🧩 System Design Explanation

This system is designed as an **event-driven, AI-powered customer feedback automation pipeline** with a strong focus on reliability, scalability, and human-like communication.

At a high level, the architecture follows a **Listen → Analyze → Decide → Act → Record** pattern.

1️⃣ Event Ingestion Layer (Trigger)

Component: Gmail – Watch Emails
Design Choice: Event-driven polling

The system continuously monitors the Gmail inbox for new incoming emails. A subject-based filter ensures that only relevant customer feedback messages (e.g., *"customer review"*, *"customer feedback"*) enter the pipeline.

Why this matters

- Prevents unnecessary AI calls
- Reduces cost and noise
- Keeps the workflow lightweight and efficient

2️⃣ Intelligence Layer (AI Processing)

Component: Google Gemini 2.5 Flash-Lite
Design Pattern: Dual-stage LLM processing

The AI layer is intentionally split into **two independent responsibilities**:

A. Response Generation Engine

- Reads the full customer message
- Generates a polite, empathetic, and concise reply
- Personalizes the response using the customer's first name
- Enforces tone constraints (non-rude, calming, professional)

This ensures customers receive **human-quality replies**, not robotic autoresponses.

B. Issue Classification Engine

- Analyzes feedback content
- Categorizes issues into predefined buckets:
    - Late Delivery
    - Poor Chat Support
    - Poor Quality of Work
    - All Good

- Supports multi-label classification if multiple issues are detected

Why this separation matters

- Improves explainability
- Allows independent tuning of prompts
- Makes future model upgrades easier

3️⃣ Orchestration Layer (Decision Control)

Component: Router (Make.com)
Design Pattern: Parallel branching

Once AI processing is complete, the router splits execution into parallel paths:

- **Customer communication**
- **Data persistence**

This avoids tight coupling between actions and improves fault tolerance.

If one branch fails, the other can still succeed.

4️⃣ Action Layer (Customer Communication)

Component: Gmail – Send Email

The system automatically replies to the customer using:

- The AI-generated response
- The original email thread (context preserved)
- Clean formatting and professional subject handling

Design Goal

- Near-real-time response
- Zero human intervention
- Consistent brand tone

5️⃣ Persistence Layer (Data Logging)

Component: Google Sheets – Add Row

All interactions are logged into a structured spreadsheet, including:

- Customer email
- Subject line

- Original feedback
- Detected issue category
- AI-generated response

Why Google Sheets

- Lightweight CRM substitute
- Easy analytics & reporting
- Low operational overhead

6️⃣ Reliability & Safety Considerations

- No destructive actions (read + respond only)
- No overwriting of customer data
- Prompt-level AI guardrails ensure safe language
- Centralized logging for audits and QA

7️⃣ Scalability & Extensibility

The architecture is intentionally modular:

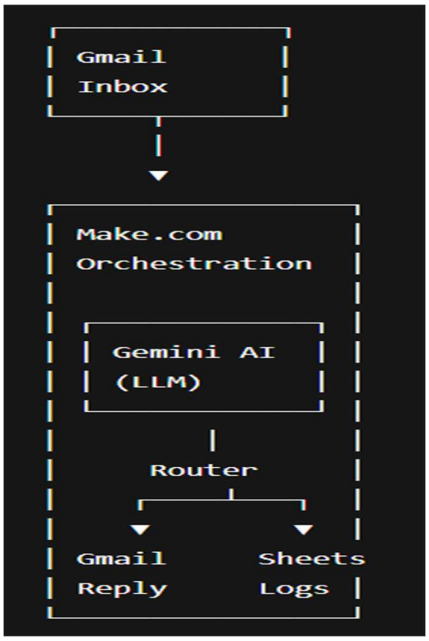
- Gmail → Can be replaced with web forms, Zendesk, Intercom
- Gemini AI → Swappable with OpenAI / Claude
- Sheets → Replaceable with CRM or database
- Router → Supports future business rules

This makes the system **enterprise-ready without rewriting the core logic**.

📊 Scalability & Growth Architecture

This system is designed with **horizontal scalability, modular expansion, and enterprise migration** in mind. Below are three progressive scalability stages showing how the workflow evolves as load and business complexity increase.

