

SOFTWARE

Permission-aware enterprise search and actions for a software company

Employees needed one place to ask questions across chat, email, documents and HR policies, and to act on the answer. Wexa built an assistant that answers only from what each person is allowed to open, cites its sources, and performs actions after confirmation and approval.

-80%

tokens vs sending raw data to the model (up to -82%)

18 / 20

correct answers vs 8 / 20 without Wexa

0

permission leaks across 9 probes

0.79

hit rate in the top 3 results

CLIENT

Enterprise software company (in-house deployment across chat, email, drive and HR policies)

REGION

US / India

WHEN

2026

ENGAGEMENT

Solution build

WEXA COMPONENTS

Wexa · connectors · automatic ontology discovery · Wexa memory layer · approvals

The challenge

- Knowledge scattered across chat, email, documents and policies
- Search tools that ignore who may see what
- Answers that cannot be acted on safely

What Wexa built

- ✓ Ingestion of files and messages with automatic ontology induction into the graph
- ✓ Hybrid retrieval: keyword, graph and knowledge base, run as the signed-in user
- ✓ Tool-calling agent that plans, cites and loops until the task is done
- ✓ Writes confirmed in chat, approved by a different person and executed exactly once; every step audited

How it runs on Wexa

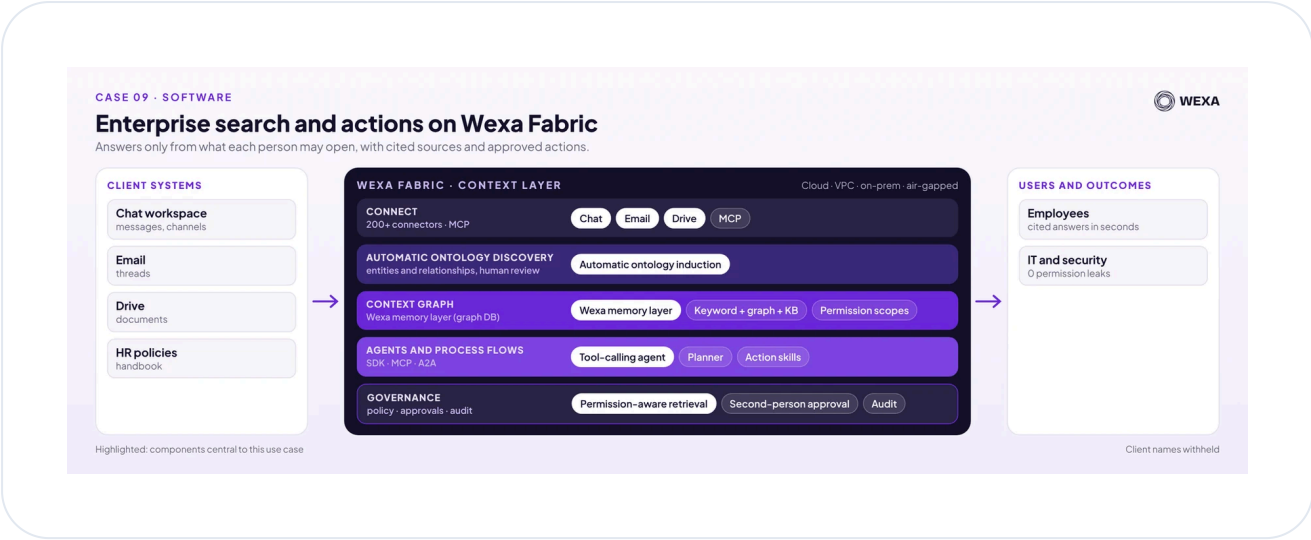


FIGURE How this use case runs on Wexa: client systems connected, ontology discovered, context graph in the Wexa memory layer, agents and process flows, and governance. Highlighted components are central to this use case.

Results

MEASURE	BEFORE / WITHOUT WEXA	WITH WEXA
Hours a week each employee spends looking for information (30 people)	2.5 hours	1 hour
Annual cost of the process	\$144K	\$58K
Time to implement the AI solution	3 months	1 week
AI build cost (one-time)	\$96K	\$4K
AI operating cost per year (models, infrastructure, upkeep, platform)	\$65K	\$29K

MEASURE	BEFORE / WITHOUT WEXA	WITH WEXA
Total AI cost in year 1	\$161K	\$33K

WITH WEXA

\$57K

net annual value to the client after Wexa's
own operating cost

12×

faster to implement

56%

lower AI operating cost

\$128K

less AI spend in year 1

- ✓ Started as the enterprise-agent track of the October 2026 hackathon
- ✓ API stress test: 47 routes, worst p95 under **380 ms** at 12 concurrent sessions, no server errors