iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83567

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Blackhole AI: Intelligent Query Routing and Cost-Optimized RAG System

Mr. Sahil A. Katkar[1], Mr. Harshal N. Barge[2], Mrs. Kajal P. Khalate[3], Mr. Prathmesh V. Ingale[4], Mr. Vedant D. Yeske[5]
Information Technology Dept. VPKBIET Baramati, SPPU Pune, Maharashtra, India

Abstract: ***The rapid adoption of large language models (LLMs) in enterprise and knowledge-intensive applications has introduced significant challenges related to inference cost, latency, and scalability. Most existing deployments rely on uniform cloud-based processing, which leads to unnecessary resource consumption for simple queries. This paper presents Blackhole AI, an adaptive query routing framework that integrates seman-tic embedding, retrieval-augmented generation, and quantitative complexity estimation to dynamically select between local and cloud-based models.***
The proposed system introduces a cost-aware routing mech-anism that evaluates query difficulty before model invocation, enabling efficient allocation of computational resources. By com-bining vector-based retrieval with adaptive decision thresholds, Blackhole AI aims to balance response accuracy with operational efficiency. The framework highlights the importance of intelligent routing strategies in achieving scalable and economically sustain-able LLM deployment.
Index Terms: Retrieval-Augmented Generation, Large Lan-guage Models, Query Routing, FAISS, Sentence-BERT, Cost Optimization, Hybrid AI Systems, Context-Aware AI.

I. INTRODUCTION

A. Background

Large Language Models (LLMs) have rapidly become cen-tral to modern intelligent systems, enabling applications such as conversational agents, document summarization, question answering, and code generation. These models rely on large-scale transformer architectures trained on massive datasets, allowing them to perform complex reasoning and language understanding tasks. However, deploying large cloud-based LLMs for every incoming query introduces significant compu-tational overhead, increased latency, and recurring operational costs.

To improve the reliability of facts, Retrieval-Augmented Generation (RAG) frameworks were introduced. RAG systems improve generative responses by retrieving relevant external knowledge before producing an answer. In a typical pipeline, a query is converted into a semantic embedding, matched against a vector database using similarity search, and the retrieved context is supplied to a generative model. This combination improves grounding and reduces hallucination compared to inference with LLM alone.

Despite these advancements, most existing RAG systems treat all queries equally, regardless of their complexity. In real-world scenarios, however, user queries vary significantly in semantic difficulty and contextual demand. Simple factual queries may not require large cloud-based reasoning, while analytical or multi-hop questions may benefit from more powerful models. Processing every query using the same computational pathway leads to inefficient resource utilization. Inspired by adaptive decision-making principles in intel-ligent systems, Blackhole AI introduces a hybrid routing framework that dynamically selects between local and cloud-based language models. Let Q denote a user query and A denote the generated response. The system aims to learn a mapping:

$$f^* : Q \rightarrow A \qquad (1)$$

Rather than relying on a single model, the system evaluates query characteristics before model invocation. The routing decision can be expressed as:

$$A = \begin{cases} M_l(R(E(Q))) & \text{if } C(Q) < T \\ M_c(R(E(Q))) & \text{otherwise} \end{cases} \qquad (2)$$

where $E(Q)$ represents the semantic embedding of the query, $R(\cdot)$ denotes the retrieval function, M_l and M_c rep-resent local and cloud-based models respectively, $C(Q)$ is a computed complexity score, and T is a routing threshold.

By integrating semantic retrieval, adaptive complexity es-timation, and selective model invocation, Blackhole AI es-tablishes a cost-aware and scalable framework for real-world AI deployment. This background sets the foundation for the proposed adaptive routing and optimization strategy discussed in subsequent sections.

B. Motivation

While Large Language Models (LLMs) have achieved im-pressive performance across a wide range of language tasks, their deployment in real-world systems presents significant operational challenges. Most production environments rely heavily on cloud-based LLM APIs, where each query incurs computational cost and latency. As query volumes scale, the cumulative financial and infrastructural burden becomes substantial [2].

Although Retrieval-Augmented Generation (RAG) improves response grounding by incorporating external knowledge, it does not inherently solve the issue of computational ineffi-ciency [2]. In most existing systems, every query—regardless of its complexity—is processed using the same large-scale model. This uniform processing strategy leads to unnecessary cloud invocations for simple queries that could otherwise be handled locally.

