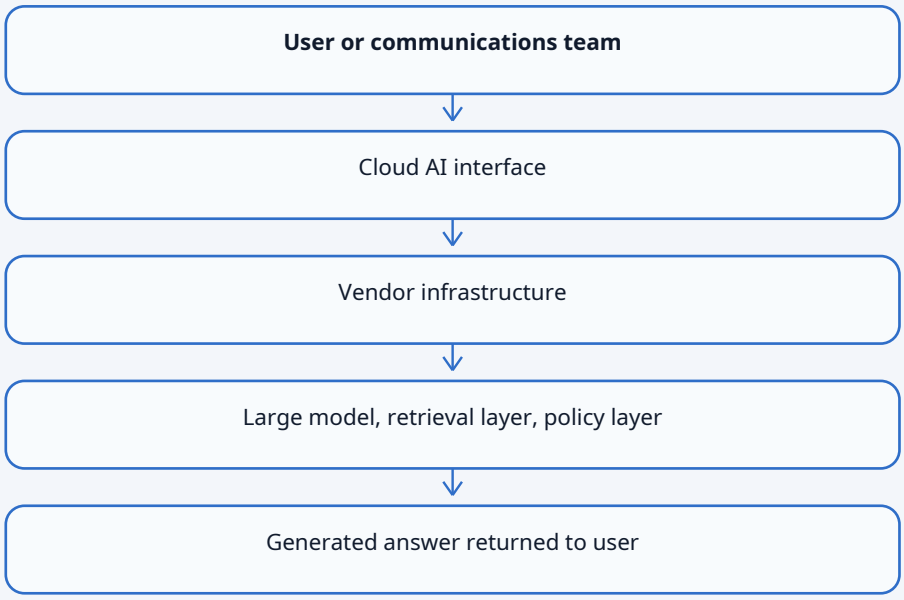
SECTION 02

Cloud and local systems.

> Cloud AI gives the department reach. Local AI gives it depth. The mature architecture knows which work belongs where.

How cloud AI works

Cloud AI begins with distance. The user interacts with a model or application that runs outside the local machine, usually in infrastructure owned, operated, or orchestrated by a vendor. The organisation sends a request across the network. The answer returns to the user interface.



This architecture removes many burdens from the organisation. The department does not need to buy GPUs, manage inference servers, update models, maintain compatibility, or optimise performance.

The price of that convenience is dependence. Modern enterprise cloud AI systems often include serious controls: Microsoft says Copilot operates inside the Microsoft 365 service boundary and scopes access to the signed-in user's permissions; OpenAI says API data is not used to train models unless the customer explicitly opts in.

[3]

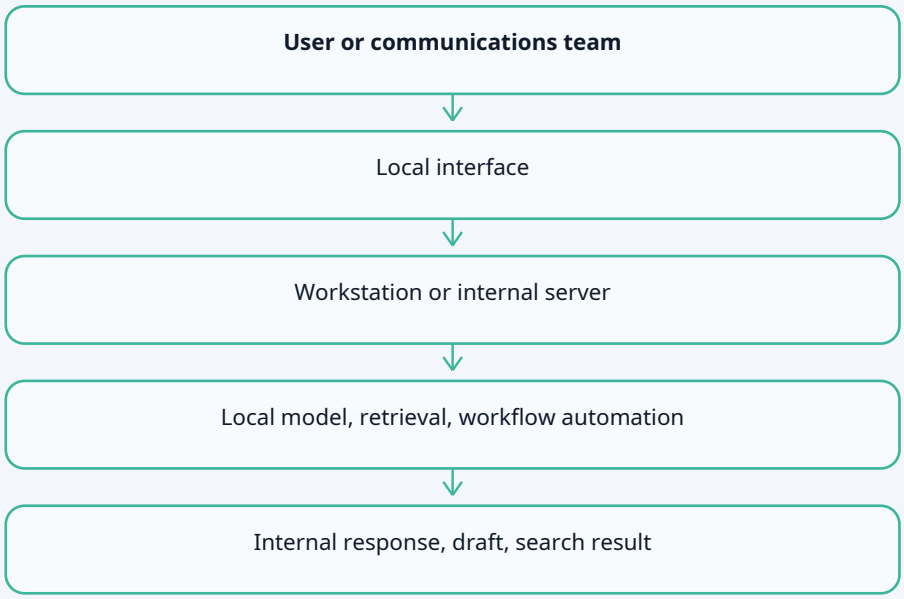
[4]

3. Microsoft Learn: https://learn.microsoft.com/en-...

4. OpenAI API data controls: https://developers.openai.com/api/docs/guides/your-data

How local AI works

Local AI begins with proximity. The model, retrieval system, automation engine, or interface runs on hardware controlled by the organisation or individual. It may still call cloud models when needed, but its centre of gravity is internal.



