BODDAN

RAG Knowledge Assistant

SOLUTION DESIGN & HANDOVER (SAMPLE)

SAMPLE · demonstration of deliverable quality · not a real client's work

ABN 97 728 052 912

Sola Fide

Prepared for

Client: Hunter & Vale Legal (fictional sample client)

Location: Newcastle, New South Wales

Brief as received: Hunter & Vale Legal asked for a private retrieval assistant over roughly 1,200 internal precedent documents, templates and policies, so that fee-earners can ask plain-English questions and get answers cited back to the source file. The assistant is to run on the firm's own Anthropic API key and hosting, with a tuned ingestion pipeline, a document-update workflow the firm can run itself, one integration into the firm intranet, a tested deployment, a full handover document and walkthrough, and two weeks of post-delivery bug-fix support.

Date: 30/06/2026

Service and tier: RAG Knowledge Assistant - Solution Design (Premium)

> This is a disclosed, fictional sample produced by BODDAN to demonstrate deliverable quality. Hunter & Vale Legal is not a real firm, and the figures, names and configuration details are illustrative. The technical approach, the trust-boundary analysis and the acceptance criteria are real and are written to the standard BODDAN applies on live engagements.

1. Executive summary

Hunter & Vale Legal holds a large, valuable body of internal know-how: precedent deeds and agreements, matter templates, file notes of approach, and firm policies. Today that knowledge is found by remembering which matter it lived on, asking a colleague, or opening files one at a time. The firm wants a private assistant that lets a fee-earner ask a question in plain English (for example, "what is our standard limitation-of-liability clause for a commercial lease, and where is the current template") and receive a direct answer with a citation pointing back to the exact source file.

This document is the solution design and handover specification for that assistant. It describes a retrieval-augmented generation (RAG) system built on three principles that matter for a law firm:

- **The corpus stays on the firm's hosting.** Documents, the search index and the access-control model never leave the firm's own infrastructure. Only the user's question and the minimal text needed to answer it are sent to the Anthropic API, under the firm's own key, at the moment a question is asked. Section 6 sets out this boundary precisely.
- **Answers are anchored to sources.** Every answer carries citations that resolve to a specific source file in the firm's document management system (DMS). The assistant is built to retrieve and cite, not to opine.
- **The firm stays in control.** A self-serve update workflow lets the firm add, supersede and remove documents without BODDAN in the loop, so a withdrawn precedent stops appearing in answers the same day it is withdrawn.

The system uses self-hosted embeddings and a self-hosted vector index on the firm's Azure tenant in the Australia East (Sydney) region, hybrid keyword-and-vector retrieval with reranking, retrieval filtered by the firm's matter-access rules so information barriers are enforced before any text reaches the model, and Anthropic's Claude (model `claude-opus-4-8`) for answer generation with native citations. It integrates into the firm's SharePoint intranet as an embedded panel using existing Microsoft Entra ID single sign-on.

The engagement is delivered as a fixed-fee build with a defined test plan, acceptance criteria, a handover document, a live walkthrough, and two weeks of post-delivery bug-fix support. Commercials and timeline are in Section 14.

A plain statement of limits sits in Section 13 and is repeated here: this assistant helps people find and read the firm's own material faster. It does not give legal advice, it does not replace a lawyer's review, and its output must be checked against the cited source before it is relied on.

2. Objectives and scope

2.1 OBJECTIVES

#	Objective	How it is met
O1	Fee-earners can ask plain-English questions across the internal corpus	Natural-language query interface in the intranet, backed by hybrid retrieval and Claude
O2	Every answer cites the source file	Native citations resolved to DMS file references via chunk metadata (Section 5.4)
O3	Confidential and privileged material is protected	Self-hosted corpus and index; query-time egress only; access-filtered retrieval (Section 6)
O4	The firm can maintain the corpus itself	Self-serve add / supersede / remove workflow (Section 7)
O5	The assistant lives where people already work	Embedded panel in the SharePoint intranet with Entra ID SSO (Section 8)
O6	The build is verifiable	Documented test plan with pass / fail acceptance criteria (Section 9)
O7	The firm can run and own the system after handover	Handover document, walkthrough, runbook and two-week support (Sections 11 and 12)

2.2 IN SCOPE

- Ingestion of approximately 1,200 documents from a single nominated DMS source location (or an export of it), in the formats listed in Section 4.
- One ingestion pipeline covering extraction, chunking, metadata capture, embedding and indexing.
- Hybrid retrieval, reranking, access-filtered retrieval, and answer generation with citations.

