WARSAMÉ AHMED

SOLUTION ARCHITECTURE · DEFTECH

GROUNDED RAG OVER YOUR DOCUMENT ESTATE

Defence Intelligence Assistant

An AI assistant that lets planning staff interrogate manuals, procedures and doctrine — with every answer grounded in source and traceable to the page.

— PREPARED FOR THE CHIEF OF STAFF & TECHNICAL ARCHITECTURE TEAM

01 The problem

Knowledge locked in PDFs, and what "good" must look like.

02 The solution

A grounded assistant, shown answering a live question.

03 Technical architecture

Retrieval pipeline, ingestion, security, evaluation.

04 Deployment & delivery

On-prem vs cloud, impact, and a phased pilot.

FIRST A quick word on **scope & key assumptions** — the ground rules this proposal rests on. An **appendix** follows with model, evaluation and glossary reference.

SCOPE · KEY ASSUMPTIONS

CONFIRMED IN PHASE 0 DISCOVERY

What we're assuming — and where it's confirmed

The brief leaves room for interpretation. These are stated openly; none changes the architecture — only its configuration.

DATA & CORPUS

- ~50k mixed PDF/DOCX manuals, procedures and doctrine in the central database.
- Predominantly **English-language** — fits `embed-english-v3.0`; multilingual model swaps in if needed.
- Text-extractable; scanned-image documents would add an OCR step.

SECURITY & IDENTITY

- Documents carry — or can be tagged with — classification & need-to-know labels.
- An existing identity source provides user clearances and compartments to integrate.
- Classified workloads require a **private / air-gapped** deployment — which Cohere supports.

PERFORMANCE & EVALUATION

- Latency and accuracy figures shown are **illustrative**, to be baselined on DefTech infrastructure.
- DefTech provides subject-matter input to build the golden evaluation set.

SCOPE & DEMO

- All demo content is **notional doctrine** for illustration — not real classified material.
- Scope is retrieval & Q&A over existing documents — not authoring or editing source material.

SCOPE · ARCHITECTURAL ASSUMPTIONS

CONFIRMED IN PHASE 0 DISCOVERY

What we assume about their environment & needs

Drawn from the brief — these shape how the assistant integrates with the existing document estate and what it must deliver.

A1 DATABASE ARCHITECTURE

Blob storage + a metadata database

Their document management system pairs blob storage for file content with a database managing metadata, access, audit and versioning. We integrate with it — we don't replace it.

A2 SCOPED ACCESS

Not everyone sees everything

Document access is restricted at the **user or department level**; the assistant inherits and enforces those same restrictions at retrieval time.

A3 AUDIT & LOGGING

Every chunk presented is logged

An audit trail is required **each time** a chunk of content is surfaced to a user — who saw what, when, and from which source.

A4 TWO CAPABILITIES

Strict citation *and* a planning tool

Beyond grounded, cited answers, the system also helps planning staff by **identifying who needs training** on specific doctrine, procedures and manuals.

SECTION 01

The problem

A finished digital transformation that hasn't yet changed how staff actually work.

THE PROBLEM · CURRENT STATE

The documents are digital. The way staff use them is not.

What changed

Manuals, procedures & doctrine digitised
All content now lives as PDF & DOCX in one central database.

A single source of record
No more physical textbooks scattered across desks and shelves.

What didn't

Staff still read to retrieve
Finding one planning factor can mean opening several 300-page documents.

Search is keyword, not meaning
A query only works if you already know the document's wording.

> CHIEF OF STAFF
>
> **"Leverage AI over the new database to increase the efficiency and output of central planning staff — while accuracy and traceability remain key."**

THE PROBLEM · WHAT "GOOD" LOOKS LIKE

Four criteria the solution is measured against

CONSTRAINT

Accuracy

Answers grounded only in retrieved source — no fabrication, no guessing.

CONSTRAINT

Traceability

Every claim links to the exact document, section and page it came from.

