International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Self-Aware Safety Middleware

Pathagunta Muniraji, M.Sowmiya, Surathi Teja,Tana Suman
Department of Electronics Communication and Engineering
T.J.S Engineering College (Affiliated to Anna University) Tamil Nadu, India

Abstract- Large Language Models (LLMs) have significantly advanced human–computer interaction by enabling natural lan-guage communication across diverse applications. However, these systems lack runtime control mechanisms, resulting in unsafe outputs, hallucinated responses, and unpredictable behaviour. This paper proposes a Self-Aware AI Middleware, an external control layer that regulates interaction between users and LLMs without modifying the model architecture. The proposed system intercepts user queries, performs intent analysis using Natural Language Processing techniques, applies policy-based decision-making, and validates generated responses before delivery. A dual-layer awareness mechanism is introduced to monitor both input queries and output responses. The architecture consists of modular components including input interception, prompt analysis, policy enforcement, LLM in-teraction, response evaluation, and self-reflection. This structured pipeline ensures safe, reliable, and controlled AI interaction. Experimental observations demonstrate reduced unsafe outputs, improved response consistency, and enhanced user trust. The system is scalable, model-agnostic, and suitable for deployment across multiple domains such as education, healthcare, and intelligent assistants.

Index Terms—Large Language Models, Middleware, AI Safety, Natural Language Processing, Policy Enforcement, Prompt Anal-ysis

I. INTRODUCTION

Large Language Models (LLMs) have revolutionized ar-tificial intelligence by enabling systems to understand and generate human-like language. These models are widely used in chatbots, educational tools, and content generation systems. Despite their capabilities, LLMs lack effective runtime control mechanisms, often producing unsafe, biased, or hallucinated outputs.

In real-world applications, especially in sensitive domains, uncontrolled responses can lead to misinformation and eth-ical concerns. Existing approaches such as fine-tuning and reinforcement learning are model-dependent and resource-intensive.
To address these challenges, this paper proposes a Self-Aware AI Middleware, which introduces an external control layer between the user and the LLM. This approach enables safe and controlled AI interaction without modifying the underlying model.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

II. PROJECT OVERVIEW

The proposed system, Self-Aware AI Middleware, is designed as an intermediate control layer that enhances the behaviour, safety, and reliability of Large Language Model (LLM)-based systems. Unlike conventional architec-tures where user queries are directly processed by the language model, the proposed system introduces a middleware layer that supervises and regulates the entire interaction pipeline.

The primary objective of the system is to enable con-trolled and responsible AI interaction by incorporating self-awareness into the processing workflow. This is achieved by intercepting user inputs, analysing their intent using Natural Language Processing (NLP) techniques, and applying policy-based decision mechanisms before forwarding the query to the language model. Additionally, the system evaluates the responses generated by the LLM to ensure that the output adheres to predefined safety and ethical guidelines.

The architecture of the system consists of multiple func-tional modules, including input interception, prompt analysis, policy enforcement, LLM interaction, response evaluation, and self-reflection. These modules work together to introduce both input-level and output-level awareness, ensuring that unsafe or inappropriate queries are filtered and that generated responses are validated before delivery.

A key feature of the proposed system is its model-agnostic design, which allows seamless integration with various LLM platforms without requiring modifications to their internal architecture. This makes the system highly scalable and adapt-able across different domains such as education, healthcare, customer support, and intelligent assistants.

The middleware follows a structured pipeline where each stage contributes to improving system reliability:

- Input Interception ensures that all queries are captured and monitored
- Prompt Analysis interprets user intent and classifies risk levels
- Policy Enforcement applies safety rules and decision logic
- LLM Interaction handles controlled communication with the model
- Response Evaluation validates outputs for safety and accuracy
- Self-Reflection logs system behaviour for continuous improvement

By introducing a structured control mechanism, the Self-Aware AI Middleware addresses key limitations of existing LLM systems, including lack of safety control, response inconsistency, and susceptibility to misuse. The system thereby contributes to the development of more reliable, trustworthy, and ethically aligned AI applications.

