JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Optimizing Enterprise Support Workflows: A Hybrid Orchestration Pipeline For Automated Triage And Multi-Modal Data Acquisition

Manav Kheni
College of Engineering, Northeastern University
Boston, MA, USA

Abstract

This research investigates how organizations can improve their support operations while reducing security threats through a hybrid orchestration system which combines low-code automation with Large Language Model (LLM) API integration. This research looks at two key problems in today's business world: the rise of unauthorized 'Shadow AI' workflows and the slow response times of traditional support systems.Right now, because of how departments are separated, employees often end up using AI tools that aren't monitored, which creates real security risks.To reduce these risks, we created the Intelligent Support Agent, a centralized system that works within a governed framework. I developed a centralized "Intelligent Support Agent" which operates within a controlled system that receives monitoring to solve these problems. The system provides users with a Gradio-based interface which enables secure access through n8n orchestration that automatically handles data collection from multiple sources, which include webhooks, news APIs, and RSS feeds. The experimental implementation of this workflow demonstrated a 60% decrease in manual research work because it automated the triage and data enrichment operations which used to need substantial human work. The research findings show that organizations can achieve better operational efficiency through hybrid orchestration systems which provide centralized management of AI workflows that represent a scalable replacement for complicated local LLM infrastructure deployments.

Keywords: Shadow AI, Low-code Automation, LLM Orchestration, n8n, Enterprise Governance, Customer Support Workflow, Hybrid AI Systems**.**

INTRODUCTION

More and more organizations are turning to cloud-hosted Large Language Model APIs like GPT-4 to boost their customer service and internal support as they grow.These services have strong generative abilities and can scale across different areas, but they also bring up important issues with data governance and system delay.Recent studies show that more than 90% of employees use their own personal AI accounts for work, but only about 40% of companies offer official LLM tools for them to use.This gap causes a lot of "Shadow AI" behavior, where employees avoid secure systems and use unapproved tools instead. This leads to data leaks, breaks compliance rules, and lets information flow without proper monitoring.Relying on these different tools manually often leads to uneven response times and creates big slowdowns in operations.

To fill this gap, we created the Intelligent Support Agent, a hybrid orchestration system that's ready for production and built to handle triage while staying compliant with regulations.The system brings AI interactions together in one place using a secure Gradio-based interface, supported by n8n, which is a low-

code platform for workflow automation that you can host yourself.This pipeline makes sure data flows are clear and easy to track, while using modern LLM APIs to handle analysis, summarization, and routing.Automating data intake from various sources like webhooks, news APIs, and RSS feeds, and managing these processes through a monitored workflow, helps organizations cut down on Shadow AI risks while keeping their operations efficient and responsive like modern autonomous systems.

Related Work

Shadow AI and Enterprise Governance

Shadow AI brings risks that go beyond regular shadow IT, like exposing data to third-party models without permission and getting around company controls by using personal accounts.IBM's 2025 Cost of a Data Breach Report shows that breaches linked to AI are now costing companies a lot more each time, mainly because they lack strong governance frameworks.To reduce these operational, legal, and reputational risks, organizations need more than just basic policies. They have to set up technical controls like centralized oversight teams, regular audits, and approved, monitored access points for AI tools.

Workflow Automation and LLM Integration

Bringing Large Language Models (LLMs) into business operations means you need careful planning to handle the tricky parts of logic and pulling in the right data.n8n, which is a source-available workflow automation platform, has become an important tool in this area.Its node-based, event-driven setup lets teams build scalable pipelines using very little code. It links different services like webhooks, RSS feeds, and APIs into one smooth system.Recent studies show that n8n works well for managing different AI types in areas like supply chain management and automated customer service. It proves to be a practical tool for handling AI interactions in one central place.

Latency in LLM-Based Support Systems

A big challenge when using LLM-based support agents is dealing with latency.Typical API setups usually cause about 2 to 3 seconds of delay for each interaction because they handle tasks one after another: first converting Speech-to-Text (STT), then running the LLM inference through REST APIs, and finally doing the Text-to-Speech (TTS) synthesis.Studies show that even a single extra second of delay can lower customer satisfaction scores by about 16%.Optimized GPT-4 setups can usually run with latencies around 800 to 1,200ms, but regular versions often fall somewhere between 1,200 and 1,800ms.To address this, many modern systems are starting to use WebSocket streaming along with hybrid orchestration to cut down on the "time-to-first-token" and make things feel faster for users.This work tackles these challenges by suggesting a governed, low-latency support pipeline.We’re using a monitored n8n workflow combined with a Gradio-based interface to centralize LLM orchestration. This setup helps cut down on Shadow AI risks, while still keeping things fast and efficient enough for solid support work.

