

WHERE TO START WITH AI · NO. 3

# Handling routine customer requests

A customer request assistant, rolled out in two stages: first it helps agents answer, then it answers the approved standard request types on its own.

Stage 1 · the agent answers, with AI support

Stage 2 · AI handles standard requests, agents the exceptions



## WHY START HERE

# It meets all four tests of a good first workflow



### High volume

Thousands of requests a month, across a handful of request types.



### Clear owner

The head of customer service, already accountable.



### Clear output

An accurate answer that cites its source.



### Known pain point

For agents and customers alike.

**A closed set of sources.** The assistant answers from data and documents the company already owns. The risk does not go away. It becomes more visible, measurable, and manageable once the sources, the scope, and the escalation rules are defined.



## ON ONE CONDITION

# Low risk, but only within a strict scope

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

Outside this scope, it becomes a different use case with a different risk profile.



THE WORKFLOW AS IT RUNS TODAY

# Six steps, all done by hand

Example Equipment Inc. · 600 employees · fictitious company

|   |                                                                      |                               |
|---|----------------------------------------------------------------------|-------------------------------|
| 1 | <b>Request received</b><br>by email, web form, or phone              |                               |
| 2 | <b>Read and categorized</b><br>in the team's shared queue            | ● A growing backlog           |
| 3 | <b>Information lookup</b><br>orders, carrier tracking, terms of sale | ● Three systems to check      |
| 4 | <b>Reply drafted</b><br>case by case                                 | ● The same reply rewritten    |
| 5 | <b>Supervisor input</b><br>when a case is out of the ordinary        | ● Handoffs without context    |
| 6 | <b>Sent and closed</b><br>reply to the customer                      | ● Request types poorly logged |

≈ 3,000

requests a month

≈ 60%

fall into three request types

≈ 1 day

first response time

● friction point · illustrative figures



**BEFORE YOU START**

# Get the sources ready

The work starts with the sources: approve the standard replies, resolve conflicts between the terms of sale, the FAQ, and the product guides, assign an owner to each document, and define which order data can be accessed. The tool comes after.

**1****Approve the standard replies****2****Resolve conflicts****3****One owner per document****4****Define data access****The tool comes after.**

## STAGE 1 · HUMAN IN THE LOOP

# The assistant drafts, the agent decides

**The request comes in**  
through the same queue as today.



**The assistant drafts a reply**  
citing the source or the order.



**An agent decides**  
reviews, edits if needed, then sends.



**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 HANDLES THE STANDARD CASES

# It answers approved request types; agents handle the exceptions

### THE ASSISTANT 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

### IT 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.



## UNDER THE HOOD · DATA, RAG, AND LEARNING LOOP

# Reference architecture: how the system stays under control



## UNDER THE HOOD · CONTROLS AND SECURITY

# The controls required before any autonomous answer

## 1 Evaluation

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. Rerun after every change to the model, prompt, or sources.

## 2 Observability

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.

## 3 Identity and access

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

## 4 Data

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.

## 5 Application security

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.



## MOVING FROM STAGE 1 TO STAGE 2

# A decision based on evidence, request type by request type

The move to stage 2 should not be decided in advance. It depends on what the customer service team observes during stage 1: the quality of the drafted replies, the corrections needed, and the system's ability to recognize requests that fall outside its scope.

---

### WHAT THE TEAM OBSERVES

- 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

The goal is not to replace agents. It is to cut response times on informational requests, stop the same replies from being rewritten, and give agents more time for customers who need to be heard or need a judgment call.

### 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)

#### Data protection

Swiss FADP: customers informed; data processed, hosting, and transfers clarified before launch.

#### Governance

The head of customer service owns the results and reviews a sample every week.



# In this example, AI is not meant to replace the agent.

It steps in first where data or a written policy already provides the answer.

So the useful question is not “Which tool should we buy?”

**It is “Which repetitive work can we have AI assist with, measure, and then delegate under clear controls?”**

**What repetitive work in your company already meets these conditions?**



**Sergio Castagna Consulting.**

sergiocastagna.com · info@sergiocastagna.com · Geneva