Overall, the proposed approach bridges the gap between powerful generative AI models and the need for controlled, safe, and interpretable AI systems suitable for real-world deployment.

III. EXISTING SYSTEM AND DRAWBACKS

A. Existing System

In recent years, Large Language Models (LLMs) have been widely adopted in various applications such as conversational chatbots, virtual assistants, educational platforms, and content generation systems.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

These systems are designed to process user queries in natural language and generate responses across multiple domains using a single unified model. Due to their open-domain nature, LLM-based systems provide flexibility and ease of access, allowing users to obtain information on diverse topics through a single interaction interface.

Most existing systems follow a direct interaction approach, where user queries are sent directly to the language model without any intermediate control or validation layer. Tech-niques such as prompt engineering, fine-tuning, and rein-forcement learning are commonly used to improve response quality and alignment with user expectations. Additionally, some platforms incorporate basic moderation tools to filter harmful content.

While these systems demonstrate significant advancements in natural language understanding and generation, they pri-marily focus on improving model performance rather than controlling interaction behaviour during runtime.

B. Drawbacks of Existing System

Despite their capabilities, existing LLM-based systems ex-hibit several critical limitations:

- Lack of Input Control: User queries are directly pro-cessed without proper validation, allowing harmful, am-biguous, or malicious inputs to reach the model.
- Unsafe and Unregulated Outputs: LLMs may generate inappropriate, biased, or misleading responses due to the absence of effective output monitoring mechanisms.
- Hallucination and Inaccuracy: Models often produce factually incorrect or fabricated information, reducing reliability in real-world applications.
- Absence of Self-Awareness: Existing systems lack the ability to analyse their own behaviour, assess risk levels, or regulate responses dynamically.
- Model-Dependent Solutions: Current improvements such as fine-tuning and reinforcement learning are tightly coupled with specific models, making them less flexible and difficult to generalize across different architectures.
- Limited Ethical Enforcement: Although moderation tools exist, they are often not integrated as a structured, real-time control layer, resulting in inconsistent enforce-ment of safety policies.

These limitations highlight the need for a robust and flexible mechanism that can regulate AI interactions without modify-ing the underlying model. To address these challenges, the proposed system introduces a Self-Aware AI Middleware Feature, which acts as an external control layer to monitor, analyse, and enforce safe and responsible AI behaviour.

IV. LITERATURE SURVEY

The rapid advancement of Large Language Models (LLMs) and Artificial Intelligence has led to significant developments in intelligent systems, particularly in areas such as conver-sational agents, educational platforms, and decision-support systems. Several research works have explored techniques to improve the accuracy, reliability, and usability of these systems. This section reviews key contributions relevant to the proposed Self-Aware AI Middleware Feature.

One of the most significant advancements in this domain is the introduction of Retrieval-Augmented Generation (RAG) by Lewis et al. (2020), which enhances language models by retrieving relevant external knowledge before generating responses. This approach improves factual accuracy and con-textual relevance, especially in knowledge-intensive applica-tions. Subsequent studies have extended RAG for

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

educational systems, where AI tutors utilise domain-specific knowledge bases to provide personalised learning support.

Wu and Tseng (2025) developed a local generative AI teach-ing assistant using RAG, demonstrating improved response accuracy and domain-specific query handling. Similarly, Vu-Minh et al. (2025) proposed a course-grounded AI tutor that integrates retrieval mechanisms with generative models to de-liver personalised academic assistance. Survey-based studies, such as those by Swacha et al. (2025) and Li et al. (2025), highlight the effectiveness of RAG-based systems in enhancing learning outcomes, reducing hallucinations, and improving contextual grounding in AI-driven educational platforms.

In addition to retrieval-based approaches, research has also focused on Intelligent Tutoring Systems (ITS) and per-sonalised learning environments. Studies by Akheel et al. (2025) and Ne´meth et al. (2025) demonstrate how AI systems can adapt learning pathways based on user behaviour and preferences using Natural Language Processing and machine learning techniques. These systems improve student engage-ment and academic performance but still rely heavily on domain-specific implementations.

Despite these advancements, several challenges remain un-resolved. Existing systems primarily focus on improving re-sponse accuracy and personalisation but do not adequately address real-time control over AI interactions. Issues such as unsafe outputs, lack of input validation, and absence of be-havioural regulation persist across most LLM-based systems. Research on AI safety, including works on guardrails, moderation systems, and prompt filtering, highlights the im-portance of controlling model behaviour. However, these ap-proaches are often integrated at the model level or imple-mented as isolated solutions, limiting their flexibility and scalability.

Recent studies in AI safety and alignment emphasise the need for external control mechanisms that can monitor and regulate both input and output interactions without modifying the internal architecture of the model. Concepts such as policy-based enforcement, input filtering, and output validation have been explored, but a unified and modular approach remains limited.

From the literature survey, it is evident that while significant progress has been made in improving the performance and accuracy of AI systems, there is a lack of a comprehensive framework that ensures safe, controlled, and self-aware inter-action. To bridge this gap, the proposed system introduces a Self-Aware AI Middleware Feature that operates as an external control layer, enabling real-time analysis, regulation, and validation of AI interactions across multiple domains.

V. OBJECTIVE OF THE PROJECT

The primary objective of the proposed system is to develop a Self-Aware AI Middleware Feature that enhances the safety, reliability, and control of Large Language Model (LLM)-based systems by introducing an intermediate supervision layer.

The system aims to regulate AI interactions by intercepting user queries, analysing their intent using Natural Language Processing (NLP) techniques, and applying policy-based de-cision mechanisms before forwarding them to the language model. Additionally, the system evaluates model-generated responses to ensure that the output adheres to predefined safety and ethical guidelines.

The key objectives of the project are as follows:

- Middleware-Based Architecture: To design a middleware-based architecture that operates independently of the underlying LLM without requiring internal model modifications.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Self-Awareness Mechanism: To implement a self-awareness mechanism capable of analysing user intent and classifying queries based on risk levels.
- Policy Enforcement Framework: To develop a policy enforcement framework that ensures controlled and ethi-cal AI interaction.
- Dual-Layer Validation: To introduce a dual-layer val-idation process that monitors both input queries and generated responses.
- Reduction of Unsafe Behaviour: To reduce unsafe outputs, hallucinations, and unpredictable behaviour in LLM-based systems.
- Scalability and Modularity: To provide a scalable and modular solution that can be integrated across multiple AI applications and platforms.
- User Trust Enhancement: To enhance user trust and reliability in AI-driven systems through controlled and transparent interaction mechanisms.

The proposed system ultimately aims to bridge the gap between uncontrolled general-purpose AI models and safe, responsible, and dependable AI systems suitable for real-world deployment.

VI. SCOPE OF THE PROJECT

The system can be deployed in various domains such as:

- Educational platforms
- Healthcare systems
- Customer support applications
- AI-powered assistants

The middleware approach ensures flexibility and adaptabil-ity across different AI models.

VII. METHODOLOGY

The proposed Self-Aware AI Middleware Feature is designed as a structured intermediate layer that regulates the interaction between the user and the Large Language Model (LLM). The methodology focuses on introducing real-time control, safety enforcement, and behavioural awareness without modifying the internal architecture of the model.

The system follows a modular pipeline approach where each stage is responsible for analysing, validating, and regulating the interaction process. The overall workflow ensures that both user inputs and model-generated outputs are continuously monitored to maintain safe and reliable AI behaviour.

A. System Architecture Overview

The architecture consists of multiple layers...
As shown in Fig. 1, the system illustrates interaction be-tween instructor, learner, and LLM.

B. Input Interception and Pre-processing

The first stage of the system involves capturing user queries through the middleware. Unlike traditional systems, the query is not directly forwarded to the LLM. Instead, it is intercepted and pre-processed to ensure that all inputs are monitored before execution.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

This stage includes:

- Query extraction from the client interface
- Text normalization and preprocessing
- Forwarding the processed input to the analysis module

This step ensures that the system has complete control over incoming data.

C. Prompt Analysis and Intent Classification

In this stage, the system analyses the intercepted query using Natural Language Processing (NLP) techniques. The objective is to understand the semantic meaning and classify the intent of the user. The analysis process includes:

- Tokenisation and keyword extraction
- Context and intent identification
- Risk classification into safe, moderate, or high-risk cate-gories

Based on the classification:

- Safe queries are allowed for further processing
- High-risk queries are blocked or modified

This stage introduces the first level of self-awareness by enabling the system to interpret user intent.

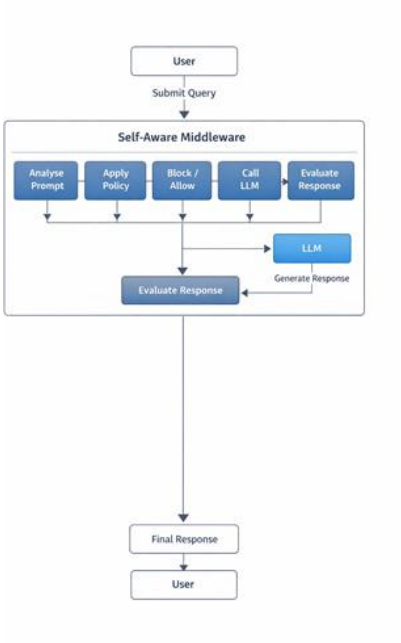

Fig. 1. Use Case Diagram Representing Interaction Between Instructor, Learner, and LLM

D. Policy Enforcement Mechanism

The Policy Enforcement Module applies predefined rules and ethical constraints to regulate system behaviour. This module acts as a decision-making engine that determines how each query should be handled.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The system performs:

- Evaluation of queries against safety policies
- Decision making (allow / reject / modify)
- Enforcement of ethical and domain-specific constraints This ensures that only valid and safe queries are forwarded

to the LLM, thereby preventing misuse and unsafe interactions.

E. LLM Interaction Module

After validation, the processed query is forwarded to the Large Language Model for response generation. The middle-ware ensures that only filtered and approved inputs are sent to the model.
This module:

- Sends validated queries to the LLM
- Receives generated responses
- Maintains controlled communication between middle-ware and model

The LLM operates as a response generator, while the middleware remains responsible for overall control.

F. Response Evaluation and Post-processing

Once the response is generated, it is passed through a validation layer to ensure safety and reliability. This stage introduces output-level self-awareness.

The system performs:

- Content analysis of generated responses
- Detection of unsafe, biased, or misleading information
- Filtering or modification of responses if necessary

This ensures that the final output adheres to safety guide-lines and improves the trustworthiness of the system.

G. Output Delivery

After validation, the final response is delivered to the user in a structured format. The middleware ensures that the output is clean, safe, and contextually relevant.

This stage provides:

- Safe and verified responses
- User-friendly output formatting
- Consistent interaction experience

H. Self-Reflection and Logging

The system incorporates a self-reflection mechanism that logs all interactions for monitoring and improvement. This module records user queries, system decisions, and generated responses.

The logging mechanism enables:

- Behaviour tracking
- Performance analysis
- Continuous system improvement

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

I. Overall Workflow Summary

The complete methodology follows a sequential pipeline: User Query → Input Interception → Prompt Analysis → Policy Enforcement → LLM Processing → Response Evaluation → Output Delivery → Logging

The proposed methodology introduces a dual-layer self-awareness mechanism by analysing both input queries and generated responses. This approach ensures controlled, safe, and reliable AI interaction, making the system suitable for real-world applications where ethical and responsible AI usage is essential.

J. System Workflow

The workflow of the proposed Self-Aware AI Middleware is illustrated in Fig. 2. The process begins with user input...

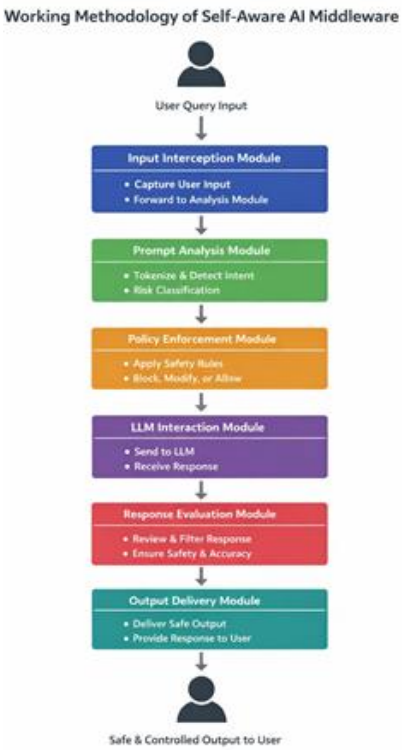

Fig. 2. Workflow of Self-Aware AI Middleware Showing Input Processing, Policy Enforcement, and Response Validation

VIII. MODULE DEVELOPMENT

The proposed Self-Aware AI Middleware Feature is developed using a modular architecture, where each module is responsible for a specific function within the interaction pipeline. This modular design improves system maintainabil-ity, scalability, and flexibility, allowing independent develop-ment and enhancement of components without affecting the overall system.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

A. Core Middleware Module

The Core Middleware Module acts as the central controller of the system. It manages the overall workflow and coordinates communication between all other modules.

Functions:

- Receives incoming user requests from the client
- Routes queries through different processing stages
- Maintains the execution flow of the system
- Integrates all modules into a unified pipeline

This module ensures that the system operates in a structured and controlled manner.

B. Input Interception Module

This module serves as the entry point of the system and captures all user queries before they reach the language model.

Functions:

- Intercepts user input from API requests
- Prevents direct access to the LLM
- Forwards input to the analysis module

This ensures that all inputs are monitored and controlled from the initial stage.

C. Prompt Analysis Module

The Prompt Analysis Module is responsible for under-standing and interpreting user queries using Natural Language Processing (NLP) techniques.

Functions:

- Performs tokenisation and keyword extraction
- Identifies user intent and context
- Classifies queries into safe, moderate, or high-risk cate-gories

This module introduces the first level of self-awareness by enabling the system to analyse the meaning and intent of the query.

D. Policy Enforcement Module

This module acts as the decision-making unit of the system by applying predefined rules and safety policies.

Functions:

- Evaluates queries against policy rules
- Decides whether to allow, block, or modify input
- Ensures compliance with ethical guidelines

This module ensures that only valid and safe queries are processed further.

E. LLM Integration Module

... your content ...

The user interface shown in Fig. 3 represents the im-plementation of the Self-Aware AI Middleware system. It demonstrates how user queries are processed through the mid-dleware layer, where unsafe

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

or restricted inputs are identified and blocked, while valid queries are processed to generate meaningful responses.

F. Response Evaluation Module

This module is responsible for validating the responses generated by the LLM.
Functions:

- Analyses output content for safety and correctness
- Detects harmful, biased, or misleading responses
- Filters or modifies responses when necessary

This module introduces output-level self-awareness, ensur-ing reliable and safe responses.

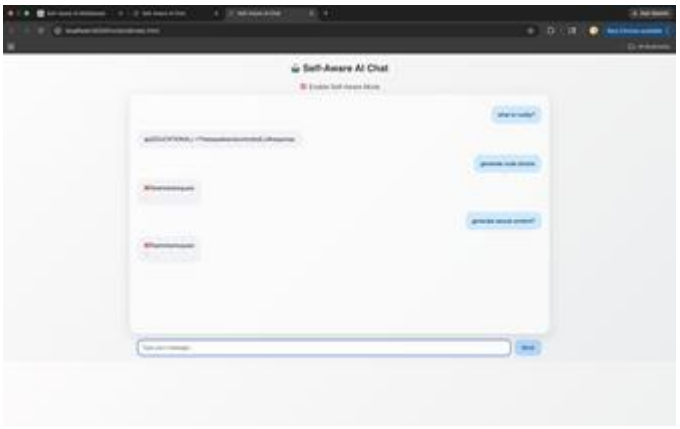

Fig. 3. User Interface of Self-Aware AI Middleware Demonstrating Safe and Controlled Interaction

