



SERVICE GUIDE · 2026

Retrieval Augmented AI

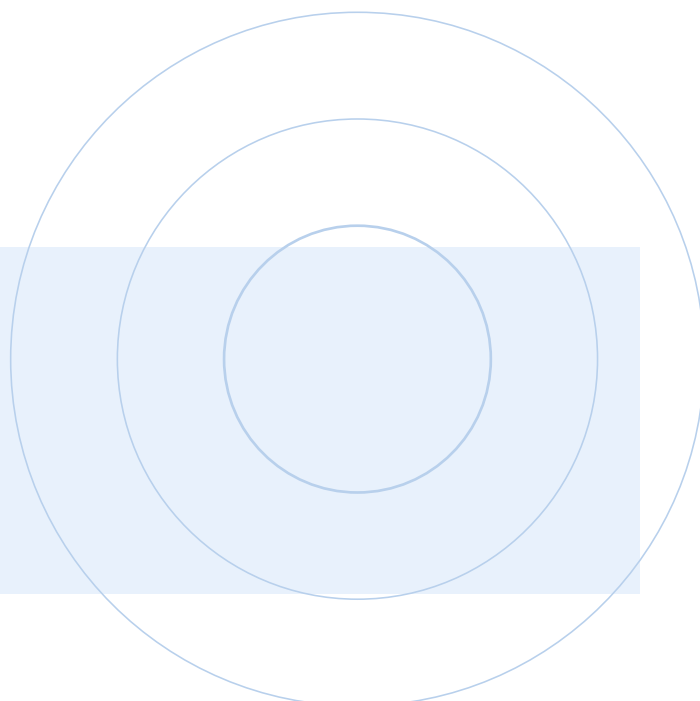
AI that answers from your own documents
— accurately, and with proof.

INSIDE THIS GUIDE

- 01 What it is
- 02 What it can do
- 03 Industry use cases
- 04 A closer look
- 05 Approaches & architecture
- 06 What we build
- 07 Oversight & Governance
- 08 Hosting & Deployment
- 09 Consulting & Advisory
- 10 Training & Support
- 11 Our Delivery Process

ENTRA AI

www.entra-ai.com
info@entra-ai.com
+44 2034 323178



WHAT IS RAG

AI grounded in your knowledge

RAG — Retrieval-Augmented Generation — is AI that **looks up information in your own documents** before answering. Instead of relying on general knowledge or making things up, it grounds every response in your real, current data — and can show you exactly where the answer came from.

THINK OF IT LIKE THIS

Imagine giving an expert assistant access to your entire filing cabinet — every policy, every report, every email. When you ask a question, they don't guess. They open the relevant file, read it, and answer based on what's actually there — pointing you to the source if you want to check.

01**How It Works**

A plain-English overview

- **Your Knowledge Is Indexed**

We organise your documents, policies, and data so the AI can search and find the right information instantly.

- **The AI Looks Up First**

When asked a question, the AI retrieves the most relevant passages from your knowledge — before forming an answer.

- **You Get Grounded Answers**

Responses are based on what's actually in your documents, with sources attached so you can verify everything.

02**What It Can Actually Do**

Real-world examples

- **Internal Knowledge Bots**

Give staff instant, accurate answers from your policies, processes, manuals, and internal documentation.

- **Policy & Compliance Lookup**

Let legal, HR, or compliance teams instantly find the right clause, rule, or precedent across thousands of documents.

- **Onboarding Assistant**

Help new joiners get up to speed by answering their questions from your handbook, wiki, and training materials.

- **Customer-Facing FAQ**

Offer customers reliable answers based on your real product information — drastically reducing simple support tickets.

- **Research & Summarisation**

Summarise long reports, papers, or regulatory documents — with citations to the exact sections used.

- **Document Q&A;**

Let anyone ask natural-language questions of contracts, reports, or technical documents — and get accurate, sourced answers.

INDUSTRY USE CASES

Real scenarios, real outcomes

Below are examples of how Retrieval-Augmented AI (RAG) is being applied across different sectors and departments. Each one replaces hours of manual work or unlocks insight that was previously out of reach.

A**Internal Knowledge & Productivity**

Find what your organisation already knows

- **Employee Help Desk**

Staff ask IT, HR, or policy questions in plain English and get accurate answers drawn from your real documentation — with the source linked.

- **Operational Playbooks**

Frontline teams query process documents and runbooks naturally — finding the right step instantly even in unusual situations.

- **Sales Enablement**

Salespeople get instant answers to product, pricing, or competitive questions from your latest internal materials — no more searching SharePoint.

