IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Enterprise AI Knowledge Bases: RAG and Egnyte Copilot

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-03-31 · enterprise ai · knowledge base · rag architecture · egnyte copilot · document retrieval · ai assistants · semantic search

Original article: https://intuitionlabs.ai/articles/enterprise-ai-knowledge-bases-rag-copilot



In this article

1. Executive Summary
2. Introduction and Background
3. The Enterprise Knowledge Stack: Components and Architecture
4. Egnyte Copilot: An In-Depth Look
5. Launch and Key Features
6. Turning Folders into Knowledge Bases
7. Architecture and Implementation (Egnyte RAG)
8. Governance and Compliance
9. Initial Adoption and Case Study
10. Integration and Ecosystem
11. Chatbots vs. Retrieval-Augmented Systems
12. Data Analysis and Industry Perspective
13. Broader Implications and Future Directions
14. Conclusion

Executive Summary

The landscape of enterprise knowledge management is undergoing a rapid transformation with the advent of generative AI. Traditional file repositories and document management systems are being repurposed as AI-driven knowledge bases, enabling employees to query internal content using natural language. A leading example is **Egnyte Copilot**, launched in July 2024, which allows organizations to “start engaging in AI-powered conversations with their own private and trusted data” (www.egnyte.com). Egnyte Copilot functions as an AI assistant embedded in the content platform, automatically **surfacing insights, summarizing documents, generating new content, and even transcribing audio/video files** (www.businesswire.com). Importantly, Egnyte and similar systems emphasize *source-grounded answers*: every AI-generated response is linked back to the originating documents, ensuring transparency and reducing hallucination. For focused Q&A, use **Create New Agent** and select the **Knowledge Base** template, which searches selected Egnyte content and cites its sources (migration notice).

This report investigates the **“new enterprise knowledge stack”** whereby file repositories become usable AI knowledge bases. We analyze the architecture and workflow of such systems (including retrieval-augmented generation, or RAG), examine multiple platforms (Egnyte, Microsoft, Google, etc.), and compare the roles of plain ChatAI versus retrieval-based approaches. Core findings include:

- **Widespread AI adoption:** Enterprise use of AI has surged in recent years. McKinsey’s early-2024 survey found that 65% of respondents reported that their organizations regularly used generative AI in at least one business function (McKinsey survey). AI-powered knowledge search and virtual assistants are key growth areas as companies seek productivity gains.

- **Retrieval-Augmented Generation (RAG) adds enterprise context:** General-purpose chatbots and RAG systems draw on different information sources. Industry sources distinguish early GenAI search based on an LLM's training data from systems that add company-document context (www.coveo.com). RAG architectures combine LLMs with indexed enterprise data to support enterprise knowledge-retrieval applications (www.coveo.com). For instance, a KG-RAG study compared knowledge-graph-enhanced retrieval with an existing embedding-retrieval production system. On 100 domain-specific questions, fully relevant responses rose from 17 to 32 (88.2%), while irrelevant responses fell from 52 to 25 (51.9%). These are relevance ratings from an LLM-based evaluation, not a general factual-correctness rate; the small, domain-specific dataset limits generalization (original paper). Egnyte's own solution demonstrates this: by enabling users to query over *their* documents (via Egnyte's file system and vector indexes) and by splitting and re-ranking content chunks, Copilot produces answers that cite real source files (www.egnyte.com).
- **Multiple industry approaches:** Beyond Egnyte, major vendors are embedding AI into their content platforms. Microsoft's Dynamics 365 Copilot **can integrate SharePoint content** as a knowledge source: after configuration, the system "uses the federated search framework … with Graph APIs to search for SharePoint content in real time" and then feeds the results to Azure OpenAI to generate answers (www.microsoft.com). Similarly, Google's Workspace has added **Gemini in Drive**, which "synthesizes information on a specific topic or project…from multiple files" across Google Docs, Sheets, etc. (workspace.google.com). Firms like Salesforce and Atlassian are also bringing generative AI into knowledge bases (e.g. Confluence's AI features), leveraging their existing repositories. These solutions all follow the same principle: treat the corporate document store as a *knowledge base* that the AI can query and explain from.
- **Security and governance:** A recurring theme is enterprise data privacy and policy compliance. Egnyte documents closed models with no cross-training (AI Assistant overview). This training commitment should be assessed separately from external processing, regional configuration, and chat retention. Microsoft likewise provides enterprise-grade security; for example, its Copilot integration can search only those SharePoint portals explicitly connected by the admin (www.microsoft.com). Google advertises its Workspace AI as "enterprise-grade" and subject to data controls. In practice, companies are implementing governance to avoid oversharing or stale content (e.g. using dashboards to find overshared SharePoint sites before enabling Copilot) (techcommunity.microsoft.com).
- **Use cases and benefits:** Egnyte's Knowledge Base template searches selected content and cites its sources (migration guide). The technology is being applied in scenarios from HR policy Q&A and contract analysis to technical support and research, wherever large pools of enterprise content exist.

In summary, enterprises are increasingly turning their document repositories (file shares, cloud drives, wikis) into **AI knowledge bases**. This stack typically includes content ingestion, search/indexing (often a vector database), LLMs, and a chat interface. The result is an AI "copilot" that answers questions with references to actual documents, accelerating knowledge work. This report explores the historical context, architectural patterns, data-driven evidence, and future implications of this trend. We also discuss the crucial question of "when is a simple chat bot enough, and when do you need a retrieval system?" – a key consideration in workshop guidance for AI deployment. All claims below are backed by industry sources and expert analysis, as cited throughout.