G. Error and Refusal Handling Module

This module manages system responses in cases of unsafe queries or errors.
Functions:

- Generates polite and informative refusal messages
- Ensures user-friendly communication
- Prevents exposure of system-level errors

Example:
"This request cannot be processed due to safety constraints."

H. Self-Reflection Module

The Self-Reflection Module enables monitoring and contin-uous improvement of the system.
Functions:

- Logs user queries, system decisions, and responses
- Tracks system behaviour and performance
- Supports debugging and future enhancements

This module strengthens the self-aware nature of the system by allowing it to analyse its own behaviour. The modular development approach ensures that the system is flexible, scalable, and easy to integrate with different Large Language Models. Each module contributes to building a controlled, safe, and reliable AI interaction framework, making the system suitable for real-world deployment.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

IX. RESULTS AND DISCUSSION

The proposed system demonstrates significant improve-ments in AI safety and reliability. The middleware effectively filters unsafe queries and ensures controlled response genera-tion.

Key observations include:

- Reduction in unsafe outputs
- Improved response consistency
- Enhanced user trust
- Better control over AI behaviour

Acknowledgment
The authors sincerely thank B.M. Yuvamaliga for their guidance, encouragement, and technical support throughout this project. Their valuable insights and problem-solving ap-proach significantly contributed to the successful development and refinement of the proposed system.

X. CONCLUSION

The proposed Self-Aware AI Middleware successfully addresses the critical challenges associated with the safety, reliability, and ethical usage of Large Language Models (LLMs). Unlike conventional AI systems that directly process user queries, the developed system introduces an intelligent middleware layer that enables real-time monitoring, intent analysis, and policy enforcement before interacting with the language model.

By incorporating a structured pipeline consisting of intent classification, policy control, and response evaluation, the system effectively prevents harmful and inappropriate content generation while maintaining the quality of legitimate and educational responses. The ability to distinguish between harmful intent and informational queries ensures that the system remains both safe and informative.
The implementation of a Self-Aware Mode Toggle further enhances the system by providing user-level control and transparency, allowing users to observe the difference between controlled and uncontrolled AI behaviour. Additionally, the integration of a confidence evaluation mechanism improves trust in the generated responses by assessing their reliability before delivery.

The experimental results demonstrate that the system:

- Successfully blocks harmful and unsafe queries
- Prevents jailbreak attempts
- Maintains meaningful and context-aware responses
- Enhances user trust and interaction experience

Overall, the Self-Aware AI Middleware transforms tradi-tional LLM-based systems into controlled, responsible, and user-centric AI systems. The modular architecture and feature-based design make it highly scalable and adaptable for integra-tion into various real-world applications such as educational platforms, chatbots, and enterprise AI systems.

The project highlights the importance of embedding self-awareness and control mechanisms in AI systems, paving the way for the development of safe, reliable, and ethically aligned artificial intelligence solutions.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

REFERENCES

1. Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," 2022.
2. Anthropic, "Guardrails for Large Language Models," 2023.
3. E. Perez and I. Ribeiro, "Prompt Injection Attacks," 2022.
4. OpenAI, "Moderation Models," 2023.
5. P. Lewis et al., "RAG," 2020.
6. Y. Ji et al., "Hallucination Survey," 2023.
7. T. Schick et al., "Toolformer," 2023.
8. K. Korbak et al., "AI Control Layers," 2023.
9. I. Gabriel, "AI Alignment," 2020.
10. E. Dinan et al., "Moderation Systems," 2022.