AnovaCloud

FREE PRACTITIONER PACK

Enterprise RAG Build Pack

Discovery questionnaire — score each item honestly; low scores are the point: they show where the program will hurt.

Pack	Enterprise RAG Build Pack
Document	Discovery questionnaire (24 questions)
Response scales	Maturity 1–5 · Yes / Partially / No · Written
Date	2026-09

AnovaCloud — Engineering the AI-Native Enterprise. Frisco, Texas. This pack is a working instrument, not a brochure: complete it with your team, score honestly, and bring it to your discovery call.

HOW TO USE

Run this like a workshop, not a survey

Complete it with the stakeholders named in each section's intro. One scribe, one room (or call), 90 minutes. Score from evidence — monitoring data, documents, demos — not opinions. Any item scoring 1–2 is a program risk: it needs an owner and a close date before you commit budget. Bring the scored questionnaire to your discovery call; it compresses weeks of fact-finding into one session.

1 · KNOWLEDGE SOURCES

Knowledge sources

RAG is a corpus business wearing a model costume.

Q1. What are the source systems (wikis, tickets, drives, code), and what is the document count and growth rate?

Evidence: source inventory with counts.

Q2. What share of the corpus is current, duplicated, or contradictory?

Evidence: sampling audit results.

Q3. Who owns each source's accuracy, and is there a content lifecycle (review/expiry)?

1	2	3	4	5

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: ownership matrix.

Q4. Are there sources that must be excluded (privileged, draft, personal data)?

☐ Yes ☐ Partially ☐ No

Evidence: exclusion list.

2 · RETRIEVAL QUALITY

Retrieval quality

If retrieval is wrong, generation is theater.

Q5. What is the expected query profile (factual lookup vs. synthesis vs. troubleshooting)?

Evidence: sample of 50 real user questions.

Q6. What chunking strategy fits the corpus (and has it been tested, not assumed)?

1	2	3	4	5

Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: chunking experiments with recall@k.

Q7. Is hybrid retrieval (vector + keyword) needed for exact terms (IDs, codes, names)?

☐ Yes ☐ Partially ☐ No

Evidence: test queries with exact-match requirements.

Q8. Is reranking planned, and on what candidate depth?

Evidence: latency/quality trade-off notes.

3 · PIPELINE & FRESHNESS

Pipeline & freshness

Stale answers erode trust faster than wrong ones.

Q9. What is the required freshness SLA per source (minutes/hours/days)?

Evidence: per-source SLA table.

Q10. Is ingestion incremental (CDC/webhooks) or full re-crawl?

☐ Yes ☐ Partially ☐ No

Evidence: ingestion architecture.

Q11. How are document updates, deletions, and permission changes propagated to the index?

1 2 3 4 5 Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: propagation design + lag measurement.

Q12. What happens to answers when a source document is corrected?

Evidence: correction-propagation test.

4 · SECURITY & ACCESS CONTROL

Security & access control

The model must not see what the user may not see.

Q13. Are document-level ACLs enforced at retrieval time (pre-filter, not post-generation)?

☐ Yes ☐ Partially ☐ No

Evidence: architecture showing ACL filter placement.

Q14. How are permission changes (role moves, terminations) reflected, and within what lag?

1 2 3 4 5 Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: lag measurement or design target.

Q15. Is PII in the corpus identified, and is redaction or access-tiering applied?

1 2 3 4 5 Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: PII scan results.

Q16. Are prompts, retrieved chunks, and answers logged — and is that logging itself compliant?

☐ Yes ☐ Partially ☐ No

Evidence: logging + retention policy.

5 · EVALUATION & GROUNDING

Evaluation & grounding

Trust is measured, not asserted.

Q17. Is there a golden Q/A set with source-grounded expected answers?

☐ Yes ☐ Partially ☐ No

Evidence: golden set with N pairs and citations.

Q18. Are faithfulness (no unsupported claims) and citation accuracy measured per release?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: eval metrics from the last run.

Q19. Is there a 'don't know' behavior spec — when must the system refuse rather than guess?

☐ Yes ☐ Partially ☐ No

Evidence: refusal test results.

Q20. Are human reviewers grading a sample of answers on a cadence?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: review cadence + rubric.

6 · SERVING & SCALE

Serving & scale

A pilot that can't get scale is a demo with a budget.

Q21. What are the concurrency, latency (p95), and availability targets?

Evidence: written SLOs.

Q22. What is the cost per 1k queries at pilot and at 10x scale?

Evidence: cost model (embeddings + LLM + infra).

Q23. How are model, embedding, and index versions managed and rolled back?

1 2 3 4 5 Maturity 1-5 (1=Ad hoc, 2=Repeatable, 3=Defined, 4=Managed, 5=Optimized)

Evidence: versioning/rollback procedure.

Q24. What is the feedback loop (thumbs down → corpus or retrieval fix)?

☐ Yes ☐ Partially ☐ No

Evidence: feedback-to-fix workflow.

SCORING

Read the pattern, not just the average

Average each section. Sections averaging below 3 are your critical path — sequence the project plan so those risks close first. A single 1 on a load-bearing question (downtime budget, rollback, access control) outweighs a 4 average elsewhere. Record owners and close dates below.

Lowest-scoring section:

Single biggest risk:

Owner:

Close date: