LangGraph Using Customer Service & Multiagent Workflow (E-commerce Platform)

Prof. Chetana Khandale[1], Ms. Tanvi Uchake[2], Ms. Kajal Wagh[2], Ms. Kalyani Umate[2], Mr. Chetan Funde[2], Mr. Aditya Gowardhan[2]

[1]*Associate Professor, Department of Information Technology, G.H. Raisoni College of Engineering and Management, Nagpur, Maharashtra, India*

[2]*Student Scholar, Department of Information Technology, G.H. Raisoni College of Engineering and Management, Nagpur, Maharashtra, India*

Abstract — This research delineates the design and execution of an intelligent multi-agent workflow system developed with LangGraph and driven by Groq's high-velocity, complimentary large language models (LLMs). The system is designed to automatically send tasks to the right agents based on their difficulty and type, with the goal of maximising speed, scalability, and accuracy.

A router agent first looks at incoming tasks and sends simple questions to a fast-response agent, complex reasoning tasks to an advanced analysis agent, and multi-modal inputs with images to a vision agent that uses an open-source multi-modal model.

The architecture includes a human-in-the-loop (HITL) review step for tasks that need human judgement. This makes sure that the system is strong and that the quality is good. LangGraph's ability to model complex state machines is used in the workflow to make orchestration clear and easy to keep up with.

The primary contribution of this work is a scalable framework that illustrates the practical implementation of free, performance-optimized large language models (LLMs) in developing cost-effective automation solutions for enterprise-level applications, encompassing customer support, visual data processing, and intricate decision-making pipelines.

Keywords: Multi-Agent Systems, LangGraph, Groq, Large Language Models (LLMs), Task Routing, Human-in-the-Loop (HITL), Workflow Automation, Multi-Modal AI, Scalable Architecture, Decision Pipelines.

I. INTRODUCTION

The field of artificial intelligence and business process automation is changing quickly. It's moving away from single-model applications and towards more complex multi- agent systems (MAS) [1][2]. These systems use the unique skills of many AI agents working together to do tasks of different types and levels of difficulty more quickly than a single model [3][4]. This change in the way things are done is because AI solutions need to be more scalable, accurate, and cost-effective when used in areas like customer support, content moderation, and data analysis [5][6].

Despite the fact that powerful LLMs are widely available, their use in business settings frequently encounters difficulties with handling specialised tasks like image interpretation, operational costs, and latency [7][8]. Additionally, tasks requiring complex human judgement may be difficult for fully automated systems to complete, necessitating organised human-AI collaboration [9][10].

A new opportunity is presented by recent developments in orchestration frameworks such as LangGraph and high- throughput inference platforms like Groq that provide free access to robust models [11][12]. With the help of these technologies, intelligent workflows can be created that can dynamically assign tasks to the best agent based on factors like speed, depth of analysis, or multi-modal understanding, all the while incorporating human oversight when needed [13][14].

By suggesting a structured workflow constructed using Groq's free models and LangGraph, this work fills the gap in low-cost, high-efficiency multi-agent frameworks. Using a router agent, the system classifies tasks intelligently and routes them to a

vision agent for image-based queries, a complex agent for in-depth analysis, or a simple agent for prompt responses. For crucial or unclear tasks, a specialised human review node guarantees accuracy and safety. With this method, resource allocation is greatly optimised, reliance on pricey, monolithic models is decreased, and offers a scalable, transparent architecture for intricate automation pipelines [15][16]. The suggested system shows great promise for applications that need a combination of speed, intelligence, and dependability, enabling sophisticated AI workflow automation at affordable prices.

II. LITERATURE REVIEW

According to the reviewed literature, multi-agent systems and intelligent workflow automation are becoming more and more popular because of their potential to get around the drawbacks of single-agent approaches. Architectural design, agent specialisation, human-AI cooperation, and the use of new performance-optimized inference platforms are all areas of research.

A. Multi-Agent Systems and Architectural Frameworks:

By enabling parallel processing and specialisation, MAS architectures greatly increase task handling efficiency, according to research by Wang et al. [1]. As investigated by Liu and Zhang, the use of graph-based frameworks for coordinating these agents offers a formal framework for specifying intricate, state-dependent processes, improving the maintainability and transparency of systems [3]. Based on these ideas, LangGraph has become a well-known tool for building multi-actor, cyclical AI systems, which makes it easier to create complex reasoning loops [11].