OUTCOME

Security

Users only ever see content their clearance and need-to-know permit.

OUTCOME

Efficiency

Minutes of reading collapse into a sourced answer in seconds.

SECTION 02

The solution

A retrieval-augmented assistant — ask in plain language, get a grounded, cited answer.

THE SOLUTION · IN ONE SENTENCE

A Cohere-powered assistant that retrieves from your documents, then writes a grounded answer with citations.

It never speaks from the model's own memory. If the answer isn't in the retrieved source, it says so.

Ask in plain language

Semantic search finds the right passage even when wording differs.

Grounded answers

Generated strictly from retrieved text, with inline citations.

Conversational follow-up

Context carries across questions; suggested next questions guide planning.

Trust by design

Click any citation to open the source at the exact page.

DEMO · A PLANNER ASKS A QUESTION

NOTIONAL CONTENT · ILLUSTRATIVE

Assistant

SIGNED IN AS

Maj. R. Okonkwo

Central Planning · Clearance: SECRET

SOURCE FILTERS

- [x] Joint doctrine
- [x] Logistics procedures
- [ ] Engineering manuals

v0.9 · pilot

What are the key planning factors when establishing a forward logistics node in a contested environment?

Four factors govern node siting in a contested environment. **Survivability** is primary — dispersion, cover and air-defence coverage are weighed before throughput [1]. Lines of communication must offer at least two viable routes to the supported force [2]. Throughput capacity is matched to the demand forecast plus a contingency reserve [2], and command & control reachback must be assured before the node is declared operational [3].

SUGGESTED FOLLOW-UPS

- How is throughput contingency calculated?
- What triggers node relocation?

DEMO · EVERY ANSWER IS TRACEABLE

NOTIONAL CONTENT · ILLUSTRATIVE

The answer references three sources

Markers 1–3 map to retrieved passages, ranked by relevance.

1. **Joint Logistics Doctrine, Vol II**
   Ch. 4.2 — Siting & Survivability · p.118
2. **Forward Sustainment Procedures**
   §3 — Lines of Communication · p.41
3. **C2 Field Handbook**
   §7.1 — Reachback Assurance · p.205

● **Groundedness check passed · 0 unsupported claims**

SOURCE PREVIEW · CITATION 1

Joint Logistics Doctrine, Vol II

Chapter 4.2 · Page 118 · Classified: SECRET

> "...the commander shall weigh survivability — achieved through dispersion, cover, and assured air-defence coverage — ahead of raw throughput when selecting a forward node location in a contested environment."