The motivation for Blackhole AI arises from three key limitations in current LLM and RAG deployments:

1) High Inference Cost: Cloud-based LLM usage is billed per token or per request, making large-scale deployment expensive, especially for high-frequency query environ-ments.
2) Uniform Processing Pipeline: Existing systems lack adaptive mechanisms to distinguish between low-complexity and high-complexity queries before model invocation.
3) Latency Constraints: Applications such as enterprise assistants and real-time support systems require rapid response times, which may be hindered by repeated cloud communication overhead.

To address these challenges, the objective of Blackhole AI is to learn an adaptive routing function:

$$f^* : Q \rightarrow A \qquad (3)$$

where Q denotes the user query and A represents the generated response. Unlike conventional systems, the proposed framework minimizes overall operational cost while preserving response quality.

The optimization objective can be formulated as:

$$\min C_{total} \quad \text{s.t.} \quad \text{Accuracy} \geq \delta \qquad (4)$$

where C_{total} represents cumulative inference cost and δ is a minimum acceptable accuracy threshold.

By integrating semantic embeddings, retrieval confidence, and complexity estimation into the routing decision, Blackhole AI aims to balance three competing factors: cost efficiency, response accuracy, and inference latency. This motivation establishes the need for a hybrid, cost-aware, and scalable architecture capable of intelligent query handling in modern AI systems.

C. Research Contributions

The primary contributions of Blackhole AI are summarized as follows:

1) A cost-aware adaptive query routing framework that dynamically selects between local and cloud-based Large Language Models (LLMs) based on query complexity.
2) A quantitative complexity estimation mechanism that combines query length, semantic hardness, and retrieval confidence to support intelligent routing decisions.
3) A hybrid retrieval architecture integrating FAISS-based vector search with web augmentation to improve contex-tual relevance and knowledge coverage.
4) An optimization-driven approach that balances response accuracy, inference latency, and operational cost within a unified framework.
5) A scalable architecture suitable for enterprise AI as-sistants, knowledge-intensive systems, and cost-sensitive LLM deployments.

The proposed framework addresses key limitations of ex-isting Retrieval-Augmented Generation (RAG) systems by introducing adaptive resource allocation and intelligent model selection, thereby enabling more efficient and economically sustainable AI services.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

II. SYSTEM ARCHITECTURE AND CORE METHODOLOGY

Blackhole AI integrates multiple sub-domains of Natural Language Processing (NLP), Information Retrieval (IR), and adaptive decision-making systems to optimize large language model deployment. Unlike conventional Retrieval-Augmented Generation (RAG) systems that follow a fixed inference pipeline, the proposed framework introduces intelligent query routing driven by semantic complexity and contextual confi-dence.

The overall architecture combines semantic embedding, hybrid retrieval, complexity estimation, adaptive routing, and response generation within a unified computational frame-work. Each component is designed to balance response quality with operational efficiency, enabling scalable and cost-aware model deployment.

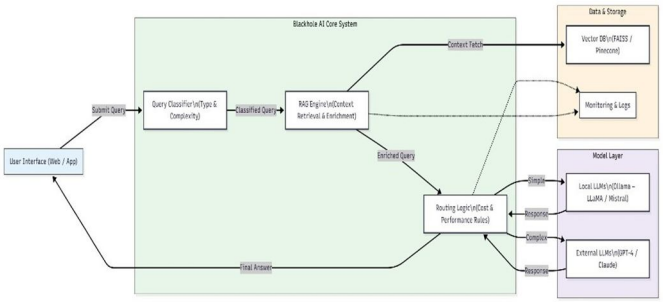


Fig. 1. Overall architecture of the Blackhole AI system integrating semantic embedding, hybrid retrieval, complexity estimation, and adaptive routing.

A. Semantic Feature Extraction

The first stage transforms a user query into a dense semantic representation using embedding models such as Sentence-BERT. This mapping converts textual input into a continuous vector space that preserves contextual and relational meaning.

Given a query Q, the embedding function E_θ produces:

$$z = E_\theta(Q) \qquad (5)$$

where z represents the semantic embedding of the query. These embeddings form the computational basis for both contextual retrieval and downstream complexity estimation.