- **Engineering & Technical Docs**

Engineers ask questions of API docs, technical specs, and incident histories — pulling the right answer out of thousands of pages.

B**Customer-Facing Use Cases**

Better service, lower cost, higher accuracy

- **Self-Service Support**

Customers ask questions and get reliable answers from your help centre and product documentation — deflecting tier-one tickets.

- **Onboarding Guides**

New customers are guided through setup with an AI assistant that answers from your real onboarding documentation and best practices.

- **Product Q&A;**

On your website or product, customers can ask detailed questions and get accurate, source-cited answers from your real materials.

- **Account & Policy Lookup**

Customers query their own data — invoices, contracts, policies — with the AI grounding answers in their actual account information.

C

Regulated & Specialised Use Cases

Where accuracy and audit trail are non-negotiable

• Legal & Contract Review

Lawyers and compliance teams query contracts, regulations, and case law — getting cited answers from documents they can verify.

• Medical & Clinical Reference

Clinicians can consult guidelines, protocols, and research summaries with full source citations — always as a support tool.

• Regulatory Compliance

Stay current with constantly-changing regulations — the AI surfaces the right rule, section, and effective date with confidence.

• Financial Research

Analysts ask questions of filings, reports, and earnings transcripts — getting summarised, sourced answers across vast document sets.

A CLOSER LOOK

How RAG actually works

RAG combines two strengths: precise retrieval from your real documents, and clear, natural answer generation. The result is AI that's both knowledgeable and verifiable.

• 1. Ingest

Your documents are gathered, cleaned, and prepared — text extracted, structure preserved, metadata captured.

• 2. Index

Content is indexed in a way that makes searching by meaning fast and accurate — not just by keyword.

• 3. Retrieve

When a question is asked, the system finds the most relevant passages from across your indexed knowledge.

• 4. Ground

Retrieved passages are passed to the AI as context — anchoring the answer in your real, current information.

• 5. Generate

The AI produces a clear answer based on the retrieved context — staying close to what the source actually says.

• 6. Cite

Sources are returned alongside the answer — so users can verify, trust, and dive deeper when needed.

RAG APPROACHES

The different ways RAG can be built

RAG isn't one fixed thing — it's a family of approaches. The right architecture depends on what you're retrieving, how your data is structured, and how complex the questions are. Here are the main approaches we use, and when each one fits.

D

How Knowledge Is Stored

Vector databases, graph databases, and traditional search

• Vector Databases

Documents are converted into numerical "meaning fingerprints" and stored in a vector database (e.g. Milvus, pgvector, Pinecone). The system can then find passages by meaning, not just keywords — finding the right answer even when the wording differs.

• Keyword & Full-Text Search

Traditional search indexes (e.g. Elasticsearch, OpenSearch) remain excellent for exact matches, codes, names, and structured filters. Often the right tool when precision matters more than semantic understanding.

• Structured Data Sources

RAG can also draw from databases, spreadsheets, and APIs — not just documents. The AI generates the right query, retrieves live data, and grounds its answer in current facts from your operational systems.

• Graph Databases

Knowledge is stored as connected entities and relationships in a graph database (e.g. Neo4j). Ideal when answers depend on links — who reports to whom, what's connected to what, how one fact relates to another.

• Hybrid Storage

We frequently combine all three — vectors for meaning, graphs for relationships, full-text for exact matches. Each approach covers the others' weaknesses, giving you robust retrieval across question types.

• Multimodal Knowledge

Beyond text — images, diagrams, tables, charts, and audio can all be indexed and retrieved. Useful for technical manuals, medical references, or any knowledge where visuals carry meaning.

E

Advanced RAG Patterns

When simple retrieval isn't enough

• Hybrid RAG

Combines vector search, graph traversal, and keyword search into a single retrieval pipeline — then re-ranks results to surface the most relevant passages. Dramatically improves accuracy over any single-method approach.

• Query Decomposition

Complex questions are broken into smaller sub-questions, each retrieved and answered separately, then combined. Essential for compound questions like "compare X and Y across regions over time".

• Citation & Confidence Scoring

Every answer carries source citations and a confidence score. Low-confidence responses are flagged for human review rather than presented as fact — keeping users in control of important decisions.

• Agentic RAG

Pairs RAG with agentic AI: the system doesn't just retrieve once and answer. It plans multi-step research, retrieves iteratively, cross-references sources, and takes actions on what it finds — closer to a research assistant than a search engine.

• Re-ranking & Filtering

An initial wide retrieval is refined by a second model that reads each candidate and ranks them by true relevance — catching cases where surface similarity doesn't mean the right answer.

• Continuous Evaluation

We run automated evaluations on a sample of real questions — measuring retrieval accuracy, answer faithfulness, and citation correctness — so quality is tracked, not just hoped for.

06

What We Build For You

The key components of your system

- **Document Ingestion Pipelines**

Automated pipelines that pull in your documents from wherever they live — file shares, intranets, databases, document management systems.

- **Grounded Answer Generation**

AI that uses retrieved content to construct accurate, well-written answers — with sources cited every time.

- **Freshness Management**

Automatic re-indexing when documents change — so your AI always answers from the latest version, not yesterday's draft.

- **Search & Retrieval Engines**

Highly accurate retrieval systems that find the right passages instantly — even across millions of documents.

- **Permission-Aware Access**

Users only see answers from documents they're authorised to access — respecting your existing security and confidentiality rules.

- **Feedback & Improvement**

User feedback loops that surface gaps and inaccuracies — letting you continuously improve quality where it matters.

07

Oversight, Governance & Responsible AI

Built-in accountability — always included

- **Full Visibility**

See exactly what your Retrieval-Augmented AI (RAG) system is doing — every action, decision, and outcome — through a clear activity log.

- **Audit Trails**

Every action recorded with timestamp and reason — a searchable record for compliance and review.

- **Continuous Improvement**

Your system gets better over time. We review outcomes, identify improvements, and update it.

- **Guardrails & Boundaries**

Clear rules for what your system can and cannot do. Sensitive actions always require human sign-off.

- **Ongoing Monitoring**

We monitor performance over time, flagging anything unusual and ensuring continued accuracy.

- **Data Privacy & Security**

Your data stays yours. We follow rigorous practices and work within your security policies.

08

Hosting & Deployment

We handle the technical side — you focus on the results

Once your Retrieval-Augmented AI (RAG) system is built, it needs somewhere to live and run reliably. We take care of all of that for you.

- **Cloud Hosting**

Hosted on secure, enterprise-grade cloud infrastructure. Your Retrieval-Augmented AI (RAG) runs 24/7, scales with usage, maintained by our team.

- **Hybrid Setup**

Hybrid deployments combining cloud convenience with on-site control, tailored to your requirements.

- **On Your Own Systems**

If data must stay on-premises for compliance or security, we deploy your system entirely within your environment.

- **Ongoing Management**

Managed packages include monitoring, updates, performance tuning, and ongoing support as your system matures.

09

Consulting & Advisory

We help you figure out what you actually need

Not sure where to start? Most organisations know Retrieval-Augmented AI (RAG) could help — but not where, how, or what to prioritise. We work through that with you clearly, without jargon.

- **Discovery & Scoping**

We understand your operations, challenges, and goals — then identify where Retrieval-Augmented AI (RAG) delivers the most value first.

- **Roadmap Planning**

A phased plan — starting with a focused pilot that delivers early results, then expanding as confidence and ROI grow.

- **Feasibility & Risk Assessment**

A clear view of what's achievable, the risks involved, and a realistic timeline. No overpromising, no jargon.

- **Stakeholder Alignment**

We help communicate the plan internally — leadership, operations, compliance, IT — with materials for non-technical audiences.

10

Training & Adoption Support

Making sure your team feels confident

The best system only creates value if your team knows how to use it, trust it, and get the most from it. We provide tailored training and support to make adoption smooth.

- **Team Onboarding**

Practical, jargon-free sessions for people working alongside the Retrieval-Augmented AI (RAG) system every day — what it does and when to step in.

- **User Guides & Playbooks**

Clear, plain-English documentation tailored to your system — a reliable reference whenever your team needs it.

- **Leadership Briefings**

Focused sessions for senior stakeholders covering outcomes, governance, risk, and measuring success.

- **Ongoing Support**

As confidence grows and workflows evolve, we're available for follow-up sessions, Q&A, and improvements.

11

Our Delivery Process

From first conversation to live system

- **01 — Discovery Call**

We learn about your business, challenges, and where we can help most.

- **02 — Scoping & Proposal**

We map the solution, confirm feasibility, and agree on clear outcomes.

- **03 — Build & Test**

We build your system, test it thoroughly, and refine it with your input.

- **04 — Deploy & Integrate**

We go live — connecting it to your tools on your chosen infrastructure.

- **05 — Train Your Team**

Onboarding sessions and documentation so your team is ready to use it.

- **06 — Monitor & Improve**

We track performance, maintain governance, and keep improving over time.



NEXT STEP

Ready to put AI to work in your business?

Let's start with a conversation. We will listen to where you are, what you are trying to achieve, and show you what is possible.

GET IN TOUCH

WEB

www.entra-ai.com

EMAIL

info@entra-ai.com

PHONE

+44 2034 323178