Open document | Copy reference

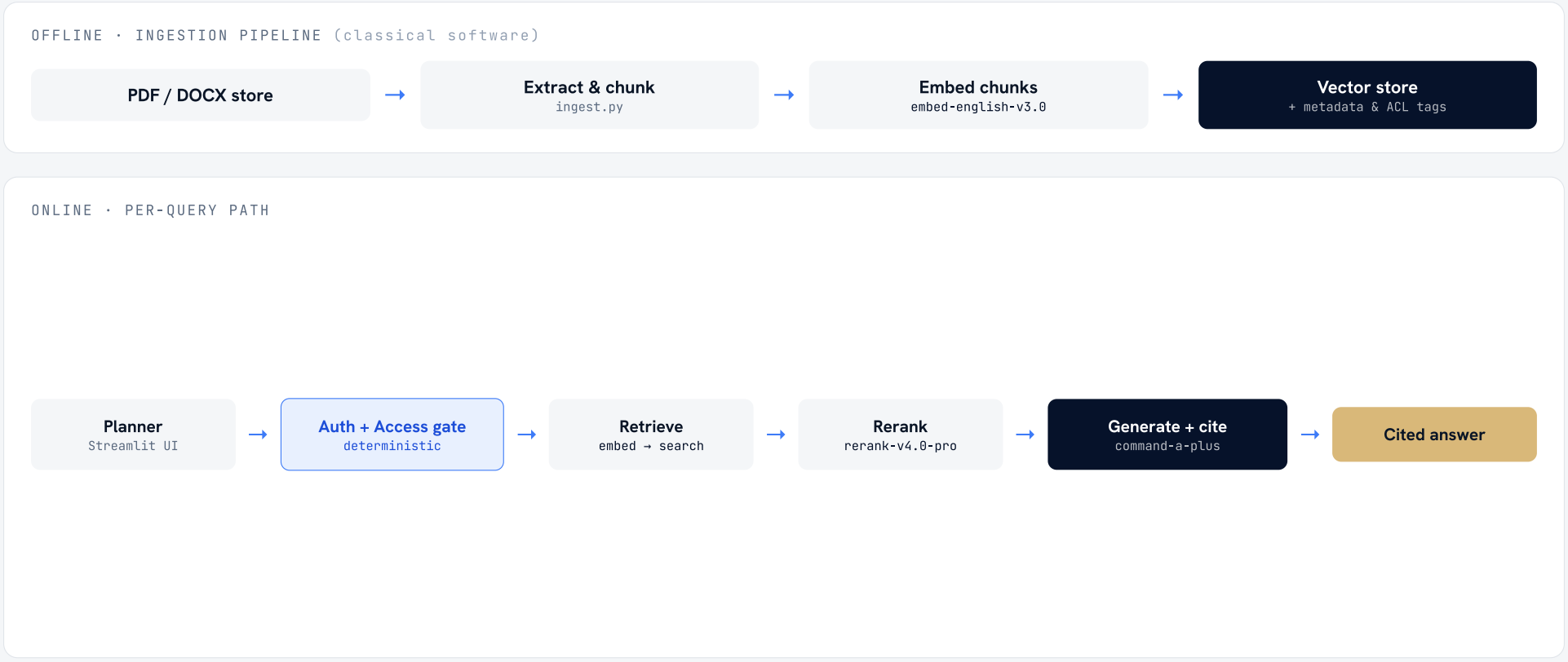
SECTION 03

Technical architecture

A RAG pipeline at the core, wrapped in deterministic security and governance.

ARCHITECTURE · END TO END

Ingestion runs offline; retrieval and generation run per query



ARCHITECTURE · THE RETRIEVAL PIPELINE

Cast a wide net, then narrow hard

Two-stage retrieve-then-rerank is what turns "plausible" into "precise" — and it's why the citations are trustworthy.

STEP 1

Embed query

`embed-english-v3.0`

Query → 1024-dim vector in the same space as the chunks.

→

STEP 2

Vector search

`ANN over allowed set`

Retrieve **top 50** candidates. High recall, low cost.

→

STEP 3

Rerank

`rerank-v4.0-pro`

Cross-encoder rescores 50 → keep **top 8**. Precision.

→

STEP 4

Generate + cite

`command-a-plus`

Answers strictly over the 8 passages, emitting inline citations.

MEMORY Follow-up questions are rewritten against conversation history before Step 1, so "how is that calculated?" resolves to a fully-specified query.

ARCHITECTURE · INGESTION

From raw documents to governed, searchable chunks

Mechanical parsing — no model judgement. Metadata captured here is what the security layer enforces at query time.

01

Extract text

PDF & DOCX parsed to clean text, tables and headings preserved.

ingest.py

02

Chunk

Split into overlapping passages sized for retrieval, kept section-aware.

~512 tokens · overlap

03

Tag & attach metadata

Source, section, page — and classification + need-to-know.

ACL tags

04

Embed & index

Each chunk embedded and written to the vector store with its tags.

embed-english-v3.0

ASSUMPTION Modelled on ~50k mixed documents across multiple classification tiers; re-ingestion is incremental on change.

ARCHITECTURE · SECURITY & ACCESS CONTROL

Access is filtered before retrieval — and it isn't AI

How the gate works, per query

User identity
clearances + compartments

∩

Chunk ACL tags
classification + need-to-know

↓

Allowed chunk set — the *only* candidates retrieval and the model ever see

`auth.py · session tokens` `SQL join + set intersection` `full audit log`

Deterministic

Same inputs, same result, every time — no probabilistic judgement on access.

