

Handling routine customer requests

THE WORKFLOW AS IT RUNS TODAY

≈ 3,000 requests a month · ≈ 60% fall into three request types · ≈ 1 day first response time (illustrative, approximate figures)

● friction point



BEFORE YOU START

Get the sources ready.

Approve the standard replies, resolve conflicts, assign an owner to each document, and define which data can be accessed. The tool comes after.

UNDER THE HOOD

A hybrid architecture: order data accessed through a read-only API, limited to the authenticated customer; hybrid search (vectors + BM25) with reranking across the terms of sale, the FAQ, and the product guides; an LLM instructed to rely on this material and cite the source; input and output guardrails with routing to agents. A learning loop connects the log, team review, and the evaluation set. Diagram and controls on page 2.

STAGE 1 HUMAN IN THE LOOP

The assistant drafts, the agent decides

- 1 The request comes in — through the same queue as today.
- 2 The assistant drafts a reply — citing the source or the order.
- 3 An agent decides — reviews, edits if needed, then sends.
- 4 Every correction is logged — request type by request type.

What this stage produces: a measure, by request type, of how good the drafted replies are. Estimated duration: 8 to 12 weeks.

STAGE 2 THE ASSISTANT ANSWERS STANDARD REQUESTS

It answers approved request types; agents handle the exceptions

ANSWERS WHEN

- The request falls under an approved type
- The customer is authenticated and the data is current
- The reply can cite its source or the order

ROUTES TO AN AGENT WHEN

- It is a complaint, a dispute, or an ambiguous case
- The source or data is missing, conflicting, or outdated
- The request involves an issue such as a refund, goodwill credit, damage, or product safety

Every week, the customer service team reviews a sample of replies and keeps the sources current.

WHY HERE, AND UNDER WHAT CONDITIONS

High volume · Clear owner · Clear output · Known pain point

- **Informational requests** order status, terms of sale, product guides.
- **Order data** read-only access, authenticated customers only.
- **No commercial decisions** no refunds or goodwill credits left to AI.
- **Complaints, disputes, and sensitive cases** routed to an agent.

The risk does not go away, but it becomes more visible, measurable, and manageable.

MOVING FROM STAGE 1 TO STAGE 2

Decided by request type, never in advance, based on what the team observes during stage 1:

- Accuracy of the reply and the cited source
- Corrections needed, by request type
- Exceptions routed correctly
- No significant incident observed within the tested scope
- Current sources, each with an owner

Order status can move to stage 2 while product guides stay in stage 1.

THE GOAL, AND HOW TO MEASURE IT

Not to replace agents, but to cut response times on informational requests, stop rewriting the same replies, and free up time for cases that call for judgment.

First response time on a standard request

≈ 1 day → ≈ 10 min

today (illustrative) → stage 2 target · also tracked: the share of requests handled without an agent, 40–50% (illustrative target)



A reference set of about 150 real, anonymized requests. Metrics: accuracy, faithfulness to sources and data, correct citation, and the share of complaints, disputes, and out-of-scope cases detected and correctly routed (routing recall). Rerun after every change to the model, prompt, or sources.

For each request, only what is needed is logged: the minimized request, references to the data and passages retrieved, the reply, control signals, and the routing decision. Access is limited to authorized staff, with a defined retention period.

Customer authenticated before any order data is shown. Access limited to the customer's own orders, enforced at retrieval, not after generation.

Hosting and data transfers verified for the chosen solution; contractual terms barring use of the data for model training; encryption; log retention and access defined; Swiss FADP assessment documented for the system as deployed.

Prompt injection defenses, including in incoming emails. The orchestration layer limits outbound actions to two: reply or route.

Three paths: the AI module of the existing customer service platform, a managed RAG platform connected to order data, or a custom build (orchestration, vector database, model API). The choice comes after the source audit, not before.