This changes the relationship between AI and organisational memory. A cloud chatbot is usually experienced as a place to ask questions. A local AI system can become a place where the department's own knowledge is indexed, retrieved, structured, and used.

WHAT LOCAL SYSTEMS CAN SIT CLOSE TO

Message houses, leadership language, past speeches, crisis statements, employee communications, analyst briefings, stakeholder maps, tone guidelines, transcripts, meeting notes, and the informal archive of how the organisation actually talks when the pressure rises.

The trade-off is not simple

There is a temptation to turn the comparison into ideology. Cloud AI is sometimes treated as careless outsourcing. Local AI is sometimes treated as automatically sovereign, private, and superior. Neither position is adequate.

Dimension	Cloud AI systems	Local AI systems
Adoption speed	Fast to start because infrastructure is provided by the vendor.	Slower because hardware, models, storage, and access patterns must be configured.
Model capability	Usually strongest for frontier reasoning, multimodal work, coding, and broad knowledge.	Depends on hardware, model size, quantisation, and local optimisation.
Data posture	Data moves through an external service boundary, even with strong controls.	Data can remain on controlled machines or networks if designed correctly.
Governance	Shared between internal rules and provider policies, contracts, and service boundaries.	Internal governance is required for access, retention, permissions, updates, and auditability.
Strategic value	Excellent for rapid augmentation and outside capability.	Stronger as a capability layer attached to organisational memory and workflow.

> The mature communications department will not ask for one universal answer. It will build a routing logic.

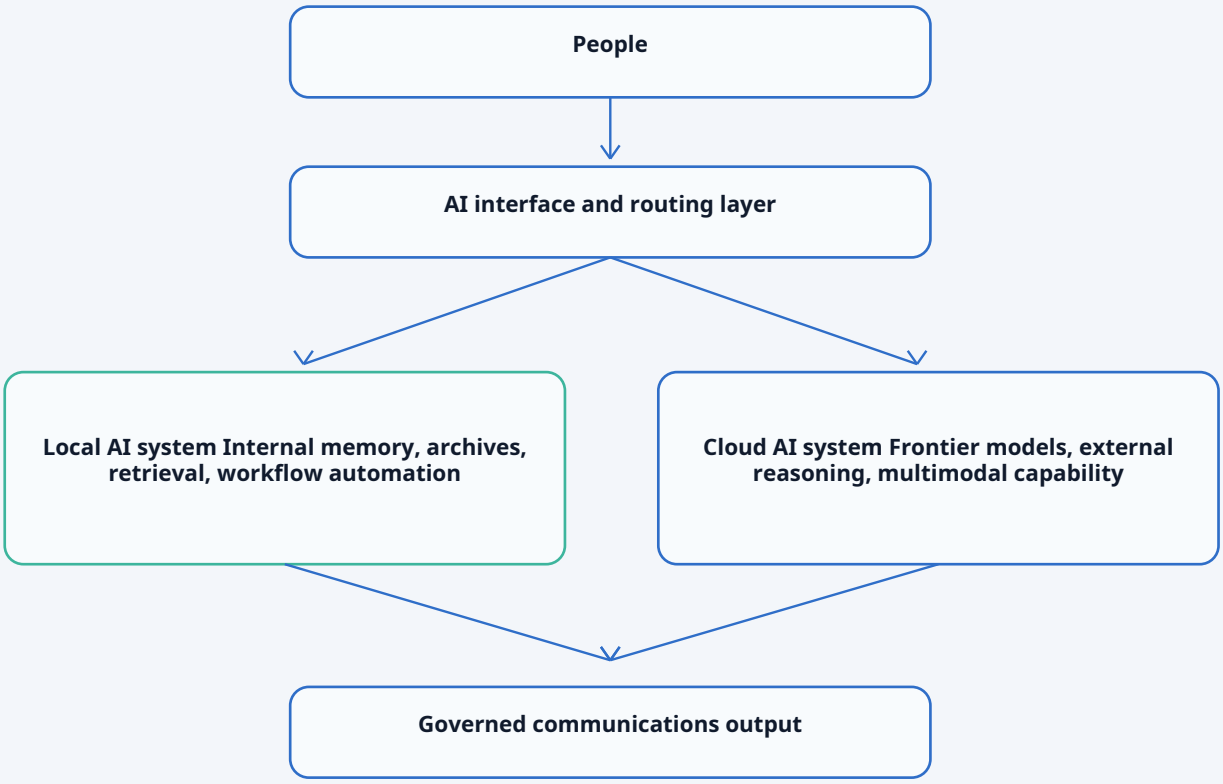
SECTION 03

The hybrid layer.