- One intranet integration: an embedded assistant panel in the firm's existing SharePoint intranet.
- A self-serve document-update workflow with a simple operator interface and a runbook.
- Test execution against a firm-provided question set, with acceptance sign-off.
- Deployment into the firm's Azure tenant, configured to use the firm's Anthropic API key.
- Handover documentation, a recorded and live walkthrough, and 10 business days of post-delivery bug-fix support.

2.3 OUT OF SCOPE (UNLESS SEPARATELY AGREED)

- Connecting live matter files or email; the assistant indexes the nominated know-how corpus, not the whole DMS.
- Drafting, automated document assembly, or any feature that produces client-facing legal work product.
- Practice-management, billing or time-recording integration.
- Migration or restructuring of the firm's DMS.
- Additional intranet or third-party integrations beyond the one in scope.
- Ongoing managed service beyond the two-week support window (available separately as a retainer).
- Provision of the Anthropic API spend itself; the firm pays Anthropic directly under its own account (see Section 14 and Appendix B).

2.4 SUCCESS CRITERIA

The build is accepted when every acceptance criterion in Section 9 passes against the firm-provided question set, the self-serve workflow has been demonstrated end to end by a firm operator, the intranet panel works under live SSO for the pilot group, and the handover walkthrough is complete. Retrieval and citation quality is measured against agreed thresholds on that question set, not asserted as a guaranteed outcome.

---

3. Solution architecture overview

3.1 ARCHITECTURE AT A GLANCE

The system has two planes. The **knowledge plane** (everything that holds or indexes firm documents) sits entirely on the firm's hosting. The **answer plane** makes a single outbound call per question to the Anthropic API to turn retrieved text into a cited answer.

```
Fee-earner
|
v
SharePoint intranet panel ----(Entra ID SSO)----.
| |
v |
```

RAG service (firm Azure tenant, Australia East) |

| 1. authenticate + resolve user's matter access

| 2. hybrid retrieve (keyword + vector)

| 3. access-filter chunks

| 4. rerank

5. build prompt (question + retrieved chunks)
------ outbound, query time only ---------> Anthropic API
(question + retrieved snippets, (claude-opus-4-8,
firm's own API key) citations enabled)
<----- cited answer ----------'

v

Answer + citations resolved to DMS source files

```

Supporting the knowledge plane, all on firm hosting:

- **Document source:** the nominated DMS location (for example, an iManage or equivalent know-how workspace), or a controlled export of it.
- **Ingestion service:** a Python service that extracts, chunks, captures metadata, embeds and indexes documents.
- **Embedding model:** a self-hosted open-weights text-embedding model running on the firm's infrastructure, so document text is never sent to a third party to be embedded.
- **Vector index:** PostgreSQL with the `pgvector` extension, holding chunk vectors, text and metadata.
- **Keyword index:** a lexical (BM25-style) index over the same chunks for hybrid retrieval.
- **Reranker:** a self-hosted cross-encoder reranking model.

## 3.2 COMPONENTS

| Component | Technology (indicative) | Where it runs | Why |
|---|---|---|---|
| Intranet panel | SharePoint Framework (SPFx) web part / embedded panel | Firm SharePoint Online | Meets people where they work; uses existing SSO |
| RAG service API | Python (FastAPI), containerised | Firm Azure tenant (Australia East) | Orchestrates auth, retrieval and generation |
| Embedding model | Self-hosted open-weights embedding model (1024-dimension), e.g. `bge-large-en-v1.5` | Firm Azure tenant | Keeps document text on firm hosting at embed time |
| Vector + metadata store | PostgreSQL 16 + `pgvector` | Firm Azure tenant | On-prem-style control; ample for the corpus size |
```

Component	Technology (indicative)	Where it runs	Why
Keyword index	BM25 (PostgreSQL full-text or OpenSearch)	Firm Azure tenant	Catches exact terms, clause names, defined terms
Reranker	Self-hosted cross-encoder, e.g. `bge-reranker-v2`	Firm Azure tenant	Sharpens the top results before generation
Generation	Anthropic Claude `claude-opus-4-8` (Citations enabled)	Anthropic API, firm's key	Turns retrieved text into a cited, readable answer
Identity	Microsoft Entra ID	Firm tenant	SSO and group-based access

Model choice is set out in Section 5.4; `claude-opus-4-8` is the default for answer quality, with `claude-sonnet-4-6` available as a lower-cost option for routine lookups (Appendix B).

3.3 HOSTING AND DATA RESIDENCY

The knowledge plane and the RAG service are deployed in the firm's own Microsoft Azure tenant in the Australia East (Sydney) region, so the corpus, the indexes and the application logs reside in Australia on infrastructure the firm controls. The single exception is the query-time call to the Anthropic API, which is the subject of Section 6.5 (including an in-country inference option if the firm requires inference itself to remain in Australia).

