instinctools

AI Agents-Based Customer Support For an Online Store Serving 3+ Million Customers Yearly

How delegating the support ticket triage groundwork to an AI multi-agent system empowered a US online store to process support requests **4 times faster** and increase first-response time by **75%** without hiring more customer support specialists.

Industry:
Ecommerce

AI DEVELOPMENT

Business challenge

When the number of customer support requests hits 5,000 per day and spikes up to 10,000 on Black Friday and holidays, mostly manual handling becomes a bottleneck. Delays, inconsistencies, and escalating costs snowball by the day. Our client, a US-based online store with over three million shoppers, felt the pressure where it hurt most: in operational efficiency and customer satisfaction score (CSAT).

What issues pushed them to **look beyond simply hiring more people?**

Slow and costly process

On average, consumers expect their query to be answered within 5 minutes*. In the client's case, the waiting time could stretch to 12 minutes. This gap put the online store at risk of losing half their customer base, as only 1 in 2* shoppers is willing to stay with the company after a single poor support experience.

*According to Zendesk

Operational capacity lagged far behind demand on all fronts. Besides spending 6–12 minutes per request, the client's support specialist managed to process only 50–70 tickets daily. Meanwhile, to satisfy the demand, they needed to cover 200+ tickets daily and at a much faster pace. Tripling the headcount might have closed the volume gap, but this costly measure still wouldn't resolve the slow response time issue.

Error-prone ticket categorization and prioritization

Human agents sometimes mislabelled issues or overlooked important details when prioritizing tickets. These mistakes affected the first-response time, as the requests had to be recategorized and re-prioritized.

Slow knowledge scalability

Each time new communication guidelines, return and refund policies, and other updates came in, customer support staff needed time to adjust their daily routine accordingly.

The client considered using **AI virtual workers** to address the pressing issues, but needed a reliable AI engineering partner to act on the idea. Inctinctools proved to be the one up for the task and assembled a dedicated team of AI engineers, QA specialists, and a project manager.

Solution

Among the many forms AI virtual workers can take, we selected microservices-based agentic AI to transform the client's customer support.

AI agents facilitate:

- **Real-time ticket triage.** Tickets are categorized and routed in milliseconds upon arrival. Urgent issues are instantly flagged, improving the time to response.
- **Consistent categorization and prioritization.** AI agents treat every ticket objectively against the same standards, reducing errors.
- **Always up-to-date expertise.** New guidelines, policies, and rules are embedded in the AI models' databases and apply to support requests instantly.

Microservice architecture, where each agent is implemented as an independent microservice, enables:

- **Fast development.** The dedicated team can craft several AI agents in parallel.
- **No tech stack lock-in.** AI engineers build each agent using the ML models, programming languages, and libraries best suited for a particular role.
- **On-demand system scalability.** Microservices auto-scale up and down, readjusting to the workload, and can process thousands of tickets per minute.
- **High system stability.** If there're issues with one of the AI agents, it can be isolated without bringing down the whole system.

After analyzing the client's customer support workflow, we suggested crafting **six AI agents** working within a multi-agent (MAS) pipeline:

→ 01 **An ingestion and pre-processing agent** removes personally identifiable information (PII) and other sensitive data and attaches metadata important for prioritization, such as account tier.