Introduction and Background

Knowledge management has long been a critical concern for organizations. Historically, companies attempted to capture institutional knowledge in static forms: **knowledge bases**, FAQs, enterprise search indices, and hierarchical file shares. Early approaches centered on either structured databases or full-text search. For example, by the 2000s many enterprises adopted intranet search engines or knowledge-base software (e.g. SharePoint, Confluence, Zendesk). These systems allowed employees to find documents via keywords, document metadata or curated articles. However, despite tech investments, **information often remained siloed** and hard to find. Employees frequently lamented spending excessive time hunting for the right document or waiting for expert answers.

As databases and search matured, some firms built **knowledge graphs** or ontologies to represent relationships among concepts (e.g. IBM's Watson Knowledge Studio, or Google's Knowledge Graph for news articles). While useful, these solutions required significant manual curation. They generally did not allow free-form Q&A over arbitrary text corpora. In parallel, machine learning and NLP advancements brought chatbots and "virtual assistants" to the customer service space (e.g. rule-based FAQ bots, then LSTM-based chatbots). Yet these were confined to predefined flows or limited scopes, not open-ended company knowledge.

The breakthrough came with **large language models (LLMs)** and generative AI (2020 onwards). Models like GPT-3/4, Claude, PaLM, etc. demonstrated an ability to understand queries and generate human-like answers. Early media and tech hype (2022–2023) proclaimed that AI agents could replace search entirely. For enterprise use, the distinction centers on two aspects of how context is supplied:

- **Organization-specific context**: A general-purpose LLM responds from the information available in the conversation. For a question such as "What is our new expense policy?", the relevant company policy needs to be included in that context.
- **Source context**: Enterprise workflows may require answers to be reviewed against the underlying corporate documents. Retrieval systems can attach document references to generated responses.

These requirements have led enterprises to use hybrid **retrieval-augmented generation (RAG)** architectures. Instead of letting the AI guess, the system maintains an index of actual documents (or even a knowledge graph), and at query time it retrieves relevant content to feed the LLM. This ensures that answers are *grounded* in real information. Among industry analysts, this approach has come to be seen as the "next evolution" of AI for knowledge work. For example, search vendor Coveo distinguishes early GenAI solutions based on an LLM's training data from systems that add enterprise data through retrieval (www.coveo.com). By contrast, modern RAG systems explicitly combine the LLM with the enterprise's data, "improving ... knowledge retrieval applications" (www.coveo.com). Similarly, Salesforce recommends "grounding" generative models in a curated knowledge base to make them "more accurate and specific to your organization" (www.salesforce.com).

Thus, the **New Enterprise Knowledge Stack** can be conceptualized as follows:

1. **Content Layer (Documents, Files, Repositories)**. This is the raw data: Word docs, PDFs, spreadsheets, slides, email, CRM notes, internal websites, etc. It includes both structured and unstructured information. For instance, an HR folder might have policy manuals (text) and training videos (media). Egnyte, Google Drive, SharePoint, and similar platforms serve as examples of such repositories.
2. **Indexing / Retrieval Layer**. Here content is ingested and made searchable. Modern systems use vector embedding databases (like FAISS, Pinecone, etc.) to enable semantic search. They often also maintain keyword indices (Elasticsearch, Lucene) for exact matching. Egnyte's engineers experimented with FAISS for FAQ retrieval (www.egnyte.com). They found that *combining* keyword (Sparse) and embedding (Dense) retrieval gave the best results (www.egnyte.com), leveraging both metadata/terms and semantic similarity. This layer handles query rewriting, chunking large documents, reranking results, and finally retrieving a set of relevant "chunks" of content.
3. **AI / LLM Layer**. The retrieved content is fed into a large language model. This stage might rephrase the user's question (query rewriting) (www.egnyte.com), then calls the LLM with the question and the retrieved text chunks. The LLM produces a natural-language answer. In a well-designed system, the answer will include citations or references to the source chunks (www.egnyte.com). Egnyte's Part 2 blog details this "Generation" phase – it even maintains chunk IDs to ensure each fact in the answer can be traced back to a file (www.egnyte.com).
4. **User Interface / Application Layer**. Finally, there is a chat or Q&A interface where an end-user interacts with the AI assistant. This may be integrated into an existing platform (e.g. Egnyte's web UI or Slack) or a new app. Importantly, the UI often lets users select which "knowledge bases" or folders to include, view source citations, regenerate answers, provide feedback, etc. For example, Egnyte's current creation workflow uses **Create New Agent** and the **Knowledge Base** template (migration notice).

In summary, the content layer provides data; the retrieval layer ensures quick lookup and relevance; the AI layer synthesizes answers; and the UI layer delivers answers with context and provenance. **These components together form the knowledge AI stack**. Egnyte documents closed models with no cross-training (AI Assistant overview); that commitment does not by itself establish data confinement or zero retention.

The remainder of this report examines each element of this stack in detail. We analyze specific implementations (with a focus on Egnyte Copilot), compare with other platforms, review empirical evidence of benefits, and discuss when general-purpose chat models suffice or when full RAG systems are necessary. We also address challenges (data quality, security, cost) and future directions (e.g. structured data extraction). Throughout, we ground the discussion in figures, case studies, and expert commentary from credible sources.

The Enterprise Knowledge Stack: Components and Architecture

1. Content and Repositories: At the base of the stack is the raw content, typically scattered across file systems, cloud storage, collaboration tools, and legacy databases. Examples include:

- **Egnyte (files and folders):** Egnyte's platform can house terabytes of data in hierarchical folders. In practice, users centralize contracts, project files, policies etc. Egnyte's **Knowledge Base** template searches selected Egnyte content and cites its sources (migration notice). In effect, Egnyte turns its file repository into the knowledge base for AI queries.
- **SharePoint / OneDrive:** Microsoft 365 customers often store documents in SharePoint sites and OneDrive. SharePoint can similarly group content by site or library. Microsoft's Copilot for Customer Service (D365) now allows admins to pick SharePoint as a knowledge source (www.microsoft.com).
- **Google Drive/Workspace:** Any files (Docs, Sheets, PDFs) in Google Drive can be accessed if permitted. Google's Gemini in Drive feature explicitly taps Drive content (workspace.google.com).
- **Other sources:** Corporate Wikis (e.g., Confluence pages), CRM notes, internal websites, and even third-party knowledge bases (e.g., support portals) can be pulled in. Some systems allow connections to databases or APIs.

All this content may include text, images, audio/video, or even structured data. For knowledge AI to work, the system must parse and index diverse formats. Egnyte's structured data effort shows they had to extend their RAG pipeline to handle tables, charts, and complex layouts (www.egnyte.com). This is a frontier area: standard RAG pipelines excel on plain text but struggle to interpret the spatial meaning of a table or the semantics of an embedded diagram. Egnyte reports dealing with issues like "positional understanding" in tables and implicit calculations (www.egnyte.com). Thus, enterprises increasingly invest in preprocessing (e.g., OCR, PDF parsing, table extraction) so that all content becomes "machine-readable" text. The better the ingestion quality, the more reliable the AI's answers will be.

2. Retrieval Infrastructure: Once content is ingested, it is indexed for retrieval. Modern architectures use a hybrid approach:

- **Sparse (Keyword) Search:** Traditional search engines, like Elasticsearch or database text search, index actual words and metadata. Queries matching keywords yield documents or doc-snippets that contain those words. This is fast and exact, especially for factual lookups.
- **Dense (Semantic) Search:** A newer layer generates vector embeddings of text chunks. For example, Egnyte's blog describes using an "embeddings generator" (likely an LLM or specialized model) to convert text chunks into 1024-dimensional vectors (www.egnyte.com). These vectors are stored in a vector index (Faiss, HNSW, etc.). At query time, the user's question is also embedded and we perform a nearest-neighbor search. This finds documents whose semantic content matches even if they don't share exact keywords.

- **Hybrid Approaches:** Egnyte found that **combining both keyword and embedding search** outperforms using one alone (www.egnyte.com). Indeed, a query often benefits from filtering (keywords) plus broader semantic matching. Many systems reorder results by combining scores. For example, after retrieving candidates by one method, a transformer-based cross-encoder may rerank them by how well they semantically match the query (www.egnyte.com).

Technically, Egnyte's pipeline (as revealed in their engineering blog) goes through steps: rephrase the query (improve retrieval), search the index (sparse & dense), split long documents into chunks, rerank with a cross-encoder, choose the top-10 chunks, then run the LLM on them (www.egnyte.com). They store "chunk offsets" so that if a chunk is used in an answer, the system knows exactly which file and offset it came from (www.egnyte.com). This enables citing sources by file name or ID.

3. Generation and LLM: The heart of the stack is the LLM (e.g. GPT-4, Anthropic, custom model). The LLM is typically *not* retrained on the corpus; instead, it consumes retrieved content at query time. RAG itself does not establish data-confinement, retention, or freshness guarantees.

Egnyte's AI policy describes separate deployments by region, with storage residency tied to the customer's domain configuration. Its subprocessor list identifies Google, Anthropic and OpenAI as AI service providers, with the USA or other customer-selected regions listed as processing locations. Review the AI policy and administrator settings for the applicable deployment. Egnyte notes that embedding generation (the vector embedding step) is the most expensive part, so they even considered running that via cloud LLMs (www.egnyte.com). The LLM then outputs a concise answer. Critical to enterprise use, Egnyte ensures the **answer includes citations**. They even parse the LLM's formatted output (initially as JSON) to extract citations and link them to source files (www.egnyte.com). The result is a response like:

> **Q:** "What are the main compliance requirements for our new HR policy?" **A:** "According to our Employee Handbook (Section 3) and the HR Guidelines document (page 5), the key requirements are [summarized insights] (www.egnyte.com)."

By design, the system wraps up answers with references so that users can click through to verify. This **source-grounding** dramatically increases trust. Indeed, Vineet Jain of Egnyte highlights authentic, data-backed responses as a copilot goal (www.businesswire.com).

4. Conversation Interface: Finally, the answer is presented through a user interface. This might be a chat window in the Egnyte web app, a Slack integration, or another client. Features at this layer include:

- **Prompt templates / Suggested Questions:** Egnyte Copilot provides prewritten prompts or questions to guide new users (helpdesk.egnyte.com).
- **Multi-turn Context:** The user can ask follow-ups; Egnyte retains chat history.
- **Agent Selection:** After a domain migrates, former Knowledge Bases appear under **Agents** on the AI Assistant page (migration notice).
- **Feedback Tools:** Users can mark answers as correct/incorrect, prompting retraining or re-tuning.
- **Security Controls:** Egnyte's AI Safeguards policies control which files AI can use to generate responses and which users can access AI-powered features (AI Safeguards overview).

In Egnyte's case, all these operate under their content governance framework (www.businesswire.com) (www.egnyte.com), meaning that usual permissions (who can see which folder) still apply to Copilot queries. This is vital: an AI assistant will not bypass access controls. (Microsoft's system similarly respects Graph permission settings (www.microsoft.com).)

In aggregate, this stack **turns conventional file storage into an indexed, retrieval-enabled knowledge base** (Egnyte architecture). Rather than searching manually, an employee can simply "chat" with the system, which in turn fetches and synthesizes the needed facts. KMWorld's 2025 AI 100 introduction presents a list of vendors for readers to explore (www.kmworld.com).

The next sections will illustrate this architecture in action, and contrast it with a purely generic chatbot or a traditional search.

Egnyte Copilot: An In-Depth Look

Egnyte Copilot is a concrete instantiation of this knowledge stack, built by an established enterprise content platform. Understanding Egnyte's approach sheds light on how organizations are crafting their own AI knowledge bases.