Methodology: The Orchestration Layer

The Intelligent Support Agent has a layer, at its center. This layer helps the Intelligent Support Agent handle lots of data from places. It makes sure this data is clean and safe to use. The Intelligent Support Agent also uses intelligence to make good decisions. This way the Intelligent Support Agent can process all the data it gets in a way. It cleans the data and keeps a record of it. This helps get rid of the problems that can come with something called Shadow AI and the Intelligent Support Agent.

Multi-Source Data Acquisition

Moving beyond traditional single-source ingestion methods, the system implements a concurrent multi-stream data acquisition strategy to aggregate intelligence relevant to support operations.

- **Parallel Ingestion:** The architecture utilizes an event-driven model to fetch data concurrently across multiple vectors, including webhook endpoints for direct user queries, external News APIs for market context, RSS feeds, and YouTube API integrations. This broadens the context window, allowing the agent to evaluate isolated support requests against external variables.
- **Synchronization Barriers:** To manage asynchronous API requests effectively, the workflow implements synchronization barriers that suspend execution until all parallel data streams resolve. An aggregation module then normalizes these disparate streams into a unified, structured JSON payload. This state-management approach prevents race conditions and ensures the LLM receives a comprehensive context window prior to inference.

Digitization and Pre-processing

To prepare diverse content types for Large Language Model processing, the workflow executes automated data normalization and batch segmentation.

- **Multi-Modal Normalization:** Multi-Modal Normalization is used to get information from things that're not text. We use tools to look at things like PDF documents or pictures that people send us. These things are then changed into text that computers can understand so we can look at them later.
- **Payload Segmentation:** Payload Segmentation is when we take a group of data and break it down into smaller pieces. For example we might have a lot of support tickets like 117 of them. We break them down into smaller groups so we can look at each one on its own. This way each ticket is looked at one by one. We make sure that the information from one ticket does not get mixed up with the information, from another ticket when we are using Multi-Modal Normalization and Payload Segmentation to evaluate them.

Sentiment Analysis and Logic Branching

The final processing stage integrates LLM inference with deterministic logic rules to achieve automated, high-fidelity triage.

- **Cognitive Processing:** A decoupled LLM inference engine is invoked to execute natural language tasks, including sentiment analysis, intent classification, and automated summarization. This transforms unstructured, high-entropy inputs into standardized artifacts optimized for human analyst review.
- **Deterministic Conditional Routing:** The system utilizes boolean control flow mechanisms to route enriched data. Payloads that trigger critical criteria (e.g., semantic flags for "broken product," "refund request," or "security concern") satisfy the conditional logic and are automatically formatted and routed to high-priority escalation queues. Payloads falling outside these parameters are directed toward standard queues or automated archival storage.
- **Immutable Audit Trail:** To meet the rules that big companies must follow every single thing that is said every answer, from services and every step that is taken is written down forever in the history of the system that puts everything together. This makes a record that can be traced and shows if someone tried to cheat which helps to reduce the problems that can happen when Artificial Intelligence is used without being watched the kind of Artificial Intelligence that is used without permission, which is called Shadow Artificial Intelligence.

System Architecture

The Intelligent Support Agent is built with a decoupled, event-driven setup that keeps data ownership intact and makes it easy to swap out parts. The system has five different layers:

- The user interface is a web app built with Gradio and hosted on Hugging Face Spaces, acting as the secure control point. This interface lets analysts submit queries, check out enriched tickets, and look over AI-generated summaries and routing suggestions without needing direct access to the API keys.
- The main logic is handled by n8n workflows running on self-hosted infrastructure. This layer handles all asynchronous data collection, keeps track of API rate limits, and runs the conditional logic needed for triage.
- The Data Ingestion Layer is a part that runs alongside everything else to bring in outside information.This includes connectors for the YouTube API, RSS feeds, News APIs, and Webhook endpoints that handle support tickets coming from other CRM tools.
- The system relies on the OpenAI GPT-4 API to tackle tricky language tasks such as figuring out sentiment, summarizing content, and classifying intent.The design keeps this layer separate so the LLM acts as a tool rather than just a place to store information.
- To handle Shadow AI risks, n8n keeps all workflow execution logs, prompt inputs, and model outputs saved in its internal database. This gives a detailed record to check for compliance and governance.

We built this system to be modular so it can grow and change as technology or privacy rules shift.The architecture isn't a fixed, all-in-one setup; it's made up of separate layers that can be swapped out easily.This means organizations can switch out individual parts without messing up the rest of the workflow.

For example, the Cognitive Layer isn't tied to just one provider.A company might switch out the OpenAI API for a local model like Ollama that focuses on privacy, to follow tougher data security rules, without changing the user interface or how data is collected at all.This flexibility helps the system stay useful and compliant as infrastructure needs change over time.

Results and Discussion

Deployment and Evaluation

We tested the Intelligent Support Agent in a simulated support setup that handled customer questions, product problems, and feedback through different channels.The evaluation looked at three main things: how much it cut down on manual work, how much smoother operations got overall, and how steady the system latency stayed when it was under load.

Comparative Performance

Based on system implementation and deployment data, performance improvements are summarized in below table:

Table 1 – Performance comparison

Metric	**n8n + Hybrid Pipeline**	**Standard Cloud API / Manual Flow**

Manual Work Reduction	60% (via automated triage and enrichment)	0% (fully manual)
Operational Efficiency	35% increase in end-to-end workflow efficiency	Baseline
Latency Management	Stable, workflow-controlled timing via Wait/Merge nodes	Variable, network-dependent (1,200–2,000+ ms)

User Testing Insights

Based on evaluation scenarios documented in the system design board:

- In Scenario 1, when dealing with a "Broken Product Claim," the system picked up on the negative sentiment just right, pulled out the important product details, and sent the ticket straight to the priority escalation queue.This cut the average analyst review time from about 8 minutes, which involved manually searching through systems, down to roughly 2 minutes by just going over the AI-generated summary—a 75% drop in time-to-resolution.
- In scenario 2, the system was able to identify low-priority informational requests and gave summaries that included helpful context.This let agents write responses right away without needing to do any extra background research.
- In Scenario 3, the system was able to take in image attachments like photos showing product damage and made sure analysts could access them easily through the interface.Right now, extraction works with files, but in the future, local multi-modal models will be added to automatically handle visual evidence.

Future Work

Future work focuses on progressively shifting from cloud-hosted LLM APIs to localized, multi-modal inference while preserving the orchestration pattern:

- We want to switch from using cloud-based APIs to running local multi-modal models like LLaVA or BakLLaVA through Ollama.This lets the agent understand visual info on its own, like pictures of broken items or error screenshots, without sending any private user images to outside servers.
- Right now, workflows mostly just keep a passive record, but in the future, they'll include active Policy-as-Code checks.This means setting up a second "Auditor Agent" inside n8n that watches decisions as they happen and flags any answers that don’t follow the regulatory rules before the analyst sees them.
- Research will focus on making inference run better on edge hardware by using NPUs to handle quantized models directly on the device.The goal here is to cut down on energy use and make sure latency stays predictable, no matter what’s happening with the network.
- We want to look into Federated Learning methods to adjust the routing logic while keeping privacy intact across various support teams.This way, the system can learn from support patterns across the whole company without sharing the raw customer data from any one department.

Conclusion

This work shows that the Intelligent Support Agent a hybrid orchestration pipeline combining low-code automation (n8n) with governed LLM API access can improve enterprise support workflows while reducing Shadow AI risks.The system cut down manual research time by 60% and boosted operational efficiency by 35% thanks to automated triage, gathering info from multiple sources, and centralizing control.The architecture managed to keep response times steady, swapping out unpredictable network delays for processing controlled by the workflow.By keeping complete audit trails and controlling LLM access through one orchestration layer, organizations can enjoy the productivity gains of modern AI while meeting tough compliance and security rules.As more businesses start using AI on a bigger scale, relying on governed, auditable orchestration platforms will be key to keeping a balance between pushing new ideas and staying within the rules.

Acknowledgments

Thanks to Northeastern University advisors for guidance on system design and evaluation methodology.

References

[1]. EPAM Systems. (2026). "Shadow AI: The Enterprise Risk That Can No Longer Be Ignored." *EPAM Newsroom*, Feb. 11, 2026.

[2]. ISACA. (2025). "The Rise of Shadow AI: Auditing Unauthorized AI Tools in the Enterprise." *ISACA Resources*, Sept. 25, 2025.

[3]. WorkOS. (2025). "n8n: The workflow automation tool for the AI age." *WorkOS Blog*, Mar. 23, 2025.

[4]. Towards Data Science. (2026). "The Hidden Opportunity in AI Workflow Automation with n8n for Low-Tech Companies." *Medium*, Jan. 18, 2026.

[5]. SapientPro. (2025). "Automating Call Centers with AI Agents: Achieving 700ms Latency." *Dev.to*, May 20, 2025.

Project Links

- Live demo: https://huggingface.co/spaces/manavkheni/support-agent-demo
- GitHub repository: https://github.com/manavkheni1/smart-customer-router