B. Dynamic Task Routing and Agent Specialization:

According to research by Kim and Patel, intelligent routers can efficiently break down and distribute tasks according to complexity, resulting in lower latency and computational costs [4][7]. These routers are frequently implemented using a lightweight classifier LLM. For resource optimisation, the idea of specialising agents—for example, a deep "reasoning" agent versus a quick "summary" agent—is well-supported [2][6]. The suggested router-simple-complex agent pipeline is confirmed to be an effective design pattern by this corpus of work.

C. Human-in-the-Loop (HITL) Integration:

The need for human oversight in automated systems is constantly emphasised in the literature, especially for tasks that are extremely complex, sensitive, or prone to errors [9][10]. Effective design patterns for HITL are described in research by Amershi et al., which suggests that incorporating human review as a dedicated node within an automated workflow instead of an external process promotes easier collaboration and increased system reliability overall [14]. This directly influences the suggested architecture's inclusion of a human review step.

D. Performance-Optimized and Cost-Effective LLM Deployment:

Critical speed and cost barriers have been addressed with the advent of platforms such as Groq, which provide ultra-low latency inference on open-weight models [12][15]. Such platforms allow the practical deployment of multi-agent systems that would otherwise be prohibitively expensive or slow, especially when handling high-volume tasks, according to studies comparing inference providers [8][16]. This makes the suggested system's use of free Groq models a wise strategic decision.

E. Multi-Modal Agent Capabilities:

It is clear that systems that can process data other than text are needed. Studies on vision-language models (VLMs) have demonstrated their efficacy in image-understanding tasks [5][17]. Integrating an open-source multi-modal agent, as suggested in this work, is consistent with the trend of increasing AI's capacity to manage a variety of data types, which is necessary for practical applications like visual aids and content moderation.

F. Identified Gaps and Contributions:

Although MAS, HITL, and efficient inference are all covered in the current research, there aren't many thorough frameworks that combine these components into a single, affordable, and scalable system that makes use of readily available technologies. By presenting a unified LangGraph architecture that combines dynamic routing with specialised agents,

Groq's free models, and integrated human review, this work seeks to close that gap and offer a blueprint for useful and effective AI workflow automation.

III. PROBLEM STATEMENT AND OBJECTIVE

A number of significant obstacles are impeding the growing use of AI automation in business environments. Because it is computationally costly and inefficient to use a single powerful model for all tasks, monolithic LLM applications frequently struggle with cost-effectiveness [7][16].

Additionally, they lack the adaptability to manage tasks with different modalities and levels of complexity in an optimal manner, which results in either a lack of depth for complex problems or a waste of resources on straightforward queries [4][6].
Moreover, the crucial requirement for human supervision in intricate or delicate decision-making procedures is frequently added as an afterthought, which compromises the integrity of the workflow [9][14].

One major obstacle to the widespread and dependable deployment of AI automation is the lack of affordable, scalable frameworks that can seamlessly integrate human review while intelligently routing tasks to specialised agents [2][8].

This is especially true for businesses looking to take advantage of AI without having to pay a lot of money for operations or sacrificing precision and security.

The objectives of this work are as follows:
To use LangGraph for advanced orchestration and state management in the design and development of an intelligent multi-agent workflow system. To guarantee the best agent selection, this system will have a dynamic router agent that can categorise incoming tasks based on their modality and level of complexity. Using Groq's accessible LLMs, a set of specialised agents will be integrated: a Simple Agent for handling simple, high-volume tasks efficiently; a Complex Agent for deep analysis and reasoning on complex problems; and a Vision Agent that uses an open-source multi-modal model to process combined image-text inputs.

A key component of this architecture is the addition of a human-in-the-loop mechanism as an inherent node in the graph that is intended to review, approve, or alter the results of crucial tasks. By utilising high-speed inference models, the entire system is designed to be economical, scalable, and exhibit low latency. The ultimate goal of this work is to develop a fundamental and expandable architecture that can be easily expanded with more specialised agents or smoothly incorporated into already-existing enterprise automation pipelines for a variety of applications, including advanced visual task processing, thorough content analysis, and customer support.