Launch and Key Features

Egnyte announced Copilot in July 2024 (www.egnyte.com), positioning it as "a generative AI-based Copilot to speed up knowledge work and unlock enterprise value from content." The press release and documentation highlight its core capabilities:

- **Summarization:** Egnyte Copilot can summarize individual documents or entire folders of content. For instance, asking it to "summarize project Apex training materials" would yield a condensed overview of all docs in the designated folder. This allows quick gleaning of key points without reading long reports. (This feature is explicitly listed: "summarization" (www.businesswire.com).)
- **Knowledge Base Q&A:** In the current workflow, use **Create New Agent** and select the **Knowledge Base** template to search selected Egnyte content with source citations (migration notice).
- **Content Generation:** In addition to retrieving info, Egnyte Copilot can *generate* new content. For example, it can draft email responses, meeting notes, or policy text based on the knowledge base. This leverages the creative power of LLMs to accelerate writing tasks, albeit still grounded in the source data.
- **Transcription:** Egnyte can transcribe audio or video recordings stored in the system. Copilot will process a meeting recording and provide a text transcript and a summary of the discussion. This brings multimedia into the knowledge fold – now even verbal knowledge is searchable. (www.businesswire.com).

Egnyte documents saved memories and the use of past chats in AI Assistant and Agents. Users can save memories, and administrators can enable or disable personalization and reference to past chats. These are retention and personalization settings to review separately from model-training commitments; this is not a zero-retention workflow (Personalization in AI Assistant and Agents).

Egnyte also provides **domain-specific prompt libraries**. This means companies can customize how Copilot behaves in a given knowledge base. For instance, a healthcare firm might have a prompt library emphasizing regulatory compliance language, whereas a tech firm might use one focused on engineering terminology. Copilot's admin UI lets IT teams upload or select such prompts per base. This fine-tuning ensures the AI's voice and focus align with business needs.

Turning Folders into Knowledge Bases

At its 2024 launch, Egnyte Copilot used designated file groupings called Knowledge Bases. The product is now called **AI Assistant** (overview). Egnyte's migration notice says that, from August 5, 2026, new Knowledge Base creation opens the Agent flow with the Knowledge Base template selected. Use **Create New Agent** and choose that template. Existing Knowledge Bases are being converted domain by domain, preserving their sources, behavior and settings. By October 22, 2026, the notice schedules removal of the separate Knowledge Bases interface and cessation of legacy Knowledge Base API endpoints and MCP tools on migrated domains; integrations need updating (migration notice).

Unlike aftermarket solutions that force organizations to migrate data into a new platform, Egnyte Copilot *works on the existing Egnyte repository*. No data extraction or triple-ownership is needed. This reduces friction – users keep working in the same folders and systems they know, but can now query them with AI.

For example, a sales team could use **Create New Agent**, select the **Knowledge Base** template, and configure its sales content as the source (migration notice). Sales reps can then ask Copilot "What is the renewal date for Acme Corp's contract?" and Copilot will find the Acme contract file, locate the date, and reply. Similarly, a "Legal Policies" knowledge base might answer questions about compliance if the employee asks annotation queries like "What do we do if a data breach occurs?"

This flexibility is powerful. Egnyte's blog recounts how they moved from single-document Q&A to full-folder Q&A to support multi-document scenarios (www.egnyte.com). Initially they used FAISS to vector-search within one document, but realized it couldn't answer queries spanning multiple docs. By marking an entire folder as a base, Copilot can gather relevant snippets from any file in that folder and amalgamate an answer (www.egnyte.com).

Architecture and Implementation (Egnyte RAG)

Egnyte publicly shares the high-level design of Copilot's RAG pipeline in its developer blog series (www.egnyte.com). Key architectural points include:

- **Query Rewriting:** The user's question is first reformulated into different phrasings to improve retrieval. This broadens the search and catches nuances.
- **Hybrid Retrieval:** The query is issued simultaneously to a keyword search index and a semantic embedding index. Egnyte tested various approaches but settled on a hybrid model. They cite research (Sawarkar et al. 2024) showing hybrid search outperforms purely sparse or purely dense methods (www.egnyte.com).

- **Chunking and Indexing:** Large files are broken into smaller text chunks (e.g. paragraphs). Each chunk is embedded into a vector space and also indexed by keywords. Egnyte quantifies the cost: embedding a 1 MB document at 1024 dimensions consumes ~1 MB of storage (www.egnyte.com), so they made tradeoffs in chunk size and embedding strategy.
- **Reranking:** For each query, the system retrieves a set of candidate chunks (potentially up to hundreds or thousands). A cross-encoder neural model reranks these to pick the top-10 most relevant chunks (www.egnyte.com). This helps filter out noise and ensures the LLM gets high-quality context.
- **LLM Answer Generation:** The top-ranked chunks are fed, along with the question, into an LLM prompt. Egnyte reports that they use closed-source models (e.g. via APIs) tuned for Q&A. The LLM outputs an answer and also metadata about which chunk IDs contributed facts.
- **Citation Linking:** Crucially, the system maps each chunk ID back to the original file and location. The final answer text is annotated with links or references to those source files (www.egnyte.com). Egnyte ensures that *"the source of each piece of information is clearly indicated"* in every response (www.egnyte.com). Users thus see the answer inline with footnotes or clickable document names, enabling them to verify facts. If the model is uncertain or an answer seems off, the user can follow the citation to read the original document snippet.

This pipeline is offered as a turnkey solution; Egnyte customers need not manage the complexity themselves. With a Copilot license, files in a knowledge base are continuously ingested and kept indexed. The user experience is virtually identical to chatting with the data: employees ask questions, get chatty replies with "references to their corresponding files" (www.egnyte.com), and can refine queries (for example, clicking "regenerate" to retry the answer if unsatisfied).