Auditable

Every access decision is logged and explainable to an accreditation officer.

"You wouldn't want an LLM deciding who sees classified documents — that decision must be deterministic and auditable."

ARCHITECTURE · A DELIBERATE DIVISION

What's genuinely AI — and what's deliberately not

AI where it adds judgement and recall; deterministic engineering where correctness and security must be guaranteed.

LLM / NLP

Genuinely AI-driven

- NLP **Embedding generation** — text → vectors for semantic search
  `embed-english-v3.0`
- NLP **Reranking** — transformer scores passage relevance
  `rerank-v4.0-pro`
- LLM **Grounded answer generation** with citations
  `command-a-plus`
- LLM **Conversation memory** across follow-up questions
- LLM **Follow-up question generation**

The substance of "leverage AI over the database" — the full RAG pipeline.

CLASSICAL SOFTWARE

Deliberately deterministic

Authentication — credential check + session tokens `auth.py`

Access-control filtering — SQL join + set intersection

Ingestion — PDF/DOCX text extraction `ingest.py`

Presentation — the Streamlit UI

Security controls that sit *around* the AI — correct, exact, and auditable.

ARCHITECTURE · ACCURACY & TRACEABILITY

Accuracy is engineered, then measured

Grounded prompting

The model is instructed to answer only from retrieved passages, and to say when the source doesn't cover it.

Mandatory citations

Every claim is bound to a passage; uncited statements are flagged before display.

Evaluation harness

A golden Q&A set scored on every release, with human review on a sample.

Groundedness

≥98%
claims supported by source

100%
answers carry citations

recall@8
tracked per release

TARGETS ARE ILLUSTRATIVE PILOT GOALS — BASELINED AGAINST DEFTECH'S OWN GOLDEN SET.

ARCHITECTURE · LATENCY BUDGET

Sourced answers in seconds, not minutes

~0.8s
to first streamed token

~2–3s
to full cited answer

ILLUSTRATIVE · VARIES WITH DEPLOYMENT & CONCURRENCY

SECTION 04

Deployment & delivery

Where it runs, what it gives back, and how we get there.

DELIVERY · DEPLOYMENT OPTIONS

The same pipeline runs where accreditation requires

Cohere deploys privately — start in cloud for the pilot, migrate on-prem without re-architecting.

Private / air-gapped

CLASSIFIED FIT

- ✓ Models run inside your security boundary — no data egress
- ✓ Full control of data residency and the update cycle
- ✓ Aligns with accreditation for classified material
- ✓ Runs on your hardware or sovereign environment

Higher setup effort · maximum control

Dedicated cloud / VPC

PILOT FIT

- ✓ Fastest path to a working pilot
- ✓ Single-tenant, network-isolated deployment
- ✓ Elastic scaling, managed updates
- ✓ Ideal for an unclassified corpus to prove value

Lower setup effort · quickest to value

HOW IT ADDRESSES THE PROBLEM

Every success criterion maps to a named mechanism

Accuracy	Grounded generation over reranked passages — the model answers only from source, with no fabrication.
Traceability	Mandatory inline citations open the exact document, section and page behind every claim.
Security	Deterministic access gate filters by clearance and need-to-know *before* retrieval, with a full audit log.
Efficiency	A sourced answer in seconds replaces opening and reading multiple long documents.

HOW IT ADDRESSES THE PROBLEM · EFFICIENCY

Reading time becomes planning time

~~~30 min~~~ to locate one sourced factor by hand

<1 min to the same answer, with citations

Hours / week

recovered per planner, redirected to planning judgement.

Consistency

the same sourced answer regardless of who asks.

ILLUSTRATIVE PILOT HYPOTHESES — MEASURED AGAINST DEFTECH'S OWN BASELINE.
~~~

DELIVERY · A PHASED PILOT

Prove value first, accredit second, scale third

PHASE 0

Discovery

Baseline current effort. Assemble the golden Q&A set. Confirm corpus & ACL model.

~2 WEEKS

PHASE 1

Pilot

One directorate, unclassified corpus, cloud. Measure against the baseline.

~6 WEEKS

PHASE 2

Accredit & secure

Move to private deployment. Onboard classified material under full ACL.

ACCREDITATION-LED

PHASE 3

Scale

Roll out across directorates. Expand corpus, monitor evaluation continuously.

ONGOING

EACH PHASE HAS A CLEAR EXIT GATE · DefTech decides to proceed at every step.

WARSAMÉ AHMED

THE OBJECTIVE

Grounded, traceable AI over your documents — safe to accredit.

Accuracy and traceability built into the architecture; security kept deterministic by design. Demonstrating how Cohere's models solve a real DefTech problem.

Next step → a 2-week discovery & pilot scope

FOR THE RECORD

Appendix

Model reference, evaluation methodology, the risk register, and a glossary of terms.

A1 · Model reference

A2 · Evaluation methodology

A3 · Risks & mitigations

A4 · Glossary

APPENDIX A1 · COHERE MODEL REFERENCE

The three models, and what each one does

embed-english-v3.0 NLP · EMBEDDINGS	Turns each chunk and each query into a 1024-dimension vector so semantic search can match on meaning, not keywords.	1024-dim ingestion + query
rerank-v4.0-pro NLP · RELEVANCE	A cross-encoder that rescores the top-50 candidates against the full query and keeps the best 8 — the precision step.	50 → 8 per query
command-a-plus LLM · GENERATION	Reads the 8 retrieved passages and writes the grounded answer with inline citations; also drives conversation memory and follow-ups.	grounded + citations

EXACT CONTEXT LIMITS & DIMENSIONS CONFIRMED AGAINST THE DEPLOYED MODEL VERSION AT INTEGRATION.

APPENDIX A2 · EVALUATION METHODOLOGY

How quality is measured every release

The golden set

1. SMEs author representative questions with known-correct answers and the source passages that support them.
2. The set spans easy lookups, multi-document synthesis, and deliberate "not in the corpus" cases.
3. It grows as real usage surfaces new question patterns.

Groundedness
Share of claims actually supported by the cited passage.

Citation accuracy
Whether each citation points to the right source location.

Retrieval recall@8
Whether the right passage reached the final 8 after rerank.

RELEASE GATE Automated scores run on every release; a human reviews a sample. A regression below threshold blocks the release.

APPENDIX A3 · RISK REGISTER

Risks we track — and the control for each

Hallucination / fabrication

Model asserts something not in the source.

CONTROL Grounded prompting + uncited-claim flagging before display.

Data leakage across clearances

A user sees content above their clearance.

CONTROL Deterministic pre-retrieval gate; non-cleared chunks never candidates.

Stale or superseded answers

Doctrine changes; the index lags.

CONTROL Incremental re-ingestion on change; version & date shown on every source.

Poor retrieval (over/under)

Wrong or missing passages reach the model.

CONTROL Reranker cut-off tuned on the golden set; recall@k tracked per release.

APPENDIX A4 · GLOSSARY

Shared vocabulary

RAG	Retrieval-Augmented Generation — the model answers from retrieved source, not memory.
ANN search	Approximate Nearest Neighbour — fast retrieval of the closest vectors at scale.
Groundedness	The property that every claim in an answer is supported by retrieved source.
ACL	Access Control List — the tags deciding which users may see a given chunk.
Embedding	A numeric vector representing text meaning, enabling semantic comparison.
Cross-encoder	A reranking model that scores a query and passage together for precise relevance.
Chunk	A passage-sized slice of a document, the unit that is embedded and retrieved.
Need-to-know	Compartment check layered on clearance level; both must pass for access.