---

4. Ingestion pipeline

The ingestion pipeline turns firm documents into retrievable, citeable chunks. It is built to be run repeatedly and incrementally, which is what makes the self-serve update workflow in Section 7 possible.

4.1 SOURCE CONNECTION AND DOCUMENT INTAKE

The pipeline reads from one nominated DMS location. Each document is pulled with its DMS identifier, title, document type, matter or practice-area tags where present, author, and last-modified date. Where direct DMS API access is not available for the first build, a controlled one-time export plus a watched folder for ongoing changes is used instead. The pipeline records, for each document, the DMS reference it came from, because that reference is what every later citation resolves back to.

Supported formats: PDF (including scanned PDF via OCR), Microsoft Word (.docx and .doc), plain text and Markdown, and Microsoft PowerPoint (.pptx) for policy decks. Spreadsheets are out of scope for the first build unless specifically nominated, because tabular content needs a different chunking treatment.

4.2 TEXT EXTRACTION AND NORMALISATION

Each document is converted to clean text with its structure preserved as far as the format allows: headings, clause numbering, lists and tables are kept as markers so that a retrieved chunk reads sensibly and a citation lands on the right clause. Scanned PDFs are passed through OCR;

documents that OCR poorly are flagged in an ingestion report for human attention rather than silently indexed at low quality. Boilerplate that adds noise (repeated headers, footers and page furniture) is stripped.

4.3 CHUNKING STRATEGY

Documents are split into overlapping chunks sized for retrieval precision. The design uses a target chunk size of **600 tokens** with **100 tokens of overlap** between consecutive chunks, and splits are made at structural boundaries (clause, heading or paragraph) wherever possible rather than mid-sentence. Overlap means a clause that straddles a boundary is still wholly present in at least one chunk. Each chunk keeps a pointer to its position in the parent document so a citation can name the section, not just the file.

Chunk size is a tunable. During the build it is validated against the firm question set (Section 9); if answers are landing on too-narrow or too-broad spans, the size and overlap are adjusted and the corpus re-indexed before acceptance.

4.4 METADATA SCHEMA

Every chunk carries metadata that drives both filtering and citation. The full field reference is in Appendix A. The fields that matter most are:

- **Source reference:** DMS document identifier, file name and document version.
- **Section locator:** heading or clause path within the document.
- **Document type:** precedent, template, policy, file note, or other.
- **Practice area and matter tags:** used for access filtering (Section 5.3) and for narrowing searches.
- **Access labels:** the groups or matter-security markers that govern who may see the chunk.
- **Lifecycle status:** current, superseded, or withdrawn (Section 7).
- **Ingested-at and source-modified-at timestamps.**

4.5 EMBEDDINGS

Each chunk is converted to a vector by a self-hosted open-weights embedding model running on the firm's infrastructure. This is a deliberate confidentiality choice. Anthropic does not offer a first-party embeddings endpoint, and the common managed alternative (a third-party embeddings provider such as Voyage AI) would mean sending document text to that provider to be embedded. By self-hosting the embedding model, the document text used to build the index never leaves the firm's hosting. Using a managed embeddings provider instead is possible and would reduce the firm's compute footprint, but it is presented as an explicit tradeoff, not the default, because it widens the set of parties that see document text.

4.6 INDEX BUILD AND SIZING

Vectors, chunk text and metadata are written to PostgreSQL with `pgvector`, and the same chunks are added to the keyword index. The build produces an ingestion report: documents processed, chunks created, OCR-flagged documents, and any failures.

Sizing for this corpus, derived from a single set of assumptions (see Appendix B for the same figures used in the cost model):

Quantity	Value	Basis
Documents	1,200	Brief
Average tokens per document	7,500	Estimate across precedents, templates and policies
Effective tokens per chunk (size minus overlap)	500	600-token chunk, 100-token overlap
Chunks per document	15	7,500 / 500
Total chunks indexed	**18,000**	1,200 x 15

The chunk total is an estimate; the true figure depends on real document lengths and is reported precisely after the first full ingestion. A corpus of this scale sits comfortably within a single PostgreSQL instance and needs no specialist vector-database infrastructure.

5. Retrieval and citation

5.1 HYBRID RETRIEVAL

When a question arrives, the system runs two searches in parallel:

- **Vector search** over the chunk embeddings, which finds passages that are semantically close to the question even when the wording differs.
- **Keyword (BM25) search** over the same chunks, which catches exact terms that vector search can miss: a specific clause name, a defined term, a section number, or a statute reference.

The two result sets are merged. Hybrid retrieval is used because legal know-how is full of precise terms where an exact-match miss is costly, and also full of concepts a user will phrase loosely. Neither search alone is sufficient.