By automating this entire flow, Egnyte Copilot lets companies deploy AI assistants *without* building custom data pipelines from scratch. The user skipping "having to build their own complex data pipelines" is a core selling point (www.egnyte.com). In effect, Egnyte provides the content, connectors, indexing engine, LLM integration, and UI as a cohesive knowledge stack.

Governance and Compliance

An enterprise knowledge AI must respect governance policies. Egnyte's AI Safeguards let administrators control which files AI can use in responses and which users can access AI features (AI Safeguards overview). Configure these controls for the intended source scope rather than treating a "no external sharing" folder setting as a blanket prohibition on AI answers.

Even within an organization, administrators can control which content is included. The Microsoft Dynamics post notes that before enabling Copilot, admins should **prepare their content**: e.g., identify overshared or obsolete documents that could produce stale or risky answers (techcommunity.microsoft.com). Egnyte, likewise, will allow admins to exclude certain folders or data types from being queried.

In practice, this means companies use Copilot for well-structured data sets at first. For instance, an M&A team might cut a "Deal Memories" knowledge base of closed-file transactions, rather than allowing Copilot to roam the entire drive. This careful scoping reduces hallucinations and information leakage.

Initial Adoption and Case Study

Egnyte's launch release quotes Alan Matthews, head of IT at **Dalmore Capital**: "Egnyte Copilot was a very practical option for us" (Egnyte launch release).

Integration and Ecosystem

Egnyte Copilot does not exist in isolation. It is part of a broader trend of **AI-enabled enterprise content platforms**. We briefly outline comparable solutions from other vendors:

- **Microsoft 365 Copilot:** The work-data chat comparison in Table 1 concerns Microsoft 365 Copilot, whose architecture uses Microsoft Graph for grounding. The August 1, 2023 post instead announced a preview of SharePoint integration specifically for **Copilot in Dynamics 365 Customer Service**; it should be read as a historical description of that product. **Agents built in SharePoint** were announced generally available on November 19, 2024 (Microsoft announcement). The separately named **Copilot in SharePoint** has a documented general-availability rollout beginning September 30, 2026, with tenant availability progressing after that date (current documentation).
- **Google Workspace (Gemini in Drive):** Google has built an AI assistant (Gemini) into Workspace. The **Gemini in Google Drive** feature enables queries over Drive: "Ask Gemini natural language questions to find quick facts about any project … without needing to find or read the relevant files" (workspace.google.com). Users can also ask for summaries of folders or themes across multiple docs (workspace.google.com). In other words, Google itself is turning Drive into a knowledge base. Google touts that *business users now get over 2 billion AI assists per month* in Workspace, showing high usage (workspace.google.com).
- **Atlassian Confluence Intelligence:** Atlassian added AI features to Confluence (their wiki) in late 2023. Atlassian Intelligence can summarize pages and answer questions "using the content from [Confluence] and Jira" context. It's effectively an on-site RAG for Confluence's docs. For example, it can highlight relevant wiki sections in response to a query. Atlassian calls this "Contextual Q&A." Still, it follows the same pattern: treat Confluence pages as the knowledge base for an internal chatbot.
- **Other Tools:** Startups and niche vendors have emerged exactly in this space. For example, Kyndi and Glean offer "enterprise search" with AI layers; Leena AI markets an "agentic RAG architecture" for corporate search (leena.ai). Apps like Notion AI allow Q&A over Notion docs. Salesforce's Einstein GPT can pull from Salesforce records and corporate knowledge bases. Salesforce's 2023 Slack GPT announcement described plans to build AI natively into Slack (Salesforce announcement). Amazon's Bedrock and Kendra services enable custom knowledge bases on AWS. In short, virtually every large knowledge platform is integrating generative AI to turn stored content into an interactive knowledge assistant. Table 1 compares a few prominent approaches.

Feature / Aspect	Egnyte Copilot	Microsoft 365 Copilot (work-data chat)	Google Gemini (Drive)
Content Sources	Selected Egnyte content configured as Agent sources (migration notice)	User-accessible emails, chats and documents through Microsoft Graph (architecture)	Google Drive (Docs, Sheets, Slides, PDFs) (workspace.google.com)
Core Functions	Natural-language Q&A across files; document summarization; multi-file project summaries; media transcription (www.businesswire.com)	Questions over work data in Copilot Chat; access depends on subscription and enabled sources (source controls)	File search by query; summarization of documents/folders; thematic Q&A (workspace.google.com)
AI Integration	RAG pipeline with vector DB, cross-encoder reranking; LLM (closed) with source citation (www.egnyte.com)	Grounds the prompt through Microsoft Graph, then sends it to the LLM (architecture)	Gemini LLM infused into Drive UI; uses Google's ML (details proprietary).
User Interface	Use **Create New Agent** and the **Knowledge Base** template; former Knowledge Bases appear under **Agents** after the domain migrates (migration notice)	Copilot Chat, with inline citations and a Sources pane (source review)	Chatbox-like UI in Drive/Workspace; Google emphasizes ease ("just type question").
Security via Control	Closed models with no cross-training; external processing and regional terms require separate review (AI Assistant overview) (AI policy)	Access scoped to the signed-in user's permissions (architecture)	Enterprise security controls via Google Cloud; private Workspace data not used to train models. (Gemini Enterprise compliance available) (workspace.google.com)
Source Citations	Yes – responses are annotated with file references; user can "verify sources" (helpdesk.egnyte.com) (www.egnyte.com)	Copilot Chat includes inline source citations and lets users open source files (source review)	In Drive projects, Gemini cites relevant documents when searching beyond pinned sources; citation cards can add documents as project sources (Drive project help)
Deployment	Provided as an add-on feature on Egnyte platform (cloud or on-prem gateway) (helpdesk.egnyte.com)	Work-data access in Copilot Chat requires a Microsoft Copilot subscription (source controls)	Cloud SaaS (requires Google Workspace subscription).
Notable Case / Customer	Dalmore Capital (asset mgmt team retrieving contracts) (www.egnyte.com)	No customer example compared here.	Google cites broad adoption (Air Liquide, Etsy, etc.) but no specific case cited.
Differentiator	Deep integration with existing content; fine-grained prompt libraries per domain (www.businesswire.com) (helpdesk.egnyte.com)	Microsoft 365 app integration and Microsoft Graph grounding (architecture)	Ubiquity and ease (AI in everyday Drive use; multi-file synthesis).

Table 1: Comparison of prominent enterprise AI copilot/knowledge-base solutions. All systems ground answers in organizational data, but differ in platforms and interfaces (sources: Egnyte press/release (www.businesswire.com) (www.microsoft.com) (workspace.google.com)).

Chatbots vs. Retrieval-Augmented Systems

A central question for practitioners is: **When does a generic chatbot suffice, and when is a structured knowledge retrieval system necessary?** That is, could an employee just ask ChatGPT or Claude (possibly fine-tuned on some docs) questions, or do they need a full RAG knowledge base? The answer depends on the use case:

- **General vs Specific Knowledge:** If the query is about *common knowledge* or broad topics (e.g. “Explain Newton’s laws of motion,” or “Draft an email thanking a customer”), a plain LLM suffices. These questions don’t rely on private corpora. For *company-specific* information (“What is our travel reimbursement limit?”, “What did we write in the 2023 Q4 sales presentation?”), the relevant internal knowledge must be supplied in the conversation or through retrieval.
- **Up-to-Date or Regulatory Details:** Supply the relevant current guideline directly in context when it fits, or retrieve it from a maintained index. Choose the approach based on corpus size, context capacity, cost, and update needs. Check ingestion/indexing delays, missed retrievals, cited document versions, and the generated answer. A long-context comparison found that sufficiently resourced long-context models outperformed RAG on average in its evaluation (study).
- **Evidence and Auditability:** In sectors like finance, healthcare, law or security, blind trust in AI is unacceptable. Users need *source citations*. In workflows that require provenance, retrieval-based assistants can connect generated answers to source material. Egnyte explicitly provides “verify sources” links (helpdesk.egnyte.com). Microsoft also emphasizes compliance with internal knowledge, though its UI for Copilot is still evolving.
- **Complex Multi-Document Queries:** If a question requires synthesizing information from multiple documents, RAG is often needed. For large document collections, RAG can select relevant passages from the available sources before generation. For instance, merging insights from three customer contracts may go beyond what a GPT prompt can hold. Egnyte’s RAG can pull chunks from all three contracts seamlessly.
- **Speed & Efficiency:** Paradoxically, chat-only LLMs can sometimes be faster for trivial queries (no retrieval step, just a single model call). For document-based queries, an indexed RAG chatbot adds a retrieval step and can return answers with source references.
- **Control and Customization:** RAG systems allow tailored prompt libraries and domain-specific tuning. Egnyte supports custom question prompts per knowledge base, which lets organizations inject their jargon or context (helpdesk.egnyte.com). General-purpose LLMs can be guided with prompts, while retrieval supplies selected organizational context at query time.

Nevertheless, there are downsides to RAG:

- **Complexity and Cost:** Building and maintaining indexes, embedding pipelines, and handling updates requires significant engineering effort (though platforms like Egnyte abstract much of this). The initial compute cost (e.g. indexing terabytes of docs) can be high. Chatbots are simpler to spin up with an API call.

- **Answer Fluency:** ChatGPT-like models excel at fluent prose and can handle nuance when they *do* know the facts. Sometimes a retrieval-augmented answer can be more stilted or overly literal, depending on prompt. Enterprises mitigate this by combining RAG with strong LLMs for generation, aiming for the best of both worlds.

In practice, many organizations do both: they use a chat-first approach for generic queries or ideation ("brainstorm marketing slogans") and use RAG assistants for factual inquiries. Guidance for users often emphasizes, *"If you need an answer based on company data, use Copilot; for general tasks, ChatGPT is fine"*. Egnyte's phrasing of "Knowledge Work" vs generic work echoes this distinction.

The KG-RAG paper also reported a 71% increase in moderately relevant responses (original paper). Its results concern response relevance, rather than establishing that indexed retrieval is necessary for every document-based task.

Ultimately, select an architecture for the available documents, context limits, costs, update needs, and evaluation results. Verify source versions and answers, and use human review where the stakes warrant it. Egnyte's AI policy explicitly assigns customers responsibility for human review of AI-generated output (AI policy).

A summary table contrasts these modes:

Dimension	Chatbot (LLM only)	Retrieval-Augmented (RAG)
Knowledge Source	Model context and information supplied in the conversation	Company documents supplied through an index
Answer Context	Uses the information available to the model and conversation (www.coveo.com)	Uses retrieved company documents; citations provide source context (www.egnyte.com)
Current Information	Can use current information supplied in the conversation	Can use files available in the indexed repository
Source Context	Depends on the sources supplied in the conversation	Can include explicit links to file IDs or pages (www.egnyte.com)
Implementation Complexity	Simple (just use an LLM API)	Complex (need content pipeline, search, vector DB, etc.)
Latency / Cost	Lower initial setup; per-query LLM cost	Higher initial ingestions; retrieval + LLM cost (often mitigated by smarter filtering)
Best For	Ideation, drafting, Q&A on general knowledge	Factual Q&A on corporate data, support, research, where accuracy is critical

Table 2: Comparing plain Chatbots versus Retrieval-Augmented Generation systems in enterprise settings (see text) (www.coveo.com) (www.egnyte.com).

Data Analysis and Industry Perspective

Evidence of the **value** of this new stack is emerging. We summarize key data and expert insights:

- **Adoption Metrics:** McKinsey's May 2024 report found that **65% of respondents said their organizations regularly used generative AI in at least one business function** (early-2024 survey). Its March 2025 report describes increased reported AI use during 2024 (subsequent survey). TechTarget reported that by the end of 2024, about one-third of businesses had put generative AI into production, with even more planning pilots (www.techtarget.com). These figures suggest generative AI (and by extension, AI-enhanced knowledge systems) is no longer experimental – it is mainstream.
- **Benefits Realized:** Surveyed organizations report tangible improvements: faster decision-making, increased efficiency, and better customer experience. McKinsey notes that business units deploying generative AI saw both cost reductions and revenue gains (www.mckinsey.com). One blog notes a reported **3.7× ROI per dollar spent on generative AI** (www.netguru.com). While exact breakdowns are proprietary, many companies cite significant time savings from AI-augmented search. For example, former Microsoft execs claim knowledge workers spend ~20% of their time searching for information; AI offers to cut that dramatically.
- **Emerging Challenges:** Surveys also highlight obstacles. TechTarget's enterprise AI survey (Oct 2024) found that while AI adoption climbed, **data quality, lack of expertise, and regulatory issues** remain top hurdles (www.techtarget.com). Egnyte separately documents AI activity logs showing what content AI accessed, what was restricted from generation, and when (AI Safeguards). Do not treat answer citations as evidence that applicable regulatory audit-trail requirements have been met; assess those requirements and controls separately.
- **Vendor Landscape:** Analysts identify a burgeoning market for "AI knowledge management" tools. For example, Gartner's *Peer Insights* lists dozens of generative AI knowledge management apps (including IBM Watson, MS Copilot, Google Gemini, smaller AI search startups, etc.). *KMWorld AI 100 (2025)* features many companies (like Atlassian, Box, Coveo, etc.) as "empowering intelligent KM" (www.kmworld.com). This reflects broad recognition that content+AI is the next frontier.
- **Expert Commentary:** Industry blogs and papers emphasize the technical improvements of RAG. For instance, a corporate AI blog states: "One key to achieving output accuracy is using a graph-based RAG system for company-specific knowledge retrieval" (writer.com). Another review by Coveo highlights that RAG can drastically reduce irrelevant or fabricated answers by grounding the model in the correct context (www.coveo.com). These expert sources advocate combining AI models with curated knowledge for best results.
- **Customer Case Studies:** Beyond Egnyte's cited example, public case studies are still emerging. Early adopters have shared anecdotes: an insurance company using a chat interface on their intranet, an IT helpdesk using RAG to cut ticket volume, etc. Many are protected by NDAs, but startups frequently blog about generic "large retailer" or "enterprise client" success stories. What is clear is that in siloed domains (e.g. banking, biotech R&D, manufacturing manuals), AI knowledge assistants are gaining traction as a productivity tool.

Broader Implications and Future Directions

The shift toward AI-driven knowledge bases has several significant implications:

- **Workforce Productivity:** By reducing the time to find information, these systems can boost employee productivity. Employees no longer need to know exactly where information lives or how to phrase a search; they simply ask the question. This flattens knowledge curves and democratizes expertise. However, organizations must manage change: employees need to trust and adopt AI tools. Early feedback indicates minimal user training is needed with Copilots (www.businesswire.com), since the interface is intuitive like chatting.
- **Knowledge Preservation:** An AI knowledge base can help preserve institutional memory. When key employees retire or leave, their know-how (if documented) remains queryable. For example, an experienced engineer's design notes tucked in a shared folder could still be accessible via AI prompts. Egnyte describes indexing chunk offsets alongside embeddings (engineering account); avoid assuming that this process constructs or curates an entity-and-relation knowledge graph.
- **Bias and Hallucination Mitigation:** RAG helps mitigate model bias by constraining answers to documented facts. Yet hallucinations can still occur (e.g. if the query yields no exact content, the LLM may hallucinate to fill gaps). Enterprises must remain vigilant: some Copilot implementations offer conservative responses ("I don't see that information" if none found). Continuous evaluation and user feedback loops are important. Research is evolving: e.g. Adobe's KG-RAG work (preprints.kavlimeetings.org) tries to further reduce hallucinations with knowledge graph filtering and confidence scoring.
- **Data Security and Privacy:** Entrusting AI with sensitive data raises new risks. Even if the model doesn't train on the data, careless prompts could expose confidential info. Hence, the deployment includes governance (as discussed), encryption, monitoring, and in some cases use of on-prem LLMs for extra control. Companies are also considering compliance for personal data (GDPR, HIPAA) within AI queries. This is an active legal and technical area; for instance, Microsoft's Copilot Terms clarify how data is handled in chat.
- **The Evolution of Search:** Enterprise search may be subsumed by these AI tools. Instead of scrolling result pages, users click on an inbox-like chat, indicating a fundamental UX shift. Traditional search logs may give way to conversation transcripts. Enterprises may redefine metrics of search success (accuracy of answers, speed to correct response) rather than recall/R-precision.
- **Future Features:** We expect richer integration. Multimodal RAG will become standard: not just text, but also retrieving from diagrams, images, databases, even real-time sensors. Egnyte's work on extracting structures shows one path. Some predictions include virtual "agents" that proactively surface relevant knowledge (e.g. AI reminding a sales rep about a contract term during a call). Also, as models improve, we may see self-optimizing pipelines that refine embeddings or prompts automatically, reducing human oversight.
- **Competitive Differentiation:** For knowledge-centric businesses, AI-powered knowledge bases become a strategic asset. The speed of internal decision-making and innovation can confer competitive advantage. Companies investing early in building clean, AI-friendly knowledge repositories (well-organized, tagged, and current) will find their employees able to tap into institutional knowledge faster. Conversely, those who lag may find their data trapped in silos.

