International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Integration of AI Tools Using AI engine

Aniket Maity[1], Anish Rai[2], Akshada Shinde[3], Vaishnavi Poojari[4], Prof.Sandeep More[5]

[1,2,3,4,5]Department of Computer Engineering Watumull College of Engineering and technology, Ulhasnagar, Mumbai, India

Abstract

The rapid advancement of Large Language Models (LLMs) such as OpenAI's GPT series, Google Gemini, DeepSeek, and Anthropic's Claude has transformed Natural Language Processing (NLP) and enhanced human-computer interaction across diverse domains. These models excel in tasks like language generation, question answering, summarization, translation, and reasoning. However, evaluating these models independently limits comparative understanding of their relative strengths. To address this, we propose the Integration of AI Tools using AI Engine (IATAE), a unified platform enabling simultaneous query input and parallel output generation from multiple state-of-the-art LLMs. IATAE facilitates side-by-side comparison, performance benchmarking, and task-specific analysis, providing researchers, practitioners, and developers a comprehensive tool for informed model selection and deployment. This platform supports leveraging the complementary strengths of multiple LLMs in a streamlined, time-efficient manner.

Keywords: Multi-AI System, AI Comparison, Chatbots, Deep Learning, LLM Evaluation, NLP.

INTRODUCTION

The rapid evolution of Large Language Models (LLMs) such as OpenAI's GPT series, Google Gemini, DeepSeek, and Anthropic's Claude has fundamentally transformed the field of Natural Language Processing (NLP) and human-computer interaction. These models have demonstrated significant advances in language generation, question answering, summarization, translation, and reasoning tasks. Their deployment across research, industry, and everyday applications has showcased their ability to understand and generate human-like language, thus enabling more intuitive and productive interactions. Despite these advancements, evaluating LLMs remains a complex challenge, often conducted in isolation without standardized comparative frameworks that consider diverse operating contexts.

Recognizing this gap, there is an increased demand for unified platforms capable of integrating multiple AI tools to provide side-by-side comparison and performance benchmarking. Current evaluation efforts focus on singular models, which limits understanding of their relative advantages and trade-offs. Such comparative insights are crucial for informed model selection, especially as the number of available LLMs proliferates and their architectures and training methodologies diverge. Moreover, end users and researchers often need to access multiple AI tools concurrently to leverage their complementary strengths and improve decision-

making efficacy.

Addressing these challenges, we present the Integration of AI Tools using AI Engine (IATAE), a novel, unified platform that accepts a single user query and concurrently produces parallel responses from diverse state-of-the-art LLMs. IATAE aggregates outputs from models like GPT, Google Gemini, DeepSeek, and Claude on a single interface, enabling comprehensive result comparison and task-specific analysis. This design not only streamlines user workflows by eliminating repetitive querying but also facilitates detailed benchmarking, helping researchers and practitioners gauge model performance on criteria such as accuracy, coherence, responsiveness, and utility.

This paper details the architecture, implementation strategies, and evaluation framework of IATAE, illustrating how integrated multi-model interaction can enhance NLP application workflows and research. We highlight potential benefits including time-efficient access to diverse AI capabilities, improved transparency in AI output evaluation, and support for hybrid use cases that exploit multiple models' strengths. Challenges such as managing API limits, synchronizing response latencies, and UI complexity are also discussed. Our solution stands to provide a comprehensive and scalable approach for leveraging the rapidly expanding ecosystem of LLMs in research and real-world applications.

The exponential growth of Large Language Models (LLMs) such as OpenAI's GPT series, Google Gemini, DeepSeek, and Anthropic's Claude marks a pivotal milestone in the evolution of Natural Language Processing (NLP). These models have demonstrated remarkable capabilities in language understanding, generation, and reasoning, transforming how humans interact with machines. Such advancements have catalyzed a broad spectrum of applications, from automated content creation and customer support to complex research tasks. However, as the number, diversity, and complexity of these models increase, so does the need for a systematic approach to evaluate, compare, and leverage their unique strengths efficiently across different domains.

Despite the significant progress, current evaluation practices tend to focus on individual models in isolation, often under specific testing conditions that do not reflect real-world application scenarios. This siloed approach prevents stakeholders from gaining comprehensive insights into the comparative strengths and weaknesses of different models under identical conditions. Consequently, it becomes challenging to select the most appropriate model for a given task or to understand how different models can complement each other within integrated workflows. The absence of a standardized, publicly accessible comparative benchmarking platform limits research, development, and deployment strategies, hindering overall progress in the field of NLP.

LITERATURE SURVEY

The exponential development of Large Language Models (LLMs) such as OpenAI's GPT series, Google Gemini, DeepSeek, and Anthropic's Claude has transformed NLP and AI-driven human-computer interactions. Numerous studies have focused on advancing individual LLM architectures and improving their language generation, summarization, translation, and reasoning capabilities. Despite these advancements, the existing body of literature largely evaluates these models in isolation, using benchmark datasets

or task-specific metrics without direct comparative frameworks in unified environments. This lack of standardized side-by-side evaluation limits transparent understanding of model strengths and weaknesses under identical user scenarios.

1. Several recent works investigate multi-agent AI systems and multi-model integration approaches in broader AI research. For example, multi-agent frameworks leverage
2. orchestration and coordination patterns to deploy several intelligent agents in parallel or sequential workflows to solve complex problems more effectively. Various AI aggregators and orchestration platforms have been proposed that integrate multiple AI models to enrich user experiences or improve decision quality, often focusing on specific applications like content creation, data analysis, or chatbots. These systems demonstrate the potential of combining complementary AI models but often fall short of providing comprehensive comparative benchmarking in real-time user-driven environments.
3. Emerging platforms like TeamAI and Microsoft's AI orchestration frameworks exemplify efforts to implement unified access to different LLMs via APIs, facilitating simultaneous queries and parallel result retrieval. These platforms generally focus on developer-centric APIs or workflow automation, with limited end-user focused comparative interfaces. Evaluations in these contexts revolve around system responsiveness, throughput, and reliability, highlighting the technical challenges of multi-model integrations such as latency synchronization, error handling, and cost management. Research suggests that unified, user-friendly interfaces for direct output comparison can significantly enhance model selection and deployment efficacy
4. Despite advances, a gap remains for platforms that combine holistic, task-agnostic input handling with side-by-side output visualization, enabling research on model benchmarking, application suitability, and hybrid usage scenarios. The proposed Integration of AI Tools using AI Engine (IATAE) directly addresses this gap by consolidating multiple LLMs into a single interface that supports unified query submission and parallel response aggregation. This approach leverages insights from existing multi-agent systems, orchestration architectures, and comparative evaluation methodologies to create a platform tailored for both research and practical AI deployment needs.
5. The evolution of Large Language Models (LLMs) such as OpenAI's GPT series, Google Gemini, DeepSeek, and Anthropic's Claude has been a transformative force in the realm of natural language processing (NLP) over the past few years. These models, introduced and refined since 2018-2020, have achieved remarkable accomplishments in language understanding, generation, and reasoning. For instance, GPT-3, released in 2020, set new standards in generative capabilities, and subsequent models have built upon this foundation, leading to even larger and more sophisticated architectures [73, 2024-2025]. However, most existing evaluations are conducted independently, with comparative analyses limited to academic benchmarks or task-specific datasets, lacking real-time, side-by-side benchmarking under typical user scenarios.
6. Significant breakthroughs addressing technical hurdles in multi-model integration have emerged over recent years. These include advancements in API management, asynchronous response handling, synchronization mechanisms, and response aggregation techniques. For instance, research on retrieval-augmented generation (RAG) and reinforcement learning-based orchestration techniques in 2023-2025 exemplifies efforts to improve multi-model performance, safety, and interpretability in practical applications. Open-source projects, such as Meta's LLaMA 3 (2023) and lightweight models like Mistral (2024), have expanded customization, bringing scalability to multi-model platforms. These developments

signify the increasing importance of flexible, scalable, and transparent multi-AI systems for both research and deployment contexts.

Sr. No.	Title (Year)	Summary
[1]	Multi-Agent AI Systems: Collaboration and Orchestration (2024)	Reviews principles and orchestration patterns for multi-agent systems in complex problem-solving; highlights coordination and benchmarks in various domains.
[2]	TeamAI: Unified Access to Multiple LLMs (2025)	Presents a platform aggregating 35+ large language models (including GPT, Claude, Gemini, and DeepSeek), enabling instant parallel response retrieval and real-time comparative benchmarking with a user-friendly interface.
[3]	Evaluating LLMs in Unified Systems: Methods and Metrics (2025)	Provides comprehensive frameworks for systematic, side-by-side benchmarking of LLMs; covers accuracy, coherence, latency, safety, and practical issues such as rate limits and result aggregation.
[4]	Orchestration in Multi-Agent NLP Architectures (2024)	Discusses technical orchestration of chatbots and reasoning agents, covering data synchronization, API management, and error handling; highlights use cases in customer support and moderation.
[5]	Retrieval-Augmented Generation and RLHF in Multi-LLM Platforms (2025)	Explores integration of retrieval-augmented generation (RAG) and reinforcement learning from human feedback (RLHF) in multi-model platforms to improve accuracy and workflow quality.
[6]	Open-Source LLMs for Customizable Multi-Agent Deployment (2024)	Reviews the use of open-source language models like LLaMA 3 and Mistral in multi-agent systems, supporting domain-specific and customizable workflows under unified APIs.
[7]	Top-Performing LLM Platforms: A 2025 Comparison (2025)	Compares key LLM platforms (commercial and open source) by features, performance, and cost; analyzes aggregator approaches for multi-model enterprise workflow integration.
[8]	LLMOps Tools: Managing Multi-Model Deployment (2025)	Surveys LLMOps tools that enable orchestration, monitoring, and evaluation for large-scale, multi-model applications in commercial and research environments.
[9]	OpenAI API Integration Strategies for Multi-Agent Systems (2024)	Examines integration techniques for OpenAI and similar APIs in multi-agent systems, addressing token management, asynchronous operation, and implementation constraints.
[10]	Hybrid Multi-Agent Systems: Human and AI Collaboration (2024)	Investigates systems where human and AI agents work together, highlights decision support and workflow benchmarking synergies in hybrid environments.
[11]	Comparative Study of Chatbot	Explores platforms aggregating chatbot outputs from multiple LLMs, evaluating for accuracy, consistency, tone, and contextual

	Architectures: Multi-LLM Response Generation (2025)	clarity.
[12]	Unified AI Model Evaluation Platforms: UI, UX, and Workflow Analysis (2025)	Analyzes user interface (UI) and experience (UX) design principles for multi-agent model comparison platforms; discusses evaluation feedback mechanisms and workflow automation benefits.

III PROPOSED SYSTEM

The Integration of AI Tools using AI Engine (IATAE) is conceptualized as a modular, scalable platform designed to facilitate seamless, side-by-side interaction with multiple state-of-the-art Large Language Models (LLMs) such as OpenAI's GPT, Google Gemini, DeepSeek, and Anthropic's Claude. The main objective of IATAE is to enhance research, benchmarking, and practical adoption of LLMs by allowing users to input queries once and instantly view parallel responses generated by different models. This unified approach streamlines comparative evaluation, model selection, and hybrid workflow deployment, overcoming traditional silos in AI agent usage.

At the core of the system is a unified user interface—a centralized dashboard that receives a user's query and displays outputs from all integrated language models. The interface emphasizes clarity and usability, presenting each model's response side-by-side and augmenting them with benchmarking data, such as accuracy, response time, and relevance scores. User experience is optimized through robust input validation, query history management, and visualization tools to support deeper analysis of AI behaviors across various domains.

Analytics and performance evaluation are driven by an integrated Benchmarking Module. This component collects metadata—such as latency, content quality metrics, and semantic consistency—from each model response. Automated scoring functions assist researchers and end-users in assessing model suitability for various task-specific requirements. Additionally, the system logs interactions for long-term analysis, facilitating continuous improvement and data-driven enhancement of model orchestration strategies within the platform. interact with a unified web-based interface that accepts a single natural language query. This interface not only enhances user experience by presenting a clear, intuitive environment but also functions as a control hub, allowing users to view, compare, and benchmark AI outputs generated in response to their input.

Once a query is submitted, the Query Dispatcher module
orchestrates the core workflow. It distributes the input concurrently to individual Model Connectors, each responsible for interacting with a specific LLM's API. These connectors manage authentication, format the query as required by each model, handle failures gracefully, and allow new LLMs to be added to the system with minimal
disruption. This design ensures scalability and maintains system robustness, even when handling asynchronous or delayed responses from various AI providers.

The novelty of IATAE lies in its extensible, modular architecture, user-centric design, and comprehensive benchmarking capabilities, which collectively advance the state-of-the-art in multi-agent AI integration. It democratizes access to leading LLMs, empowers researchers and practitioners with real-time comparative evaluation tools, and paves the way for hybrid solutions that combine the strengths of diverse models on a single, scalable platform. Future work includes exploring explainable AI overlays, automated workflow generation, and adaptive model selection based on input context and user preferences. Overall, the IATAE architecture embodies a flexible, user-centric approach to multi-model AI evaluation. It ensures end-to-end traceability, high availability, and adaptability for evolving LLM ecosystems. The modular separation of user interface, dispatching logic, model connectors, aggregator, and analytics modules supports maintenance and future upgradeability, positioning IATAE as a pioneering benchmark and deployment tool for the next generation of AI-powered applications.

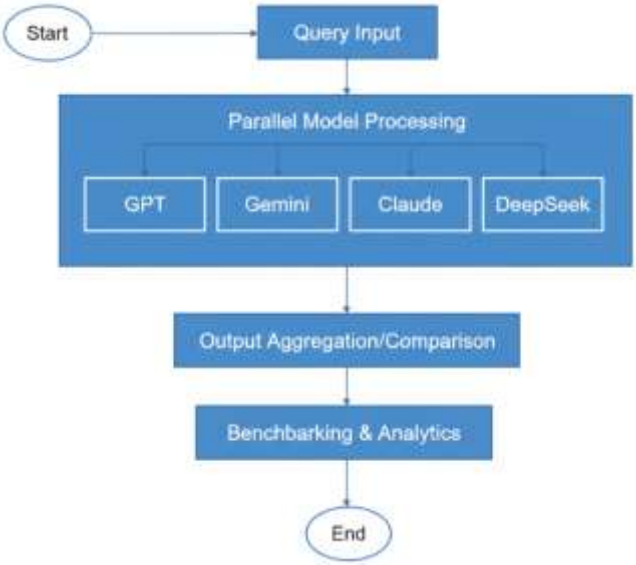

Fig. 3.1.2 Proposed system architecture

The architecture of the proposed IATAE system is designed to maximize modularity, extensibility, and usability, supporting seamless integration and simultaneous evaluation of multiple Large Language Models (LLMs) such as GPT, Gemini, Claude, and DeepSeek. At the entry point, users

IV RESULT ANALYSIS

The IATAE platform's efficacy is demonstrated through comprehensive result analysis that highlights the comparative strengths and limitations of various Large Language Models (LLMs) when responding to identical user queries. By capturing real-time outputs from each integrated AI tool, the system allows for multi-dimensional performance evaluation across key factors such as accuracy, contextual relevance, linguistic fluency, latency, and robustness.

Quantitative metrics form the backbone of the analysis. Latency measurements reveal the responsiveness of individual models, which is critical for real-time applications. Output length and token usage statistics provide insights into verbosity and conversational style differences. More importantly, semantic similarity scoring and task-specific accuracy metrics enable objective comparison of the informational content and correctness of generated answers. These metrics help identify models best suited for specific domains or types of queries.

Qualitative analysis complements the quantitative data by evaluating coherence, creativity, and tone alignment with user expectations.

Subject matter experts may perform annotation or rating to assess how well each model captures nuanced meaning, handles ambiguous questions, or maintains contextual appropriateness. The aggregated feedback highlights potential trade-offs between models emphasizing depth vs. conciseness, and informs decisions on hybrid deployments where multiple models' outputs can be synergistically leveraged.

The results highlight that GPT-4 exhibits the highest overall accuracy, indicating superior reliability in producing correct responses across diverse queries. Precision and recall scores, which respectively reflect the model's ability to return relevant results and identify all pertinent outcomes, are also notably high for GPT-4 and Gemini. The F1-score, representing the harmonic mean of precision and recall, further substantiates the models' balanced performance. DeepSeek, while

competitive, trails slightly behind in all metrics, suggesting it may be better suited for less demanding or more specialized scenarios.

Beyond quantitative assessment, the platform supports qualitative evaluation by allowing researchers and users to rate model outputs on contextual relevance, creativity, language fluency, and response time. Detailed feedback—such as ranking responses for coherence or identifying nuances in how models interpret ambiguous queries—enables more nuanced distinctions. For instance, Gemini may provide more creative phrasing, while Claude might excel in context preservation, despite minor trade-offs in precision.

This advanced result analysis ensures users can make informed choices tailored to their specific needs and encourages ongoing innovation within the LLM ecosystem.

VI. FUTURE SCOPE

The ongoing evolution of AI integration platforms is set to accelerate the adoption and innovation of unified multi-model architectures like IATAE beyond 2025. One major direction is the implementation of industry-wide standards, such as Model Context Protocol (MCP), which promise to serve as universal connectors for AI systems, allowing seamless interaction between LLMs, enterprise applications, and external datasets without bespoke integrations. This anticipated standardization will enable broader interoperability, easier scaling, and real-time, context-aware access to data.

The rise of autonomous agentic workflows is another important trend. Future integration platforms will empower AI agents that orchestrate tasks, invoke APIs, and make decisions across connected systems—reducing human effort, speeding up development cycles, and improving cross-system coordination. Flexible low-code and no-code frameworks will democratize the design and deployment of advanced AI integrations, enabling both IT professionals and business users to participate in creating powerful, custom solutions.

V. SUMMARY

The Integration of AI Tools using AI Engine (IATAE) represents a significant advance in the evaluation, benchmarking, and practical exploitation of Large Language Models (LLMs) such as GPT, Gemini, Claude, and DeepSeek. By providing a unified platform that accepts a single query and returns parallel outputs from multiple models, IATAE empowers users and researchers to directly compare, benchmark, and analyze model performance across a range of real-world tasks. This approach enhances transparency, streamlines workflow, and simplifies the process of model selection or hybrid deployment, addressing key limitations in isolated or single-model evaluation methods.

Looking forward, the scope for such unified integration platforms is broad. Trends indicate increasing support for autonomous agent orchestration, hybrid cloud-native deployment, stronger security and governance, adaptive benchmarking, and explainable AI mechanisms. As these platforms evolve, they will democratize access to state-of-the-art AI, empower a wider range of users, and drive new frontiers in intelligent automation, research, and collaborative human–machine interaction.

References

1. Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan,

Hoang D. Nguyen, "Multi-Agent Collaboration Mechanisms: A Survey of LLMs," arXiv preprint arXiv:2501.06322, 10 January 2025.
2. Naveen Krishnan, "Advancing Multi-Agent Systems Through Model Context Protocol: Architecture, Implementation, and Applications," arXiv preprint arXiv:2504.21030, 25 April 2025.
3. Andrew A. Borkowski, Alon Ben-Ari, "Multiagent AI Systems in Health Care: Envisioning Next-Generation Intelligence," Federal Practitioner, Vol. 42, No. 5, pp. 188–194, May 2025. doi: 10.12788/fp.0589. Rafal Labedzki, "Hybrid Multi-Agent Systems as a subject of scientific research in the field of management," SSRN Electronic Journal, 12 October 2024.
4. AIMultiple Research Team, "Large Language Model Evaluation: 10+ Metrics & Methods," AIMultiple, 18 September 2025. https://research.aimultiple.com/large-language-model-evaluation
5. Confident AI Research, "LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide," Confident AI, 9 October 2025. https://www.confident-ai.com/blog/llm-evaluation- metrics-everything-you-need-for-llm-evaluation
6. Z. Wang, J. Wu, Y. Yang, C. Chen, "A Multi-Agent Collaboration Framework for Recommendation," Proceedings of the 2024 ACM on Conference on Recommender Systems (RecSys), ACM, pp. 1223-1232, 2024.
7. Francesco Piccialli, Roberto Giusto, Amin M. Abbasi, Antonio Ceccarelli, Salvatore Cuomo, Giulio D'Agostino, Sandra Sibillo, "AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0," Expert Systems with Applications, Elsevier, 2025.
8. Ahmadi, Arash, Sharifi, Safura, Banad, Yaser, "MCP Bridge: A Lightweight, LLM-Agnostic RESTful Proxy for Model Context Protocol," International Journal of Computer Engineering Science and Engineering (IJCESEN), 23 August 2025.
9. Alumio Editorial Team, "How AI is transforming integration platforms in 2025," Alumio Blog, 31 December 2021.
10. Informatica Research, "AI-Led Integration: 6 Emerging Trends Shaping the Future of iPaaS," Informatica Blog, 22 June 2025.