B. Hybrid Retrieval: Vector Search and Web Augmentation

To ensure comprehensive and up-to-date knowledge cover-age, Blackhole AI employs a hybrid retrieval mechanism that combines vector-based semantic search with controlled web-based knowledge acquisition.

Initially, semantically relevant documents are retrieved from a vector database using similarity search:

$$D_k = \text{FAISS}(z) \qquad (6)$$

where D_k denotes the top-k documents most similar to the embedding z.

However, static vector indices may not always contain re-cent or domain-specific information. To address this limitation, a retrieval confidence score $R(Q)$ is computed based on sim-ilarity distribution and contextual relevance. If this confidence falls below a predefined threshold τ, the system dynamically activates an external knowledge acquisition function:

$$D_{web} = \text{WebFetch}(Q) \qquad (7)$$

The externally acquired content undergoes preprocessing steps including HTML parsing, noise filtering, duplicate re-moval, and semantic embedding to ensure reliability and semantic coherence. The refined documents are then merged with D_k to form an enriched contextual set for response generation.

This hybrid strategy enables the system to maintain the efficiency of indexed retrieval while extending its capability to incorporate real-time information when necessary.

C. Query Complexity Estimation

Unlike traditional RAG systems that apply uniform infer-ence strategies, Blackhole AI evaluates query difficulty prior to model invocation. The complexity score $C(Q)$ is computed as a weighted combination of semantic and structural factors:

$$C(Q) = \alpha L(Q) + \beta H(Q) + \gamma D_{conf} \qquad (8)$$

where:

- $L(Q)$ represents query length,

- $H(Q)$ denotes semantic hardness,
- D_{conf} represents retrieval confidence,
- α, β, γ are tunable weighting parameters.

This scoring formulation enables quantitative assessment of query difficulty and supports principled routing decisions.

D. Adaptive Model Routing

Based on the computed complexity score, the system dy-namically selects between a local language model M_l and a cloud-based model M_c:

$$\Delta - \begin{cases} M_l(D_k \cup D_{web}) & \text{if } C(Q) < T \\ M_c(D_k \cup D_{web}) & \text{otherwise} \end{cases} \quad (9)$$

where T denotes a routing threshold. Simple or low-complexity queries are processed locally to reduce latency and operational cost, while complex queries are routed to higher-capacity cloud models for enhanced reasoning performance.

This adaptive strategy ensures efficient resource allocation without compromising response quality.

E. Evaluation Metrics

System performance is evaluated using multiple quantitative metrics:

1) Response Accuracy – correctness and relevance of gen-erated answers.
2) Inference Cost – monetary expenditure per query.
3) Latency – response time measured per request.
4) Routing Efficiency – percentage of correctly routed queries based on complexity.

F. Adaptive Modeling and Query Routing Framework

After establishing the foundational AI domains in Section II, we now describe the modeling framework that enables Blackhole AI to dynamically route queries while maintaining performance and efficiency.

Unlike conventional Retrieval-Augmented Generation (RAG) systems that follow a fixed processing pipeline, the proposed approach introduces an adaptive decision layer that evaluates query characteristics before model invocation.

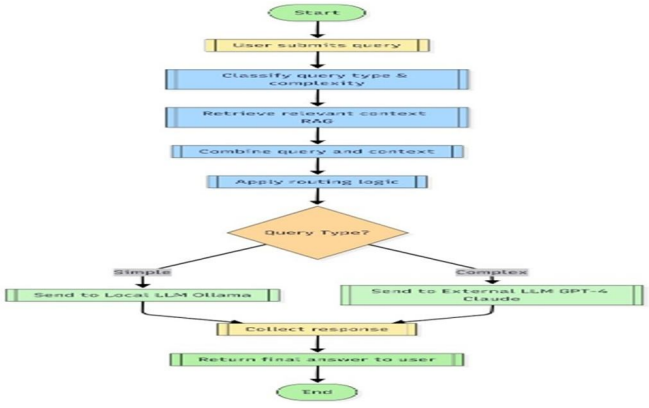


Fig. 2. Workflow of adaptive query routing distinguishing simple and complex queries.

G. Semantic Embedding and Retrieval Modeling

The first stage converts the user query into a dense semantic representation using a transformer-based embedding model. Given a query Q, the embedding function E_θ produces:

$$z = E_\theta(Q) \quad (10)$$

where z denotes the contextual embedding of the input query.

Relevant supporting documents are then retrieved using vector similarity search:

$$D_k = \text{Retrieve}(z) \quad (11)$$

where D_k represents the top-k semantically similar docu-ments from the knowledge base. This retrieval step improves response grounding and reduces hallucination.

H. Query Complexity Modeling

To avoid unnecessary cloud inference, Blackhole AI in-troduces a query complexity estimation mechanism. Instead of treating all queries uniformly, the system computes a complexity score:

$$C(Q) = \alpha L(Q) + \beta S(Q) + \gamma R(Q) \quad (12)$$

where:

- $L(Q)$ represents structural length or token count,
- $S(Q)$ captures semantic difficulty,
- $R(Q)$ reflects retrieval confidence,
- α, β, γ are weighting parameters.

This score provides a quantitative basis for routing deci-sions.

I. Adaptive Routing Mechanism

Based on the computed complexity score, the system selects the appropriate inference model:

$$A = \begin{cases} M_l(D_k \cup D_{web}) & \text{if } C(Q) < T \\ M_c(D_k \cup D_{web}) & \text{otherwise} \end{cases} \quad (13)$$

where M_l denotes the local language model, M_c denotes the cloud-based large language model, and T is a predefined routing threshold.

This adaptive mechanism ensures that computationally sim-ple queries are processed locally, while complex queries leverage higher-capacity cloud models.

J. Cost-Aware Optimization Objective

Since cloud-based inference incurs higher financial cost, the system models cumulative operational cost as:

$$C_{total} = \sum_{i=1} (c_l \cdot \mathbb{1}_{local} + c_c \cdot \mathbb{1}_{cloud}) \quad (14)$$

The optimization objective is formulated as:

$$\min C_{total} \quad \text{s.t. Accuracy} \geq \delta \quad (15)$$

where δ represents a minimum acceptable performance threshold.

By integrating semantic retrieval, complexity estimation, and adaptive model selection, Blackhole AI achieves a balance between accuracy, latency, and operational cost.

III. COMPARATIVE ANALYSIS

A comparative analysis of existing retrieval-augmented and large language model deployment strategies is presented in Table I. The comparison focuses on routing capability, cost-awareness, scalability, and practical deployment efficiency.

Traditional cloud-based LLM systems provide high accu-racy but incur significant operational cost due to uniform processing of all queries. Basic Retrieval-Augmented Gen-eration (RAG) frameworks improve factual grounding but still lack adaptive routing mechanisms. Recent routing-based approaches introduce partial decision strategies, yet many fail to explicitly optimize cost and latency simultaneously.

Blackhole AI differs by integrating semantic retrieval, quan-titative complexity estimation, and dynamic model selection within a unified framework. The proposed system balances response quality with computational efficiency, making it suitable for scalable real-world deployment.

TABLE I
COMPARISON OF QUERY PROCESSING FRAMEWORKS

Method	Adaptive Routing	Cost Aware	Latency Efficient	Scalability
Cloud-only LLM	No	No	No	Medium
Basic RAG	No	No	Moderate	High
RAGRouter-like Systems	Partial	Partial	Moderate	High
Blackhole AI (Proposed)	Yes	Yes	Yes	High

IV. PROPOSED EVALUATION FRAMEWORK

Future validation of Blackhole AI will be conducted using quantitative metrics including response accuracy, inference latency, routing efficiency, operational cost, and scalability. These metrics will be used to evaluate the effectiveness of adaptive query routing compared to traditional cloud-only and Retrieval-Augmented Generation (RAG) systems. The evaluation will focus on determining whether the pro- posed framework can reduce computational cost while main-taining acceptable response quality and system performance.

V. DISCUSSION AND RESEARCH GAPS

Despite progress in retrieval-augmented and hybrid LLM systems, several challenges remain in achieving efficient and scalable deployment.

1) Uniform Model Invocation: Many systems continue to process all queries using large models, leading to avoidable computational overhead.
2) Reliable Complexity Modeling: Accurately estimating query difficulty remains challenging, particularly for se-mantically ambiguous or multi-step queries.
3) Cost-Aware Optimization: Existing research prioritizes accuracy improvements but rarely integrates explicit cost modeling into routing strategies.