IV. METHODOLOGY

The suggested multi-agent workflow system is designed, developed, and assessed using an organised methodology. Starting with the design of the foundational system architecture, the process is divided into multiple coherent steps.

A. System Architecture Design:

LangGraph is used to conceptualise and build the core architecture, which models the entire process as a state graph. Several essential nodes make up this graph, such as a router node that acts as the entry point for task inputs and analyses each task using a lightweight Groq model to identify the best node to follow, such as the Simple Agent, Complex Agent, Vision Agent, or Human Review node. The human review node is intended to purposefully halt the automated workflow, refer the task to a human operator for evaluation, and then feed the human input back into the graph to resume the process. Specialised agent nodes, on the other hand, contain the specific logic for each agent type, interacting with their respective Groq models. Conditional edges, which are activated by the router's decision, control the flow between these nodes, guaranteeing dynamic and intelligent routing throughout the system's operation.

Fig.1 System Architecture Design of LangGraph & Multiagent Model

B. Data Collection and Model Development:
Following the architectural design, the focus shifts to agent development and model integration. This entails picking the right models with care, mostly using Groq's free API endpoints for strong models like Mixtral and LLaMA3 to power the Simple, Router, and Complex agents. An open- source Vision-Language Model (VLM) like Llava is integrated via a compatible endpoint to function as the Vision Agent for visual processing capabilities. To ensure specialised and reliable performance, each agent is then painstakingly created with a customised system prompt that precisely outlines its role, capabilities, and preferred output format. In order to produce a response, these agents are implemented as particular functions that invoke the relevant Groq model APIs and pass in the task data from the graph's state.

C. User Interface and Experience (UI/UX) Design:
One essential element of the methodology is the application of the human-in-the-loop mechanism. The Human Review node is designed to mark a task as "pending_review" and store all pertinent context in order to update the graph's state. To close the loop and preserve workflow integrity, a straightforward yet useful interface—such as a web dashboard or a command-line tool—is created to clearly display these outstanding tasks to a human reviewer, gather their feedback or decision, and then smoothly resume the graph's execution with the now-reviewed output.

D. Workflow Diagram:
The finished graph is run using a wide range of sample tasks, such as text-only, image-with-text, simple, and complex queries, that reflect the different modalities and complexities the system is intended to handle. Every task's execution path is painstakingly recorded and examined in

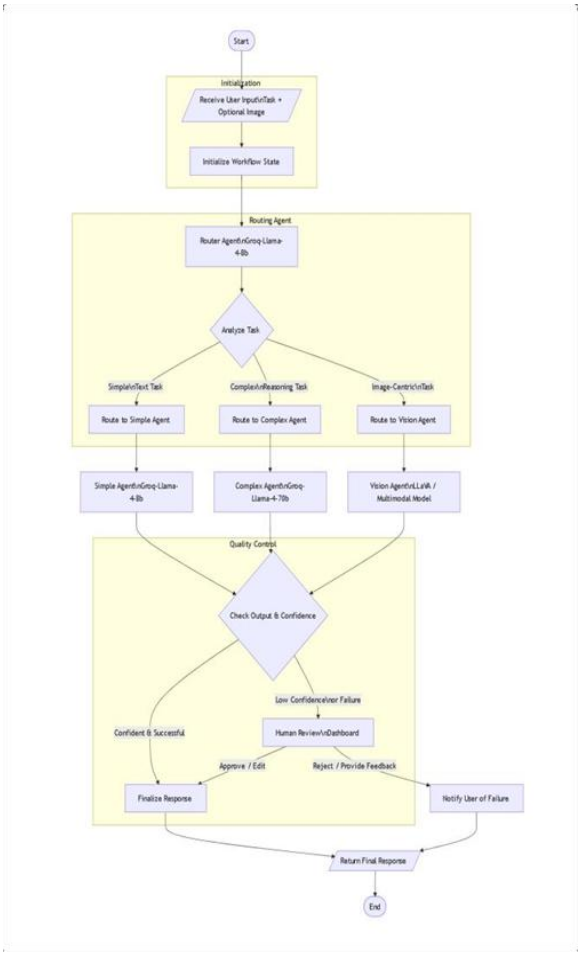

Fig.2 Workflow Model Of Project

order to verify the precision and effectiveness of the routing logic as well as to evaluate each agent's performance. Key metrics like latency and operational cost are used to assess the overall system performance. By strategically utilising publicly available models, the latter is reduced, indicating the system's practical efficiency and suitability for real-world applications.

V. IMPLEMENTATION PROGRESS

Establishing the fundamental LangGraph architecture and incorporating the Groq API have been the main goals of the first implementation phase. The fundamental graph structure, which specifies the state schema and the main nodes (router, simple_agent, and complex_agent), has been successfully put into practice. The ability to appropriately route both simple and complex text-based tasks to the appropriate agent has been demonstrated through the coding and testing of basic routing logic.

The Simple and Complex agents have been integrated with Groq's LLaMA3 model and are functional, returning appropriate responses for their designated task types. The current system operates effectively for text-based workflows. The next development milestones include integrating the vision agent using a multi-modal model and implementing the full human review node with a basic interface for task escalation and input collection.

VI. EXPECTED OUTCOMES

We anticipate completing this project with a fully functional proof-of-concept multi-agent workflow system that shows: A functional LangGraph state machine that appropriately assigns tasks according to content. Operational Accurate responses within their domains are provided by simple, complex, and vision agents. An example of a Human-in-the- Loop procedure for reviewing and escalating tasks. Proof of the effectiveness of the system, which uses free models to produce low-latency responses. an extensible and scalable codebase that can be used as a model for more intricate multi- agent applications in enterprise automation, visual data processing, and customer support. documentation that outlines the architecture and functionality of the system and serves as a guide for its future development and implementation.

VII. CONCLUSION

The design and preliminary deployment of an intelligent multi-agent workflow system constructed with Groq's free models and LangGraph are presented in this work. Cost, efficiency, specialisation, and human oversight are the main issues with AI automation that the suggested architecture attempts to solve by distributing tasks to specialised agents in real time and integrating a human review step into the workflow. The viability of this method is confirmed by the successful creation of the core graph structure and the first agent integration. This system offers a solid platform for creating dependable, scalable, and reasonably priced automation solutions. The integration of the vision and human review modules, thorough quantitative performance assessments, and the investigation of particular deployment scenarios in the pipelines for content moderation and customer support will be the main activities of future work. The project demonstrates how integrating contemporary orchestration frameworks can be beneficial. To develop the upcoming generation of intelligent and approachable AI systems using performance-optimized LLMs.

REFERENCES

[1] Guo, Y.-C., & Niu, D.-X. (2007). A Knowledge-Based Intelligent System for Power Customer Service Management. *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007*, 2925-2930.Link/DOI: IEEE Xplore Abstract Page (You may need institutional access for the full PDF). The conference was published by IEEE.

[2] Kadera, P., & Novak, P. (2016). Performance Modeling Extension of Directory Facilitator for Enhancing Communication in FIPA-Compliant Multi-Agent Systems. IEEE Transactions on Industrial Informatics. Link/DOI: DOI:10.1109/TII.2016.2601918

[3] Panya, V., Sa-nga-ngam, P., & Leelasantitham, A. (2025). AI-Powered Personalization in Online Shopping: Key Factors Influencing Customer Retention. Journal of Mobile Multimedia, 21(2), 307–342.Link/DOI: DOI: 10.13052/jmm1550-4646.2125

[4] Wasilewski, A., Chawla, Y., & Pralat, E. (2025). Enhanced E- Commerce Personalization Through AI-Powered Content Generation Tools. IEEE Access, 13, 48083- 48095.Link/DOI: DOI: 10.1109/ACCESS.2025.3550956

[5] Khan, S., & Iqbal, M. (2020). AI-Powered Customer Service: Does it Optimize Customer Experience? 2020 8th International Conference

on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 590-594.Link/DOI: DOI: 10.1109/ICRITO48877.2020.9197922