Retail Policy Q&A Assistant (RAG) — End-to-End Example

1) Discovery meeting

Goal: Align on problem, value, constraints.
Agenda: intro → current process → pain points → data tour → risks/compliance → success metrics → next steps.
Key questions:

- What questions do employees/customers ask most? What's the cost of slow/incorrect answers?
- Where do the answers live (PDFs, SharePoint, Confluence, ticketing history)? Who owns them?
- Required accuracy/latency? (e.g., ≥85% exactness)
- Who's allowed to see what? PII/PCI/PHI considerations, data residency?
- Failure modes: what should the bot do when unsure?
- Change cadence: how often do policies change? Who updates the source of truth?
- Success = …? (deflect 30% tickets, cut handle time 25%, CSAT +0.3)

Outputs: problem statement, success metrics, constraints, stakeholders, initial system boundaries.

---

2) Requirements

User stories

- *As a store associate,* I want to ask policy questions ("What's the return window for electronics?") and get a sourced, up-to-date answer in <3s.
- *As compliance manager,* I need answers to include citations and to log every response for audit.
- *As content owner,* I want to be able update a policy and see it reflected in <1 hour.

Non-functionals

- p95 latency < 2.5s; uptime 99.9%; support 5k DAU.
- Guardrails: no personal data in prompts; redact PII; refusal on out-of-scope.

  Sidenote: p95 means the **95th percentile** of a set of values.

  - **Plain English:** 95% of the observations are **at or below** this value; the slowest/top 5% are **above** it.
  - **Example (latency):** "p95 latency = 2.4s" ⇒ 95% of requests finish in ≤2.4 seconds; 5% take longer.
  - **Why it's used:** Averages hide tail pain. Percentiles (p50/median, p90, p95, p99) show real user experience, especially the slow tail.

Acceptance criteria

- ≥85% *correctness* on a blinded test set; ≥95% answers include at least one valid citation.
- Hallucination rate ≤3% (measured by human eval with rubric).
- Security sign-off passed; audit logs retained 1 year.

---

3) High-level solution

Pattern: Retrieval-Augmented Generation (RAG) with policy PDFs/FAQs as the knowledge base.

Flow:

1. User asks a question in web/Teams/Slack.
2. Query → embed → vector search (top-k + re-rank).
3. Retrieved chunks + system guardrails → LLM generates answer *with citations*.
4. Safety checks (PII redaction, toxicity/refusal) → response + links.
5. Observability logs: prompt, retrieved docs, answer, feedback.

Core components

- Document pipelines: PDF→text, chunking (e.g., 1–2k tokens with overlap), metadata tags (policy version, owner).
- Vector DB (e.g., pgvector/FAISS/managed).
- Re-ranker (cross-encoder) to improve hit quality.
- LLM (managed API) with safe-prompting + tool use (search, escalate to human).
- Admin UI: document status, re-index, feedback review.

---

4) Data & governance

- **Sources:** Policy PDFs, SharePoint pages, Zendesk macros.
- **Access:** Service account with least privilege; nightly delta crawl + on-change webhooks.
- **PII:** Strip emails/order IDs at ingestion; runtime redaction in prompts & outputs.
- **Versioning:** Each chunk tagged with `policy_id`, `version`, `effective_date`. Answers display version.

---

5) Evaluation plan

Offline (before pilot)

- Build a 150-question test set from real tickets; label gold answers + citations.
- Metrics: top-k recall, exactness, groundedness, citation validity, latency.

Online (pilot)

- A/B: legacy FAQ vs. RAG assistant to 10% users.
- Success: deflection rate, time-to-answer, CSAT, override-to-human rate.

Human-in-the-loop

- Thumbs up/down with reasons; weekly error clinic to fix bad chunks/prompts.

---

6) Implementation timeline (example 6 weeks)

- **Week 1:** Discovery, access to data, security review, draft KPIs.
- **Week 2:** Ingestion pipeline & chunking; initial embeddings; vector DB stood up.
- **Week 3:** Retrieval tuning (k, chunk size, re-rank); prompt v1; guardrails.
- **Week 4:** Admin & analytics; audit logging; offline eval > target.
- **Week 5:** Pilot rollout to 10%; collect online metrics; iterate prompts/filters.
- **Week 6:** Org training; SLOs set; scale to 100%; handover & runbook.

---

7) Risks & mitigations

- **Stale content →** owners & webhook re-index; “This policy updated on…” banner.
- **Hallucinations →** strict *cite-or-say-I-don’t-know* prompting; retrieval-only answering for policy facts; uncertainty threshold to escalate.
- **Security →** prompt scraping defenses; no data retention at vendor; redaction.
- **Adoption →** embed where users already work; fast (<2.5s) or they’ll bypass it.

---

8) Example artifacts (ready to reuse)

System prompt (excerpt)

```
You answer retail policy questions using ONLY the provided context.
If the answer isn’t fully supported, say “I don’t have enough
information.”
Always include numbered citations [1], [2] with titles and links from
the context.
Refuse questions that ask for personal data or out-of-scope topics.
```

Answer template (few-shot)

```
Q: What is the electronics return window?
A: Electronics can be returned within 30 days if unopened. Open-box
items are 15 days. [1][2]
Limits: Extended holiday returns apply Nov 1–Jan 15. [3]
If unsure: "I don't have enough information…"
```

Retrieval parameters (starting point)

- chunk_size: 1200–1600 tokens, overlap 150–200
- k: 8 initial, re-rank to top 3
- filters: `department:retail`, `doc_type:policy`

Test cases (sample)

- Edge: "Is a 32-day return allowed if receipt is lost?" → Expect refusal + escalation note.
- Ambiguous: "What about returns?" → Ask clarifying question.
- Off-policy: "Can I return worn shoes?" → Cite wear-and-tear exception.

Go-live checklist

- SOC2/DPoA approved

  Sidenote:

  - **SOC 2 approved** → The vendor has passed a **SOC 2** audit (System and Organization Controls, Type I or Type II) covering controls for **Security, Availability, Processing Integrity, Confidentiality, and/or Privacy**. In practice this means your security team has reviewed the vendor's SOC 2 report (often Type II + a current bridge letter) and signed off.
  - **DPoA approved** → This is almost certainly a shorthand/typo for **DPA** (Data Processing Addendum/Agreement). "Approved" means your legal/privacy team has executed and accepted the **DPA** with the vendor (GDPR roles, SCCs, data locations, sub-processors, retention, breach notice, etc.).

- If it truly says **DPoA**, ask what policy it refers to—some orgs use it to mean "Data Protection Office Approval" or an internal privacy approval step, but **DPA** is the standard term.

- Data cataloged; owners listed
- Rollback switch & rate limits
- Dashboards: latency, retrieval recall, hallucinations, feedback
- Runbook: incident, drift, re-index

---

9) Handover & operations

- Weekly content sync with policy owners.
- Monthly evaluation with fresh tickets; refresh test set.
- Alerting: spike in “I don’t know” or citation failures.
- Prompt & retrieval reviewed on drift or KPI regression.