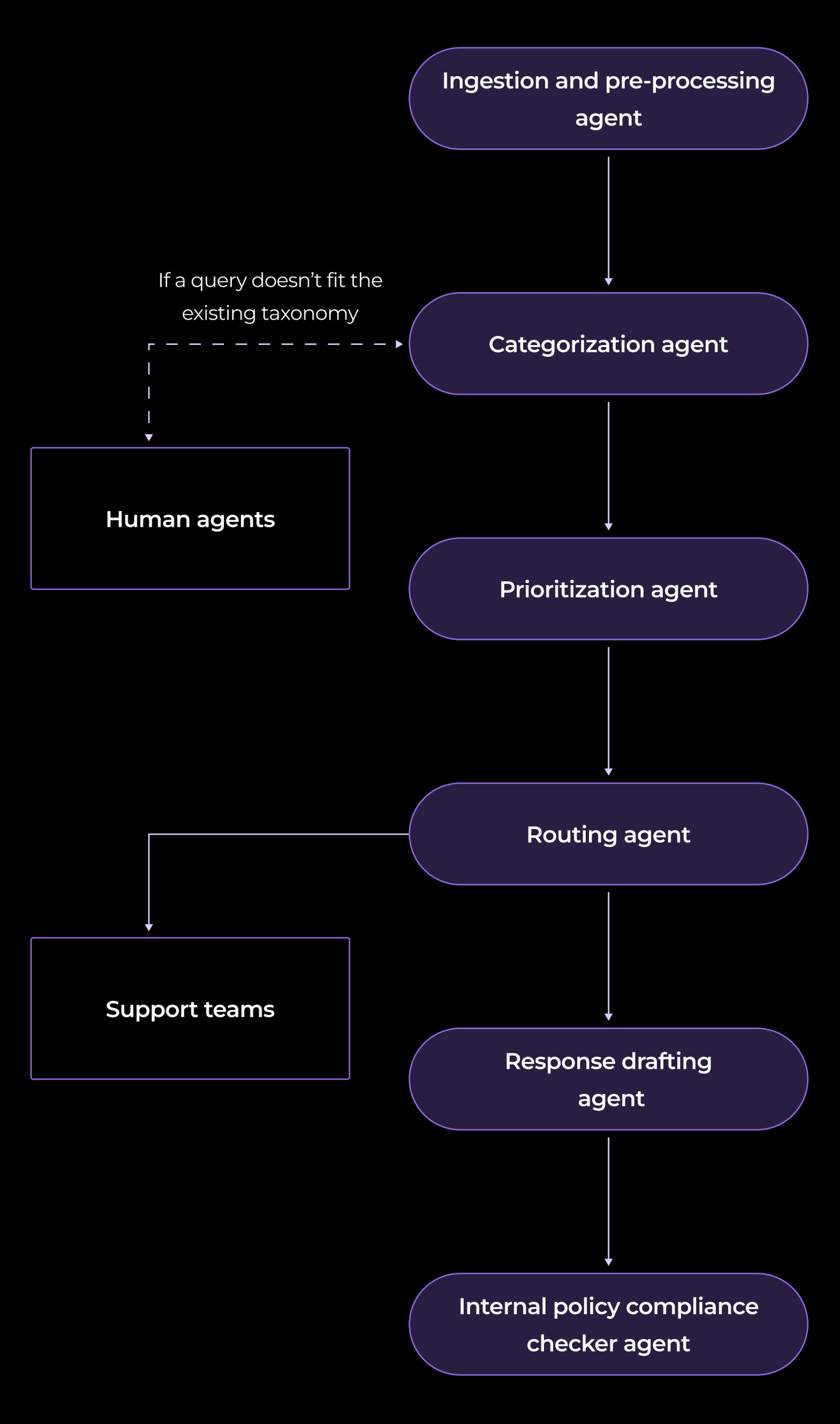
→ 02 **A categorization agent** analyzes the request and puts it into a relevant issue category, say, 'wrong product,' 'delivery status,' 'billing issue,' etc. If the query doesn't fit the existing taxonomy, the agent escalates the ticket to human supervisors.

→ 03 **A prioritization agent** checks the text for urgency markers (words and phrases such as 'ASAP,' 'immediately,' 'I've already written five times about this,' etc.), factors in metadata from the ingestion agent, and assigns priority level from P1 (Critical) to P4 (Low).

→ 04 A routing agent determines which of the available support teams or individual agents fits best for addressing the issue and assigns the task.

→ 05 **A response drafting agent** outlines a preliminary answer to the customer's query.

→ 06 **An internal policy compliance checker agent** inspects whether the proposed response aligns with the company's guidelines on communication with customers and other policies applicable to the issue, for example, return and refund rules.



01 Data classification and model training

Most AI agents mentioned above were trained on **thousands of past customer support tickets.** However, training the internal policy compliance checker required a different approach. There simply weren't enough examples of requests that violated the company's guidelines to train the model effectively. To solve this, we generated a **synthetic dataset of 2,000 conversations** that deliberately broke internal guidelines around customer communication, returns and refunds, accessibility and inclusion, and other policies.