> The point is not to choose one column. It is to build a routing layer that governs how the two work together.

The hybrid layer

The most important architecture is likely to be hybrid. Not because hybrid is a compromise, but because communications work itself is hybrid: public and private, fast and slow, strategic and tactical, human and machine, regulated and informal.



THE ROUTING RULE

Local AI is the department's memory, operating base, and control surface. Cloud AI is a selectively invoked extension of capability. Human judgment decides what can leave the building, what must stay inside it, and what should never enter an AI system at all.

SECTION 04

Department architecture.

AI changes communications from a production function into an intelligence system of people, models, memory, workflow, and judgment.

What changes inside the department

The traditional communications department was designed around outputs: requests, drafts, review, approval, publication, and passive archive. AI makes that structure untenable because it increases the speed of production without automatically increasing the coherence of the system producing it.

The communications department becomes less like a production desk and more like an intelligence system. It still writes, edits, advises leaders, and protects reputation. But those activities are embedded in a loop of sensing, retrieval, generation, evaluation, publication, and learning.

The new responsibilities

Once AI becomes architectural, the communications department needs responsibilities that were previously informal.

Responsibility	What it governs
AI routing	Which tasks go to local models, cloud models, human experts, or no AI system.
Memory architecture	The usable archive, retrieval logic, and knowledge structure of the department.
Voice governance	What the organisation sounds like across contexts, leaders, risks, and channels.
Workflow design	Repeated communications work turned into repeatable, inspectable processes.
Approval integration	How AI-assisted work enters legal, executive, regulatory, and reputational controls.
Feedback instrumentation	What happened after publication and how it feeds back into planning and memory.

These are not purely technical roles. They are communications roles under architectural conditions. IT can secure, connect, monitor, and maintain the systems. Communications must decide what the systems are for, what they may say, and what they must never do.

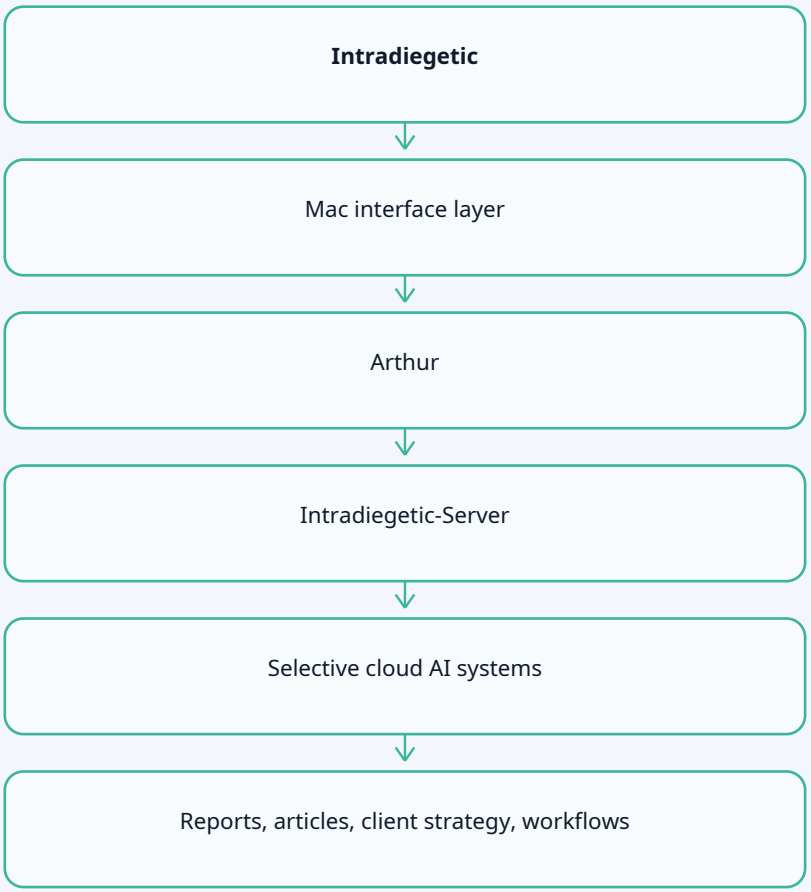
SECTION 05

Intradiegetic as a working example.

A small-scale architecture can make the principle concrete: local memory and orchestration, selective cloud extension, human authority.

Intradiegetic as a working example

Intradiegetic is a useful example because it is not starting from the assumption that AI is one application. It is closer to a small-scale version of the architecture larger organisations will eventually need.

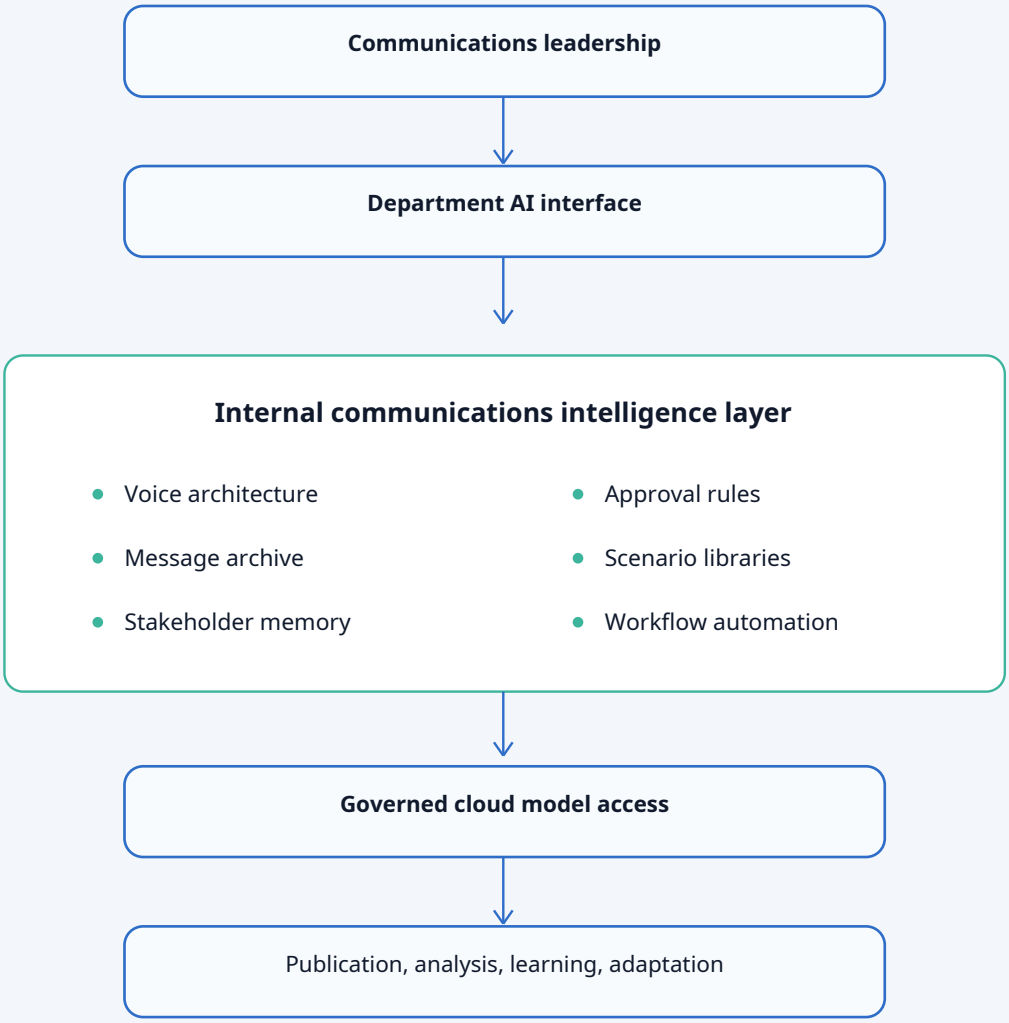


SERVER LAYER

Local files and working archive · Retrieval and knowledge base · n8n automation · Local AI models and agents.

How the pattern scales

The lesson is not that every department should copy the exact Intradiegetic configuration. The lesson is that the communications department needs an internal centre of gravity before it connects itself to external intelligence at scale.



SECTION 06

The architectural principle.

> Local and cloud AI are not tools. They are department architectures.

The architectural principle

The defining question for the next communications function is not whether AI will be used. It will be. The question is whether that use will remain a scatter of individual prompts inside vendor interfaces, or whether it will become a governed system of people, models, memory, workflow, and judgment.

CLOUD

Reach. External capability. Frontier models when their scale justifies the handoff.

LOCAL

Depth. Internal memory. Workflow control close to organisational knowledge.

HYBRID

Shape. Routing, governance, and human authority across both layers.

AI does not simply make the communications department faster. Speed is the least interesting outcome. The deeper possibility is that communications becomes an operating system for organisational meaning: a system that remembers, interprets, generates, checks, publishes, and learns.

> Without that architecture, AI produces more language. With it, AI helps build a department capable of knowing what its language is for.