🔹 **Level 1: Current Scalable Design (Single-Team / SMB)**



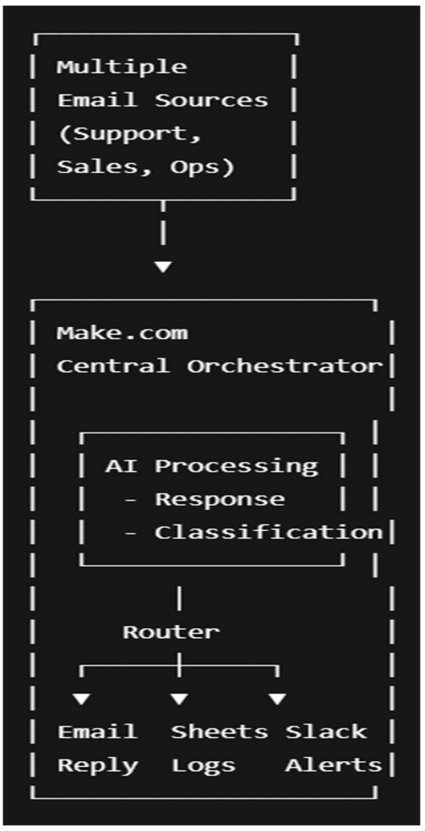
Why this scales

- Make.com handles retries, rate limits, and parallel steps
- Gemini Flash-Lite is optimized for low latency
- Google Sheets supports thousands of rows effortlessly

Capacity

- Hundreds of emails/day
- Near real-time responses
- Minimal ops overhead

🔹 Level 2: Team-Scale / Startup Growth



What changes

- Multiple inboxes feed the same logic
- Slack / Teams alerts added for critical issues
- AI prompts can differ per department

Capacity

- Thousands of emails/day
- Multiple teams supported
- Faster escalation paths

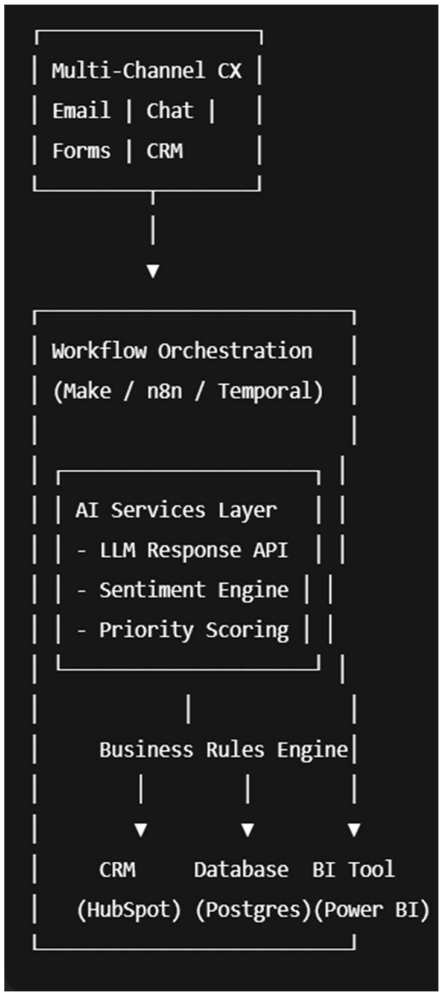
🔹 Level 3: Enterprise / Platform-Scale Architecture

Enterprise-grade benefits

- Channel-agnostic input
- Decoupled AI services
- Database-backed persistence
- BI dashboards for leadership

Capacity

- Tens of thousands of interactions/day
- SLA-based routing
- Full audit & compliance readiness



⚙️ **Horizontal Scalability Principles Used**

Principle	Implementation
Event-driven design	Gmail triggers
Stateless processing	Each email processed independently
Parallel execution	Router-based branching
AI decoupling	Separate prompts & roles
Replaceable components	Tool-agnostic design

🎱 **Design Philosophy**

“Automate the boring, preserve the human.”

The system automates repetitive support tasks while preserving empathy, personalization, and clarity—critical for customer trust.