TABLE II
COMPARATIVE ANALYSIS OF RETRIEVAL AND QUERY ROUTING FRAMEWORKS

Approach	Core Mechanism	Strengths	Limitations	Suitable Applications
Cloud-only LLM	Single large cloud model for all queries	High reasoning capability; strong performance across domains	High operational cost; in-creased latency; no adaptive routing	Enterprise assistants, research tools with low query volume
Basic RAG	Embedding + Vector Re-trieval + Single LLM	Improved factual grounding; reduced hallucination	Uniform processing; no cost-aware decision making	Knowledge-based Q&A sys-tems, document search
RAGRouter (2023) [4]	Contrastive learning -based query routing	Introduces adaptive routing; improves inference efficiency	Limited explicit cost model-ing; threshold tuning complex-ity	Large-scale AI assistants, hy-brid retrieval systems
Blackhole AI (Pro-posed)	Semantic embedding + FAISS retrieval + Com-plexity estimation + Cost-aware adaptive routing	Balances accuracy, latency, and operational cost; scalable hybrid deployment; supports local and cloud integration	Requires robust complexity es-timation; threshold optimiza-tion needed	Enterprise AI assistants, scalable LLM services, cost-sensitive deployments

4) Scalability Under High Throughput: Maintaining sta-ble routing decisions under large query volumes requires further study.
5) Adaptive Threshold Learning: Fixed routing thresholds may not generalize across domains or workloads, high-lighting the need for dynamic threshold adaptation.
6) Security and Privacy Concerns: Hybrid deployments involving cloud APIs introduce potential data exposure risks that must be carefully managed.

Addressing these challenges requires integrating complexity-aware learning, cost-performance modeling, and scalable deployment strategies within unified AI frameworks.

VI. CONCLUSION

This work presented Blackhole AI, an adaptive query rout-ing framework aimed at improving the efficiency of large lan-guage model deployments. While cloud-based LLMs provide strong reasoning capabilities, applying the same high-capacity model to every query results in unnecessary computational overhead and increased operational cost. Blackhole AI ad-dresses this challenge by combining semantic embeddings, hybrid retrieval mechanisms, query complexity modeling, and adaptive model selection within a unified system.