5.2 RERANKING

The merged candidate set (typically the top 20 to 40 chunks) is passed through a self-hosted cross-encoder reranker, which scores each candidate against the question directly and reorders them. The top results after reranking (typically six to eight chunks) are what proceed to generation. Reranking materially improves which passages reach the model, which is what most affects answer and citation quality.

5.3 ACCESS-FILTERED RETRIEVAL (ETHICAL WALLS)

This is a hard requirement for a law firm and is enforced in the retrieval layer, not only in the user interface. Before any chunk is sent to the model, the candidate set is filtered against the requesting user's access rights, resolved from Entra ID group membership and the matter-security and information-barrier labels captured in chunk metadata. A chunk the user is not entitled to see is removed from the candidate set, so it never reaches the reranker, the prompt, or the answer.

The consequence is that the model can only ever cite material the user was already entitled to open. Zero cross-barrier leakage is treated as a pass / fail acceptance gate in Section 9, not as a percentage target. The firm can also exclude entire matter classes from indexing in the first place (Section 7), so the most sensitive or ethical-walled matters need never enter the index at all.

5.4 GENERATION AND CITATION

The surviving top chunks are assembled into a prompt and sent to Claude (`claude-opus-4-8`) under the firm's API key, with Anthropic's native Citations feature enabled. Each chunk is supplied to the model as a document block, so the model's answer cites the specific spans of the specific chunks it relied on. The system prompt instructs the model to answer only from the supplied material, to say so plainly when the material does not contain the answer, and never to present itself as giving legal advice.

Citation works in two layers, which together make "cited back to the source file" technically true:

1. **Span to chunk:** Anthropic's Citations feature returns which spans of which supplied chunks the answer drew on.
2. **Chunk to source file:** the application resolves each cited chunk to its source reference and section locator (Appendix A), and renders a citation the user can click to open the exact file in the DMS at the right section.

The system prompt is stable across queries and benefits from prompt caching; the retrieved chunks differ on every query and are not cacheable, so no caching saving is claimed on the retrieval payload. Routing routine, low-ambiguity lookups to `claude-sonnet-4-6` is offered as a cost option (Appendix B); answer quality on nuanced questions is the reason `claude-opus-4-8` is the default.

5.5 WORKED EXAMPLE

A fee-earner asks: *"What is our standard limitation-of-liability clause for a commercial lease, and which template is current?"*

1. Hybrid retrieval finds chunks from the commercial-lease precedent, the firm's clause bank, and a superseded earlier template.
2. Access filtering confirms the user is entitled to the commercial-property know-how; all candidate chunks survive.
3. The reranker pushes the current clause-bank chunk and the current template chunk to the top; the superseded template chunk is down-ranked and, because it carries a "superseded" lifecycle status, is excluded by the retrieval filter (Section 7).
4. Claude answers with the clause text and names the current template, citing the clause-bank file and the current template file.
5. The user clicks the citation and opens the current template in the DMS at the limitation-of-liability clause. The superseded template never appears.

---

6. Security, confidentiality and professional obligations

This section is the spine of the design. It is written to be precise rather than reassuring, because for a law firm the precise boundary is what matters.

6.1 THE THREE DATA ZONES (TRUST BOUNDARY)

It would be inaccurate to say "nothing ever leaves the firm." It is accurate, and important, to say exactly what leaves, when, and under what terms. The system has three zones:

Zone 1 - never leaves firm hosting. The source corpus and DMS, every extracted chunk, every embedding vector, the vector and keyword indexes, the access-control model, the application code, and the application logs. All of this resides in the firm's Azure tenant in Australia. Self-hosted embeddings and a self-hosted index are chosen precisely so that building and searching the index requires nothing to leave.

Zone 2 - the egress, at query time only. When, and only when, a user asks a question, the system sends to the Anthropic API: the user's question, and the minimal retrieved snippets needed to answer it (the top chunks that survived access filtering and reranking). It does not send the whole corpus, the index, or any chunk the user was not entitled to see. The call is made under the firm's own Anthropic API key.

Zone 3 - Anthropic's processing. Anthropic processes that single request and returns a cited answer. This processing is governed by Anthropic's commercial terms for API use (Section 6.2).

Claiming this boundary, rather than an absolute, is what makes the assistant safe to reason about: the firm can state with confidence what is and is not exposed at each step.

6.2 NO TRAINING AND DATA RETENTION

Under Anthropic's published Commercial Terms of Service and Data Processing Addendum, inputs and outputs submitted through the commercial API are not used to train Anthropic's models. Anthropic's published policy is that API inputs and outputs are retained only for a limited period for trust-and-safety purposes, and zero-data-retention arrangements are available to eligible organisations. These positions should be confirmed in the agreement the firm executes with Anthropic, and the specific retention period and zero-retention eligibility verified at that time rather than assumed. BODDAN will assist the firm to record the relevant terms as part of handover, but the contracting relationship for the API is between the firm and Anthropic.

6.3 ACCESS CONTROL AND ETHICAL WALLS

Access is enforced at three points:

- **Authentication:** the intranet panel authenticates the user through existing Entra ID SSO; only signed-in firm users reach the service.
- **Authorisation at retrieval:** as described in Section 5.3, candidate chunks are filtered against the user's entitlements before reaching the model, so information barriers hold at the data layer.
- **No privilege escalation through the assistant:** because the assistant can only retrieve what the user could already open, it cannot become a side channel around matter security.

6.4 PROFESSIONAL AND REGULATORY OBLIGATIONS

The design is shaped by the obligations a NSW firm carries:

- **Confidentiality and privilege** under the Legal Profession Uniform Law (NSW) and the firm's duties to clients. The trust boundary in Section 6.1 and the access filtering in Section 6.3 are the technical expression of those duties. Client legal privilege is preserved by keeping privileged material within the firm's entitlement model and by limiting query-time egress to what is needed and permitted.
- **Privacy** under the Privacy Act 1988 (Cth) and the Australian Privacy Principles, relevant where documents contain personal information. Australian data residency for the knowledge plane and the query-time-only egress model support the firm's privacy posture.

These are the firm's obligations to discharge; BODDAN builds the system to support them and documents how, but does not provide legal or compliance advice on them.

6.5 IN-COUNTRY INFERENCE OPTION

The first-party Anthropic API does not process requests in Australia. If the firm requires that inference itself remain in-country, the honest path is to run Claude through Amazon Bedrock in the Sydney region (ap-southeast-2), which keeps the query-time processing in Australia. This trades the "firm's own Anthropic key" arrangement for AWS-managed access to the model, and is subject to the required Claude model being available in ap-southeast-2 at the time of deployment. It is presented as an explicit decision for the firm to make, weighing in-country inference against the simplicity of the firm's direct Anthropic relationship. The default design in this document uses the firm's own Anthropic API key as the brief specified.

6.6 LOGGING, AUDIT AND RETENTION

The RAG service logs questions, the documents retrieved, the citations returned, and who asked, in the firm's tenant, to support audit and quality review. Log retention is configurable by the firm. Because logs can themselves contain confidential text, they are held in Zone 1 and subject to the firm's access controls and retention policy.

7. Self-serve document-update workflow

The corpus is a living thing. Precedents are revised, templates are replaced, policies are withdrawn. The firm must be able to keep the index current without BODDAN in the loop, and a withdrawn precedent must stop appearing in answers promptly. The workflow supports three operations:

Operation	What the operator does	What the system does
Add	Place a new document in the watched location (or tag it in the DMS)	Extracts, chunks, embeds and indexes it; reports chunks created

Operation	What the operator does	What the system does
Supersede	Mark a document as superseded and point to its replacement	Sets the old document's lifecycle status to "superseded", which excludes it from retrieval, and indexes the replacement
Remove / withdraw	Mark a document withdrawn (or delete it from the source)	Sets status to "withdrawn" and removes its chunks from both indexes so it can no longer be retrieved or cited

Two points are deliberate, because in a legal corpus a stale precedent surfacing is a real harm, not a cosmetic flaw:

- **Supersession is a first-class action.** A superseded template drops out of retrieval immediately on the next index sync, while its replacement takes its place. The worked example in Section 5.5 shows this in effect.
- **Withdrawal genuinely deletes from the index.** Withdrawing a document removes its chunks from the vector and keyword indexes, so it cannot be retrieved, reranked or cited. It is not merely hidden in the interface.

Updates run incrementally, so a daily (or on-demand) sync processes only what changed. The operation is delivered with a simple operator interface, a written runbook, and a verification step the operator can run to confirm a withdrawn document no longer returns in a search. Operating the workflow is a firm task; it requires no developer involvement.

8. Intranet integration

The assistant is delivered as a single integration: an embedded panel in the firm's existing SharePoint intranet. The integration:

- Renders as a panel or page within the intranet, so fee-earners use the assistant where they already work, with no new application to learn or log in to separately.
- Uses the firm's existing **Microsoft Entra ID single sign-on**, so the signed-in user is the identity used for access-filtered retrieval. There is no separate password.
- Presents the answer with clickable citations that open the cited source file in the DMS at the relevant section.
- Shows, alongside each answer, the documents it drew on, so the user can see the basis of the answer at a glance.
- Includes a persistent, visible reminder that the assistant helps locate and read firm material and does not give legal advice (Section 13).