Conclusion

In conclusion, the **new enterprise knowledge stack** – integrating file repositories, RAG pipelines, and copilot assistants – represents a paradigm shift in how organizations access their own information. We have shown that solutions like Egnyte Copilot enable rich, source-backed, AI-driven Q&A over corporate content (www.egnyte.com). These tools promise to dramatically cut search times, improve learning curves, and extract value from previously dark data. Crucially, they combine conversational ease with the precision of traditional retrieval, addressing the practical quandary of “when chat is enough versus when retrieval is needed”.

Empirical evidence and case studies affirm the gains: companies leveraging these systems report faster information access and productivity boosts (www.egnyte.com) (www.coveo.com). Industry surveys indicate widespread acceptance of generative AI, especially as companies solve concerns about data quality and trust through grounded answers (www.techtarget.com) (www.businesswire.com).

Looking ahead, evaluate knowledge assistants against representative tasks and current source documents. Choose direct document context, indexed retrieval, or a combination based on corpus size, context capacity, update needs, cost, and measured performance. Research comparing RAG with long-context models identifies lower cost as a RAG advantage (study). Workshop guidance should teach source verification and human review for consequential answers regardless of architecture.

Ultimately, organizations that successfully align their content strategy with AI capabilities will unlock a new “golden knowledge” that was previously locked in silos. **The file share becomes the knowledge share.**

References: Egnyte Copilot press release and documentation (www.egnyte.com) (www.businesswire.com) (helpdesk.egnyte.com); Egnyte technical blogs (www.egnyte.com); Microsoft Dynamics blog on Copilot (www.microsoft.com); Google Workspace announcements (workspace.google.com); McKinsey survey (www.mckinsey.com); Netguru AI adoption stats (www.netguru.com); TechTarget enterprise AI survey (www.techtarget.com); Coveo generative AI blog (www.coveo.com); KMWorld AI100 article (www.kmworld.com); Salesforce blog on AI grounding (www.salesforce.com); Egnyte user guide and knowledge-base articles (www.egnyte.com) (helpdesk.egnyte.com) (All sources are hyperlinked above.)

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.egnyte.com/press-release/egnyte-launches-generative-ai-based-copilot-to-speed-up-knowledge-work-and-unlock-enterprise-value-from-content
2. https://www.businesswire.com/news/home/20240708375745/en/Egnyte-Launches-Generative-AI-based-Copilot-to-Speed-up-Knowledge-Work-and-Unlock-Enterprise-Value-From-Content
3. https://helpdesk.egnyte.com/hc/en-us/articles/49308831600141-Migration-of-Knowledge-Bases-to-Agents
4. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024
5. https://www.coveo.com/blog/enterprise-knowledge-retrieval
6. https://arxiv.org/html/2502.15237v1
7. https://www.egnyte.com/blog/post/part-2-how-egnyte-built-its-turnkey-retrieval-augmented-generation-solution

8. https://www.microsoft.com/en-us/dynamics-365/blog/it-professional/2023/08/01/use-copilot-with-sharepoint-knowledge-to-serve-your-customers
9. https://workspace.google.com/products/drive/ai
10. https://helpdesk.egnyte.com/hc/en-us/articles/28274380320397-Egnyte-s-AI-Assistant-Overview
11. https://techcommunity.microsoft.com/t5/sharepoint-premium-blog/announcing-sharepoint-advanced-management-innovations-for-the-ai/ba-p/4126366
12. https://www.salesforce.com/blog?p=91094
13. https://www.egnyte.com/blog/post/part-1-how-egnyte-built-its-turnkey-retrieval-augmented-generation-solution
14. https://www.egnyte.com/blog/post/beyond-plain-text-egnytes-journey-to-structured-data-extraction-in-rag-systems
15. https://www.egnyte.com/subcontractors
16. https://www.egnyte.com/generative-ai-policy
17. https://helpdesk.egnyte.com/hc/en-us/articles/28274380320397-Copilot-Overview
18. https://helpdesk.egnyte.com/hc/en-us/articles/43288758481293-AI-Safeguards-Overview
19. https://www.kmworld.com/Articles/Editorial/Features/2025-The-KMWorld-AI-100-The-Companies-Empowering-Intelligent-Knowledge-Management-169468.aspx
20. https://helpdesk.egnyte.com/hc/en-us/articles/48824437797389-Personalization-in-AI-Assistant-and-Agents
21. https://www.egnyte.com/press-releases/egnyte-launches-generative-ai-based-copilot-speed-knowledge-work
22. https://techcommunity.microsoft.com/blog/microsoft-copilot-blog/ignite-2024-agents-built-in-sharepoint-now-in-general-availability/4298746
23. https://learn.microsoft.com/en-us/sharepoint/copilot-in-sharepoint-get-started
24. https://workspace.google.com/blog/product-announcements/new-ai-drives-business-results
25. https://leena.ai/knowledge-management
26. https://www.salesforce.com/news/press-releases/2023/05/04/slack-gpt-news
27. https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-architecture
28. https://support.microsoft.com/en-us/microsoft-365-copilot/control-review-sources-copilot-chat
29. https://support.google.com/drive/answer/16684520?hl=en
30. https://arxiv.org/abs/2407.16833
31. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value
32. https://www.techtarget.com/searchEnterpriseAI/feature/Survey-Enterprise-generative-AI-adoption-ramped-up-in-2024
33. https://www.netguru.com/blog/ai-adoption-statistics
34. https://www.egnyte.com/products/ai-safeguards
35. https://writer.com/guides/generative-ai-survey
36. https://preprints.kavlimeetings.org/2025/02/21/from-documents-to-dialogue-building-kg-rag-enhanced-ai-assistants

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.