



glean

How to Give Your Workforce AI-Powered Answers Without Compromising Security

Executive Resource
2026

Introduction



Understanding how RAG-based platforms connect siloed systems while enforcing your existing permissions

Many agencies have massive document repositories but no practical way for staff to find what they need across siloed systems.

Decades of digital records. Terabytes of policy documents, case files, and institutional wisdom scattered across SharePoint, Google Drive, email archives, case management systems, ticketing systems, and a dozen other systems.

When an analyst needs an answer, like What's our precedent on this case type? Which waiver applies here? Has this vendor been vetted before?—they often spend 20% of their day hunting for it.

Not because the information doesn't exist, but it's buried in the wrong system, labeled with the wrong keywords, or locked in the head of someone who retired three years ago.

Why Search Breaks Down in Practice

The problem isn't that your search tools are old. It's that they're designed for a different job.

Keyword Search Gives You Links. Mission Work Requires Answers.

You type "parental leave policy" and get 50 PDFs ranked by keyword frequency. Now you have a new job: open each file, skim for relevance, cross-reference versions, and synthesize the answer yourself.

This is hunting, not searching. And it eats hours.

The "Walled Garden" Trap

Many agencies run hybrid environments. Some teams on Microsoft 365, others on Google Workspace. Critical data in case management systems or project management tools. Case histories in legacy systems.

But Microsoft Copilot only sees Microsoft. Google's Gemini only sees Google.

Your mission doesn't live in one app. If your search does, you're missing most of the picture.

Shadow AI: When Desperation Beats Policy

Here's the uncomfortable part: Your workforce is already using AI. They're pasting case details, policy drafts, and constituent data into public LLMs like ChatGPT, Claude, or Gemini just to get a summary or draft a response faster.

This isn't malicious. It's survival. They're optimizing for speed because the approved tools are too slow.

You can block these tools and add friction, or you can give them a better internal alternative that respects your governance model.

The Shift: From Search to Intelligence

The Shift: From Search to Intelligence

What if instead of searching for documents, your workforce could ask questions and get answers? Answers grounded in your agency's records, respectful of your permissions, and cited to the source?

This requires a different architecture. One that separates the reasoning engine from the data.

The Brain vs. The Memory

Think of modern AI in two parts:

The Brain (Large Language Model)

The reasoning engine. It understands natural language, synthesizes information, and generates coherent responses. This is the consumer LLM part: tools like ChatGPT, Claude, or Gemini that are powerful at language, but generic. It knows the internet, not your agency.

The Memory (Your Agency's Records)

The specific facts, policies, precedents, and institutional knowledge that make your agency unique. This is what the brain needs access to in order to be useful.

The mistake many teams make is thinking they need to train a custom brain. They don't. The public LLMs are already excellent at reasoning. What they lack is the right memory.

This is where **Retrieval Augmented Generation (RAG)** comes in.

How RAG Works: Grounding AI in Agency Truth

RAG connects a generic AI brain to your specific agency memory, without moving your data, without training a custom model, and without breaking your security posture.

Here's what that looks like in practice:

Step 1: Index Your Knowledge (Without Moving It)

A RAG system connects to your existing systems (SharePoint, Google Drive, email, project management tools, case management systems, IT service management systems) and builds a unified index. Not a data lake where you copy everything. An index that maps what exists, where it lives, and who can see it.

Think of it as a card catalog for your entire agency, updated in real time.

Step 2: Enforce Your Permissions

The index mirrors your existing Access Control Lists (ACLs). If an employee doesn't have permission to open a document in SharePoint, they can't retrieve it (or generate answers from it) in the AI system.

Permissions aren't bolted on after the fact. They're enforced at the index level.

Step 3: Retrieve Relevant Context

When a user asks a question like "What's our policy on telework for new hires?" the system searches the index, retrieves the most relevant documents (policy memos, email threads, HR guides), and ranks them by recency, authority, and relevance.

This happens in seconds. Across all your systems. Respecting all your rules.

Step 4: Generate a Cited Answer

The system feeds the retrieved context into the AI brain and asks it to synthesize an answer. The result isn't a list of links. It's a coherent response with inline citations pointing back to the source documents, page numbers included.

The user gets an answer. The agency gets auditability. The CISO gets proof that permissions were respected.

What This Unlocks (When It Works)

If you're responsible for workforce productivity, knowledge continuity, or AI governance, this architecture might be useful. It tends to unlock four distinct outcomes:

1. Workforce Velocity: Stop the Search Struggle

New hires ramp up faster. Analysts spend their time analyzing, not hunting for files. Senior staff stop being "human routers" for every policy question.

In practice, agencies report 40–60% reduction in time spent searching. For a 5,000–person agency at \$65/hour fully loaded cost, that's \$50M+ in reclaimed productivity annually.

2. Governed Sovereignty: Safe AI That Knows Your Rules

Your data never leaves your environment. The AI respects your permissions, your classification levels, and your retention policies. You eliminate Shadow AI risk by giving your workforce a better internal alternative.

A common outcome: 90% reduction in unauthorized AI tool usage—not because you blocked access, but because you offered something better.

3. Agnostic Unification: The Mission Doesn't Live in One App

Many knowledge platforms only connect to one vendor's ecosystem. RAG-based systems can connect to 100+ enterprise systems. Productivity suites, cloud storage platforms, case management systems, project management tools, IT service management systems, HR platforms, and so on—all indexed in one place.

Your team asks one question and gets answers from everywhere, without logging in and out of a dozen apps.

4. Institutional Memory: Wisdom Shouldn't Retire When People Do

When your senior staff retire, their expertise doesn't have to vanish. It's captured in the emails they sent, the documents they authored, and the decisions they influenced. A well-indexed system makes that tacit knowledge explicit and searchable.

This is where succession planning becomes knowledge continuity, not crisis management.

Why RAG Tends to Be Safer Than Fine-Tuning

Some vendors pitch **fine-tuning**: training a custom AI model on your agency's data. "An AI trained just for you" often sounds very appealing at the surface, but it introduces risks that are hard to govern:

Your data gets baked into the model. Once it's trained in, you can't delete it. You can't audit what it learned. You can't enforce dynamic permissions on a per-user basis.

Fine-tuned models still hallucinate. Now they do it with the confidence of your agency's voice, which makes false information harder to detect.

Expensive and brittle. Every time your policies change, you need to retrain the model. Every time you hire someone new or someone changes roles, their permissions are static in the model.

RAG is different:

Your data stays in your systems. The AI reads it on-demand, respecting live permissions. If you revoke someone's access to a file, they can't generate answers from it anymore.

Answers are cited. Every claim is traceable to a source document. Hallucinations are easier to catch because there's a citation trail.

Always current. When you update a policy in SharePoint, the next query reflects the change. No retraining required.

RAG separates the brain from the memory, and that separation is what makes it governable.

Three Common Applications

Three Common Applications

These examples show how permissions-aware AI search works across different agency workflows. Your environment will have its own high-value use cases.



Policy Research

Question: "What are the eligibility requirements for telework under the 2024 policy?"

Traditional approach: Employee searches SharePoint, gets 50+ PDFs, opens 10, skims each, cross-references, manually writes summary.
Estimated time: 2 hours.

RAG-powered system: Returns a synthesized answer with citations to HR handbook (page 12), union agreement (section 4.3), and the December 2024 memo. **Estimated time: 15 seconds.**

The difference isn't just speed. It's completeness. The system read all 50 documents. The human only got through 10.



Contract Compliance

Question: "Does this vendor proposal comply with our procurement regulations?"

Traditional approach: Employee reviews procurement manual, checks previous contracts, emails procurement officer, waits for response.
Estimated time: 4+ hours.

RAG-powered system: Retrieves procurement regulations, flags non-compliant clauses, cites specific regulation sections. **Estimated time: 2 minutes.**

Again, the value isn't just time. It's that the system checked every relevant regulation. A human under time pressure might miss one.



Case Precedent

Question: "How have we handled similar disability accommodation requests in the past 5 years?"

Traditional approach: Search case management system, read through dozens of case files, interview case managers, manually summarize patterns.
Estimated time: 6+ hours.

RAG-powered system: Retrieves 47 similar cases, summarizes common outcomes, highlights key precedents with citations. **Estimated time: 30 seconds.**

This is where institutional memory becomes tangible. The knowledge of every case manager who ever handled a similar case becomes accessible to everyone.

Where RAG Projects Stall (And How to Plan Around It)

The technical architecture is the easy part. Getting it through procurement, funded, and adopted is where most initiatives hit friction. Here's what tends to slow things down.