The integration is scoped to one intranet surface. Additional surfaces (for example, a Microsoft Teams app or a desktop DMS plug-in) are out of scope for this build and available as separate work.

9. Test plan and acceptance criteria

Testing is run against a **firm-provided question set** of representative questions with known good answers and known correct source files, assembled with the firm during discovery. Quality is measured against that set; it is not asserted in the abstract. The target thresholds below are acceptance gates agreed before testing, tuned during the build, and signed off at acceptance.

9.1 TEST TYPES

Test	What it checks
Ingestion integrity	Every in-scope document is indexed or explicitly flagged (for example, OCR failures); chunk counts reported
Retrieval relevance	For each question, the correct source document appears in the retrieved set
Citation accuracy	The answer's citation resolves to the correct source file and section
Access control (ethical walls)	A restricted user cannot retrieve or cite material outside their entitlements
Supersession and withdrawal	A superseded document stops being retrieved; a withdrawn document is gone from the index
Faithfulness	The answer is supported by the cited material and does not assert beyond it
"Not found" behaviour	When the corpus lacks the answer, the assistant says so rather than inventing one
Performance	Median end-to-end answer time within the agreed target under pilot load
Integration	The panel works under live SSO for the pilot group

9.2 ACCEPTANCE CRITERIA

#	Criterion	Threshold
A1	Ingestion coverage	100% of in-scope documents indexed or explicitly flagged with a reason
A2	Retrieval relevance	Correct source document present in the retrieved set for at least 90% of the question set
A3	Citation accuracy	Citation resolves to the correct source file for at least 90% of answered questions
A4	Access control	**Zero** cross-barrier retrievals or citations across all access-control test cases (hard pass / fail)
A5	Supersession / withdrawal	**Zero** retrievals or citations of superseded or withdrawn documents after sync (hard pass / fail)
A6	Faithfulness	At least 95% of answers supported by their cited material on review of the question set
A7	"Not found" behaviour	The assistant declines to answer (rather than fabricating) on 100% of the designed "no answer in corpus" cases

#	Criterion	Threshold
A8	Performance	Median end-to-end answer under 8 seconds at pilot load (illustrative target; the exact figure is agreed at the start of testing)
A9	Integration	Panel functions under live Entra ID SSO for all pilot users

A4, A5 and A7 are hard pass / fail, but for a different reason than A6, and the distinction is worth being explicit about. A4 (cross-barrier retrieval) and A5 (superseded or withdrawn documents) test deterministic, rule-enforced behaviour: the access filter either removes an out-of-entitlement chunk or it does not, and a withdrawn document is either gone from both indexes or it is not. These outcomes are genuinely zero-able, so a single failure is unacceptable regardless of the rate. A7 is also pass / fail, but it tests one specific binary decision: on cases that are designed to have no answer in the corpus, does the assistant decline rather than attempt one. That is a "does it decline when it should" test, not a graded measure of answer quality.

A2, A3 and A6 are quality thresholds measured on the firm question set. A6 in particular grades degree of support: "is what the answer asserts supported by the material it cites", scored across answered questions on a graded scale rather than as a single binary. A6 and A7 are therefore different tests, not the same behaviour held to two different bars: A7 asks whether the assistant declines when it should, A6 asks whether what it does assert is supported by its cited material. The 90% to 95% figures are the agreed acceptance bar for this build on that set, not a guarantee of performance on every future question.

10. Deployment

Stage	Activity
Provisioning	Stand up the RAG service, PostgreSQL + `pgvector`, embedding and reranker services in the firm's Azure tenant (Australia East)
Configuration	Configure the firm's Anthropic API key as a secret in the firm's key vault; configure Entra ID SSO; configure the DMS connection
First ingestion	Run the full ingestion; produce and review the ingestion report; resolve OCR flags
Test execution	Run the Section 9 test plan against the firm question set; tune chunking and retrieval; re-index if needed
Pilot	Release the intranet panel to a nominated pilot group of fee-earners
Acceptance	Confirm all acceptance criteria; obtain written sign-off
Go-live	Release to the wider firm

Secrets, including the Anthropic API key, are stored in the firm's own key vault and are never embedded in code or configuration files. The firm can rotate the key at any time.

11. Handover and knowledge transfer

Ownership transfers to the firm at handover. The firm receives:

- **A handover document** covering the architecture, the configuration, the data model and metadata schema, the security boundary (Section 6), and how to operate and troubleshoot the system.
- **An operator runbook** for the self-serve update workflow (Section 7), including the verification step for confirming a withdrawal.
- **A deployment and recovery runbook:** how the system is deployed, how to restart components, how to rotate the API key, and how to restore from backup.
- **A live walkthrough** with the firm's nominated operators and IT contact, recorded for later reference.
- **The source code and configuration**, in a repository the firm owns.

The intent is that after handover the firm can run, maintain and extend the system without dependence on BODDAN.

12. Two-week post-delivery support

Following acceptance, BODDAN provides **10 business days** of bug-fix support, covering defects in the delivered system: errors in the pipeline, retrieval or citation behaviour that does not meet the accepted criteria, the update workflow, or the intranet integration.

Included	Not included
Fixing defects in delivered functionality	New features or scope additions
Correcting behaviour that fails an accepted acceptance criterion	Changes to the corpus content itself (a firm task)
Help diagnosing an issue traced to the delivered system	Issues caused by changes the firm makes to its own tenant, DMS or intranet after handover
Guidance on operating the self-serve workflow	Ongoing managed service or monitoring

Defects are triaged on report and worked within the support window. Anything outside bug-fix scope is quoted separately. Ongoing support beyond the two weeks is available as a retainer.

13. Limits and honest disclosures

This section is written plainly and is repeated in the intranet panel.

- **It assists retrieval; it does not give legal advice.** The assistant helps a fee-earner find and read the firm's own material faster. It does not advise, does not exercise legal judgement, and does not replace a lawyer's review. Output must be checked against the cited source before it is relied on.

- **It can only work from what it is given.** The assistant answers from the indexed corpus. If the answer is not in the corpus, the correct behaviour is to say so, and the system is built and tested to do that (A7). It does not draw on general world knowledge to fill gaps, and it should not be treated as a general legal research tool.
- **Citations point to sources; they do not certify them.** A citation tells the user where the answer came from. It does not warrant that the source is current law or fit for the matter at hand. Currency of precedents is maintained through the update workflow (Section 7), which is a firm responsibility.
- **Quality is measured, not guaranteed.** The acceptance thresholds in Section 9 are measured against a firm-provided question set. They are the agreed bar for this build on that set. They are not a guarantee of accuracy on every future question, and no such guarantee is given.
- **A boundary exists at query time.** As Section 6 sets out, the user's question and the retrieved snippets are sent to the Anthropic API at query time under the firm's key. The firm should be comfortable with that boundary and with Anthropic's terms before go-live. The in-country inference option (Section 6.5) exists for firms that need inference to remain in Australia.
- **It is not a system of record.** The DMS remains the firm's source of truth. The assistant is a way to find and read what is already there.

14. Commercials and timeline

14.1 FEE

Fixed project fee: **AUD $24,500 ex GST.** GST at 10% is **AUD $2,450**, for a total of **AUD $26,950 inc GST**. The fee covers everything in Section 2.2, including the test plan, deployment, handover, walkthrough and the two-week support window. It does not include the firm's Anthropic API spend (Appendix B) or any Azure hosting cost, both of which the firm pays directly.

14.2 MILESTONES AND PAYMENT SCHEDULE

Milestone	Deliverable	%	Amount (ex GST)
M1	Discovery and solution-design sign-off	20%	$4,900
M2	Ingestion pipeline built and first index produced	25%	$6,125
M3	Retrieval, citation and intranet integration working	25%	$6,125
M4	Test plan passed, deployed, acceptance signed	20%	$4,900
M5	Handover, walkthrough and start of support	10%	$2,450
Total		**100%**	**$24,500**

14.3 INDICATIVE TIMELINE

Week	Focus
1	Discovery, question set assembly, access-model mapping (M1)
2 to 3	Ingestion pipeline, first ingestion, index build (M2)
4 to 5	Retrieval, reranking, access filtering, citation, intranet panel (M3)
6	Test execution, tuning, pilot, acceptance, deployment (M4)
7	Handover and walkthrough; two-week support begins (M5)

Total indicative duration is about seven weeks to handover, followed by two weeks of support. The schedule assumes the dependencies in Section 15 are met.

15. Assumptions and dependencies

- The firm provides timely access to the nominated DMS source location, or a controlled export and an agreed change feed.
- The firm provides its Anthropic API key and a key vault to hold it, and contracts directly with Anthropic for API use.
- The firm provides the target Azure tenant and the access needed to deploy into it, in the Australia East region.
- The firm provides Entra ID configuration and the SharePoint intranet surface for the panel.
- The firm assembles the test question set with BODDAN during discovery and nominates a pilot group.
- The firm nominates operators for the update workflow and an IT contact for handover.
- The corpus is approximately 1,200 documents in the supported formats; materially larger or more varied corpora may affect timeline and sizing and would be re-estimated.
- Acceptance thresholds in Section 9 are agreed at the start of testing.