Once the datasets were ready, our team took on data analysis, classification, and labeling to create high-quality datasets for ML model training. As a result, we got a clear picture of the **14 broad categories of customer support requests with 75 subcategories.**

The labeled and categorized data then served as the foundation for training the ML models at the core of AI agents. We decided to leverage the **pretrained models.**

GPT was a perfect fit for drafting a response agent, as the LLM excels at generating polite and brand-safe answers out of the box.

BERT was a natural starting point for the other tasks, given that the model was already trained on the BookCorpus of 800M words and can outperform models like Claude, Llama, Mistral, and others on sentiment analysis tasks.

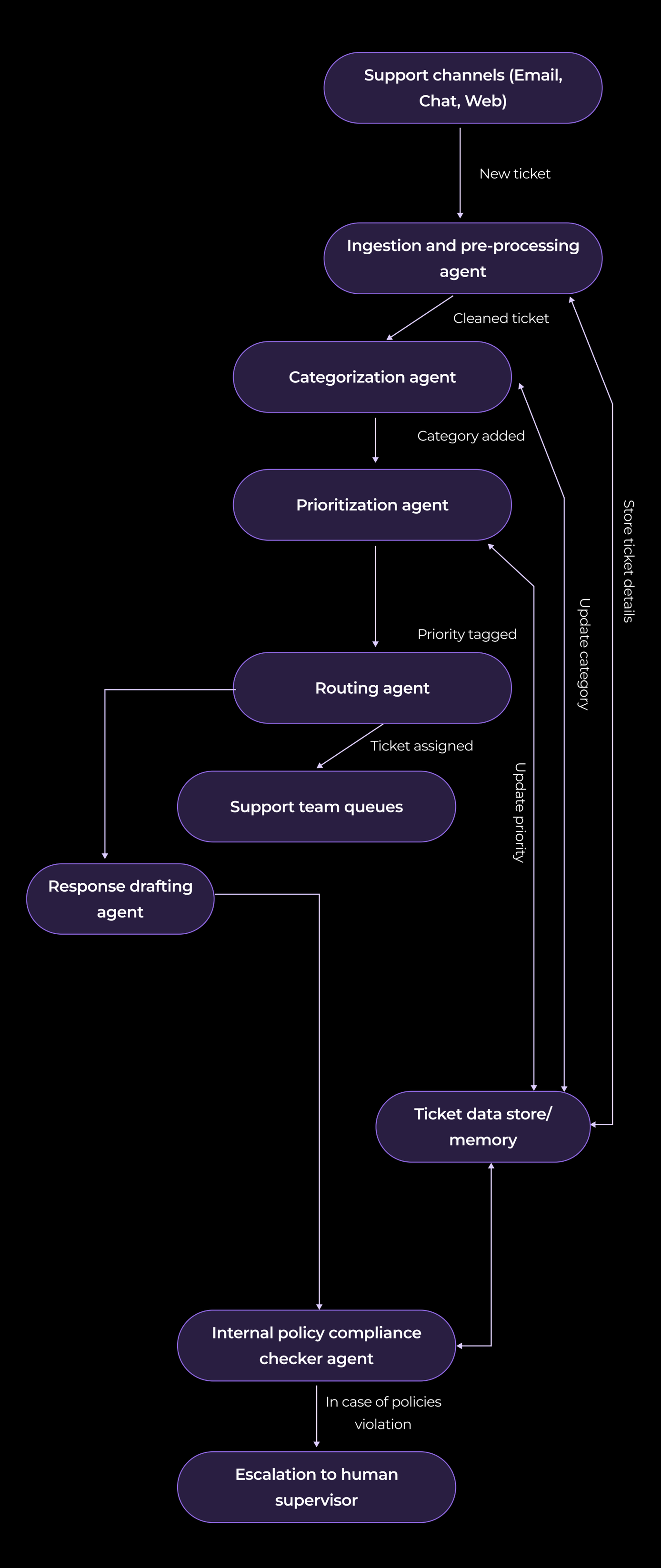
Nevertheless, BERT didn't cut it as its inference speed (350 ms) wasn't fast enough to ensure a seamless customer experience. Therefore, we reevaluated available ML models and switched to **DistilBERT**, which retains 97% of BERT's language understanding but is 60% faster. In reality, the new model showed even better results — we decreased the **inference speed to 98 ms.**

Moving further, we evaluated the model's performance. It showed **87% accuracy** in categorization, prioritization, routing tasks, and policy compliance checks.

While this level of accuracy was sufficient, we didn't stop there and leveraged retrieval augmented generation (RAG) to raise the bar even higher. For instance, we enabled more dynamic, context-aware reactions by allowing DistilBERT to access product reviews in real time. The **accuracy of AI agents' outputs spiked to 92%.**

02 How our AI multi-agent system works in practice

Here's an example of our MAS handling a customer request.



A VIP customer wrote to customer support on the website:

> *Hello, I placed an order two weeks ago. #382911-A, under jane.doe@example.com. The tracking still says 'label created'. I just checked my credit card statement and I've been charged twice — once on May 8 and again on May 9. My card ends in 3711. What's going on? Better sort it out fast.*

The **ingestion and pre-processing agent** removes the email and card's last digits as it's either personal or sensitive data. It also identifies the VIP account tier and attaches the corresponding meta tag to the request.

Moving further, the **categorization agent** identifies intent as billing issues and delayed shipment.

Next, the **prioritization agent** steps in and factors in metadata and urgency indicators in the customer's request, such as "better sort it out fast." Coupled with the VIP status, this ticket gets the highest priority level – P1.

From there, the **routing agent** takes on and finds the right team to resolve the customer's issues. Since there're two problems, it auto-creates two linked sub-tickets and routes them to the billing and logistics departments.

The **response drafting agent** generates a clear, action-oriented message designed to restore trust and resolve the issue quickly:

> *Hi Jane, thanks for reaching out, and I'm really sorry for the confusion and delay. I've located your order (#382911-A) and escalated it to our billing and logistics teams for immediate review. It appears you were accidentally charged twice, and we're processing a refund for the duplicate charge. You'll receive confirmation shortly. Meanwhile, I've flagged your shipment for manual tracking so we can get it moving ASAP. Thanks again for your patience — we're on it!*

The **internal policy compliance checker agent** verifies that the tone of voice adheres to brand standards, no PII is exposed, the refund policy is mentioned accurately, and handles the draft to a human in the loop.

All these operations, that previously took 6–12 minutes per each request, are now done within **one (!) second**. The c**ustomer support specialist only needs 1–3 minutes to review the AI-generated message and fine-tune it**, if needed.

Such well-orchestrated human-AI agents collaboration enabled the client to **reduce ticket handling time by 75%.** An unaided support specialist could process a maximum of 70 tickets daily. With AI agents covering the groundwork, humans now each manage up to 250 tickets a day within the same working hours.

Before

- Human error-prone support request categorization, prioritization, and routing
- Ticket processing time is 6–12 min
- A single customer support specialist can process 50–70 tickets daily, depending on their complexity
- Need to triple the headcount of their support team to tackle the rising number of requests efficiently

After

- Near-zero error rate at any stage of support request processing
- Ticket processing time is 1–3 min
- With an AI multi-agent system handling the groundwork, a single customer support specialist can now process 200–250 tickets daily
- Support team seamlessly handles 4 times more tickets within the same working hours

Business value

- Real-time ticket triage that was beyond the reach of mostly human-led customer support
- ×4 faster overall support request processing without increasing the number of customer support specialists
- 75% reduction in first-response time
- 15% CSAT uplift

Multiplier effect

Customer support isn't just a service touchpoint. It's what turns one-time buyers into repeat customers. In fact, 93%* of consumers are more likely to make repeat purchases from companies that offer excellent customer service.

*According to HubSpot

This impact goes far beyond retail. For companies from any industry, from healthcare to banking to logistics, the speed, quality, and responsiveness of customer support directly affect customer retention, operational efficiency, and overall business profitability.

AI agents offer a smart and cost-effective path forward, helping companies scale support operations without growing headcount.

Do you have a **similar project idea**?