Procurement Doesn't Have a Category for This

Your procurement office has contracts for productivity suites, security tools, and databases. But "AI-powered knowledge retrieval with permissions enforcement" doesn't fit neatly into an existing category.

The default instinct is often a new RFP, which can mean 6-12 months (or more) of process before you can even start, not counting the time to write requirements.

Faster paths exist that are just as rigorous:

Many enterprise platforms have already been competitively procured through cooperative contract vehicles like GSA Schedules, NASPO ValuePoint, or OMNIA Partners. If a platform is already on one of these vehicles, you may be able to leverage that existing competition rather than running your own full RFP process. This can cut months off your timeline while maintaining fair and competitive procurement standards.

Another option: Run a well-structured RFQ (Request for Quotation) instead of a full RFP. RFQs are faster, less resource-intensive, and still allow you to evaluate multiple vendors on technical merit and pricing. For emerging technology categories like this, an RFQ often gives you more flexibility to learn from vendors during the process.

If you need help structuring requirements, [RedLeif](#) maintains a library of public sector procurement resources, including model RFQs and platform evaluation checklists, that can serve as a starting point. They also work with agencies to build category-specific requirements. (Full transparency: RedLeif publishes vendor-neutral procurement tools, not vendor-specific sales materials.)

Your Data Sources Aren't Ready to Connect

RAG platforms need API access to your systems. Many legacy systems don't have modern APIs. And even when they do, your IT security team may not allow third-party tools to connect without extensive review.

If your SharePoint is locked down, your case management system is air-gapped, or your email archive requires manual exports, the technical lift to connect everything can be substantial.

What helps: Start with the systems that are easiest to connect. If your Microsoft 365 and Google Workspace environments are accessible via standard APIs, connect those first. Prove value with what's available, then tackle harder integrations once you have executive sponsorship.

Where RAG Projects Stall (And How to Plan Around It)

Change Management Gets Skipped

You can deploy the best RAG platform in the world, but if your workforce doesn't know it exists or doesn't trust it, adoption flatlines.

In our experience, the agencies that succeed treat this as a change management initiative, not a technical deployment. They train power users, run internal champions programs, and communicate relentlessly about what the system can and can't do.

What helps: Identify 10–20 influential knowledge workers who are already frustrated with search. Give them early access. Let them evangelize to their peers. Build adoption from the bottom up, not top down.

Expectations Aren't Calibrated

Some stakeholders expect the system to answer every question perfectly on day one. They test it with edge cases, find limitations, and declare the project a failure before it's fully ramped up.

The real constraints aren't usually accuracy; they're **coverage and timing**.

Coverage: If you've only connected 2 of your 10 core systems, the platform can only answer questions about the data it can see. A question that requires information from an un-indexed system will come back empty, not wrong. This looks like a failure, but it's actually incomplete implementation.

Timing: These systems get better as they index more content, learn usage patterns, and surface the most relevant sources. If you evaluate on day one with a partially indexed environment, you're not seeing the actual capability.

What helps: Set realistic expectations upfront. Baseline your current search experience first. If your existing tools force employees to hunt for 20–30 minutes per query (or give up entirely), and the new system delivers cited answers in seconds—even if it can't answer every question yet—that's transformational. The goal isn't perfection on day one. It's measurable time savings and eliminating dead-end searches.

Next Steps

Let's Make Knowledge Actionable

Your agency's greatest asset isn't the data you collect. It's the knowledge you can act on.

Across the public sector, agencies are dealing with the same challenge: knowledge is fragmented, siloed, and inaccessible. If your organization is like most, your workforce is likely spending hours each week searching for information... or worse, turning to risky public AI tools just to get work done faster.

Glean is built specifically to solve this. It's a governed, RAG-powered knowledge platform that turns your scattered records into trusted answers, without requiring data migration, custom model training, or compromising on security.

If you're responsible for productivity, AI governance, or knowledge continuity, Glean changes the tradeoff between speed and security. You don't have to choose between fast answers and safe governance anymore.

Ready to Reclaim Your Agency's Intelligence?

Next Steps:

- Learn more about Glean and enterprise AI search capabilities on [The Trading Post](#)
- Interested in a Proof of Concept? Visit Glean's Trading Post page and reach out. Mention you're interested in exploring a PoC for your organization
- Download procurement resources including evaluation checklists and model RFQs at redleif.io

Explore how Glean is supporting the public sector.



glean

2026