Appendix A: Metadata field reference

Field	Description	Used for
`source_dms_id`	DMS document identifier	Citation resolution
`file_name`	Source file name	Display and citation
`document_version`	Version of the source document	Currency and supersession
`section_locator`	Heading or clause path within the document	Section-level citation
`document_type`	Precedent, template, policy, file note, other	Filtering and display
`practice_area`	Practice area tag	Filtering and narrowing

Field	Description	Used for
`matter_tags`	Matter or workspace tags	Access filtering
`access_labels`	Groups / matter-security markers governing visibility	Access-filtered retrieval
`lifecycle_status`	Current, superseded, or withdrawn	Update workflow; retrieval exclusion
`ingested_at`	When the chunk was indexed	Operations and audit
`source_modified_at`	When the source document last changed	Incremental sync

Appendix B: Cost model worked example

This appendix derives an indicative monthly Anthropic API cost from a single set of assumptions. The firm pays Anthropic directly; this is for budgeting only and is an estimate, not a quote. Anthropic's published prices are in US dollars; figures are converted at an illustrative rate of 1 USD = 1.52 AUD, stated here only so the working is transparent. The real cost depends on actual usage and the prevailing exchange rate. The figures below are in US dollars and exclude any GST or local taxes that may apply to overseas API spend; the firm should confirm the GST treatment of its Anthropic billing with its accountant.

Per-query assumptions:

Item	Value
Input tokens per query (system prompt + retrieved chunks + question)	5,500
Output tokens per query	700
Model	`claude-opus-4-8` at USD $5.00 / 1M input, USD $25.00 / 1M output

Per-query cost on `claude-opus-4-8`:

- Input: 5,500 / 1,000,000 x $5.00 = **USD $0.0275**
- Output: 700 / 1,000,000 x $25.00 = **USD $0.0175**
- Total: **USD $0.045 per query**

This per-query figure is deliberately gross: it bills the full 5,500 input tokens at the standard input rate and applies no prompt-caching discount. In practice the stable system-prompt portion of the input would bill cache reads at roughly a tenth of the input rate (Section 5.4), while the retrieved chunks and the question are not cacheable and bill at full rate. The estimate therefore sits at the high end of likely cost rather than the low end, which suits a budgeting figure.

Monthly volume assumptions:

Item	Value
Fee-earners	40

Item	Value
Queries per fee-earner per working day	8
Working days per month	21
Queries per month	**6,720** (40 x 8 x 21)

Monthly cost:

- `claude-opus-4-8`: 6,720 x USD $0.045 = **USD $302.40**, about **AUD $460 per month** at the illustrative rate.
- `claude-sonnet-4-6` (USD $3.00 / 1M input, USD $15.00 / 1M output) for the same volume: per query input 5,500 x $3.00 / 1,000,000 = $0.0165, output 700 x $15.00 / 1,000,000 = $0.0105, total **USD $0.027 per query**; 6,720 x $0.027 = **USD $181.44**, about **AUD $276 per month**.

Routing routine lookups to `claude-sonnet-4-6` and reserving `claude-opus-4-8` for nuanced questions therefore reduces spend; the split is a firm choice between cost and answer quality. These figures exclude the firm's self-hosted compute (embedding, reranking, database) running in its Azure tenant, which is a separate hosting cost.

Appendix C: Glossary

Term	Meaning
RAG	Retrieval-augmented generation: retrieve relevant text, then have a model answer from it
Chunk	A passage of a document, sized for retrieval, that can be embedded and cited
Embedding	A numeric vector representing a chunk's meaning, used for semantic search
Vector search	Finding chunks whose embeddings are closest to the question's
BM25	A standard keyword-ranking method used for lexical search
Hybrid retrieval	Combining vector and keyword search
Reranker	A model that reorders candidate chunks by direct relevance to the question
Citations	Anthropic's feature that ties an answer to the specific source spans it used
Ethical wall / information barrier	A restriction preventing some staff from accessing certain matters
DMS	Document management system: the firm's system of record for documents
Entra ID	Microsoft's identity service, used here for single sign-on
pgvector	A PostgreSQL extension for storing and searching vectors
Supersession	Marking a document replaced so it drops out of retrieval

Term	Meaning
ZDR	Zero data retention: an arrangement under which API inputs and outputs are not retained

Prepared by BODDAN, an Australian veteran-owned professional document service. This is a disclosed fictional sample for demonstration of deliverable quality. It is not legal advice and does not describe a real client engagement.

PREPARED AND ISSUED BY

Kyle Boddan

KYLE BODDAN · BODDAN

01 July 2026

Decorative mark for presentation, not a legal signature.