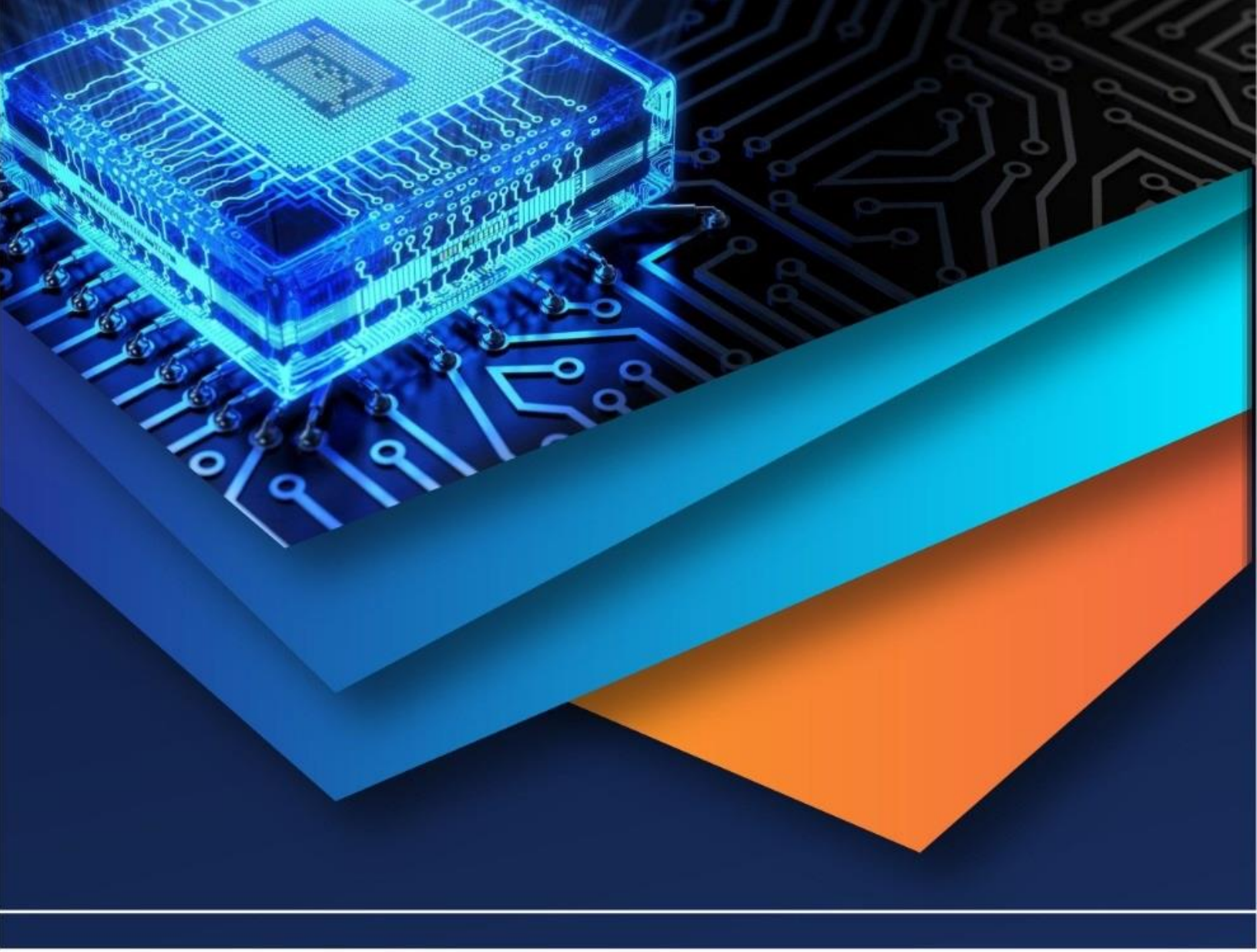
Instead of relying on a single inference strategy, the frame-work evaluates the nature of each query before execution. Low-complexity queries are processed using local models to reduce latency and cost, while more demanding tasks are routed to cloud-based models capable of deeper reasoning. The use of a quantitative complexity score enables structured decision-making and avoids arbitrary or purely heuristic rout-ing.

Although the proposed framework demonstrates a practical direction for cost-aware LLM deployment, several aspects require further investigation. Future work will focus on im-proving threshold learning mechanisms, strengthening robust-ness of complexity estimation, and validating performance un-der large-scale real-world workloads. Exploring reinforcement learning-based routing and multimodal extensions may further enhance adaptability.

In summary, Blackhole AI highlights the importance of intelligent routing in modern AI systems. As large language models continue to scale, efficient resource allocation will become as critical as model accuracy. Adaptive routing frame-works such as Blackhole AI provide a pathway toward sus-tainable, scalable, and economically viable AI services.

REFERENCES

[1] T. Izacard and P. Grave, "Leveraging Retrieval-Augmented Generation for Efficient Language Model Responses," IEEE Access, vol. 10, pp. 55231-55245, 2022.
[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Proc. NeurIPS, 2020.
[3] H. Zhou, Y. Zhang, and J. Tang, "A Survey on Routing Strategies in Large Language Models," IEEE Trans. Neural Networks Learn. Syst., 2024.
[4] R. Alfina et al., "RAGRouter: Query Routing for Retrieval-Augmented Language Models," IEEE Access, 2023.
[5] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transform-ers," Proc. NAACL, 2019.
[6] A. Vaswani et al., "Attention Is All You Need," Proc. NeurIPS, 2017.
[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese Networks," Proc. EMNLP, 2019.
[8] J. Johnson, M. Douze, and H. Jegou, "Billion-Scale Similarity Search with GPUs," IEEE Trans. Big Data, 2021.
[9] K. Cheng et al., "Efficient Batch Serving for LLM-as-a-Service," arXiv, 2024.
[10] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Approach," arXiv, 2019.
[11] T. Brown et al., "Language Models are Few-Shot Learners," Proc. NeurIPS, 2020.
[12] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain QA," Proc. EMNLP, 2020.
[13] N. Shazeer et al., "Switch Transformers: Scaling Efficient Sparse Mod-els," JMLR, 2022.
[14] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY