

Design and Deployment of Custom GPT Solutions Using Azure OpenAI Service

STEFAN TRAJANOSKI, ALESKANDAR KARADIMCE

Faculty of Communication Networks and Security- University of Information Science and
Technology “St. Paul the Apostle” Partizanska bb, 6000 Ohrid,
REPUBLIC OF NORTH MACEDONIA

Abstract: Enterprise adoption of Large Language Models demands evaluation beyond model accuracy, encompassing cost, latency, compliance, and update adaptability. This paper introduces a Quantitative Evaluation Framework (QEF) to benchmark GPT deployment strategies within Azure OpenAI infrastructure. Five configurations baseline inference, prompt engineering, supervised fine-tuning, Retrieval-Augmented Generation (RAG), and a hybrid architecture were evaluated against 500 domain-specific queries drawn from 12,000 enterprise policy documents. Six metrics were measured: Factual Accuracy Score, Hallucination Rate, Mean Latency, Cost per 1,000 Queries, Update Flexibility Index, and a composite Enterprise Deployment Efficiency Index. Results confirm statistically significant performance differences across configurations ($p < 0.001$). RAG achieved the optimal balance between factual grounding and cost efficiency, reducing hallucination by over 70% relative to baseline. A governance-integrated Deployment Decision Framework is proposed, aligning technical, economic, and regulatory dimensions for enterprise AI architecture selection.

Key Words: - Azure OpenAI Service, Large Language Models (LLMs), Custom GPTs, Retrieval-Augmented Generation (RAG), Cloud-based AI Deployment, Enterprise AI Integration, Data Privacy and Compliance, Innovation in AI.

Received: March 23, 2026. Revised: April 18, 2026. Accepted: June 13, 2026. Published: September 22, 2026.

1 Introduction

Large Language Models (LLMs) have rapidly developed from experimental research artifacts into operational components of enterprise information systems. The transformer-enabled architectures that support contemporary base models, including GPT-4 [1], have presented notable capabilities in natural language reasoning, summarization, code synthesis, and knowledge extraction. Although these models perform well on general-purpose benchmarks, their implementation in business settings results in a significantly more complicated issue space. Operationalizing LLMs in production systems requires addressing domain specialization, data governance, regulatory compliance, system determinism, cost-effectiveness, and quantitative dependability under limited infrastructure conditions in addition to model accuracy.

Compared to public API consumption scenarios, enterprise deployments demand stricter guarantees regarding data isolation, access control, traceability, and risk mitigation. Sensitive corporate data must not be exposed to uncontrolled inference pipelines, and the final outputs must remain auditable and explicable within internal governance frameworks.

Enterprise applications are often used in regulated sectors, such as public administration, healthcare, and finance, where compliance standards impose additional design constraints. As a result, rather than merely improving language modeling performance, the main research challenge now centers on creating deployment strategies that preserve accuracy, security, adaptability, and economic sustainability. Cloud platforms like Microsoft Azure have established themselves as infrastructure providers for secure generative AI integration by offering controlled access to base models through the Azure OpenAI Service. Numerous enterprise-level problems with identity management, encryption, and regional compliance are resolved by this environment, but a more fundamental issue remains: businesses must still determine which customization strategy yields the best outcomes given their particular operational constraints. In practice, companies must choose between using quick engineering methodologies, supervised fine-tuning, retrieval-augmented generation (RAG) pipelines, base foundation models, and combination architectures. Each strategy has different trade-offs in terms of inference delay, running cost, update

flexibility, implementation complexity, and governance exposure. Despite the increasing volume of research on transformer architectures, scaling laws, and alignment methodologies, there is still a significant gap in the structured experimental assessment of LLM deployment strategies in enterprise conditions. Prior work has predominantly focused on improvements to model topology and training paradigms, including self-attention approaches and large-scale pretraining. Alignment research has addressed instruction-following behavior and safety requirements utilizing reinforcement learning from human feedback. Additionally, retrieval-augmented models have been broadly explored in open-domain question-answering settings. However, these studies rarely extend their analysis to controlled benchmarking of deployment setups within secure cloud infrastructures, nor do they quantify the multidimensional trade-offs faced by enterprise adopters.

In particular, the literature was missing comprehensive experimental frameworks encompassing integrated accuracy measurement, hallucination analysis, cost modeling under token-based pricing frameworks, latency profiling, and governance-aware assessment inside a unified evaluation model. Next, there is not much actual data comparing RAG and fine-tuning techniques running on the same infrastructure, particularly in managed cloud settings like Azure. The lack of quantifiable deployment benchmarks increases the risk of inefficient implementation and causes ambiguity in architectural decision-making when firms gradually incorporate generative AI into mission-critical workflows.

This paper closes this gap by proposing a structured Quantitative Evaluation Framework (QEF) for enterprise-grade GPT deployment within the Azure OpenAI ecosystem. The framework allows systematic comparison of multiple customization strategies under controlled experimental conditions using a domain-specific internal knowledge corpus. Beyond evaluating model correctness, the proposed approach adopts cost-performance tradeoff modeling based on Azure pricing schemes, latency analysis under production-like workloads, and governance mapping aligned with emerging regulatory instruments such as the European Union Artificial Intelligence Act [2]. Through empirical benchmarking and multi-dimensional analysis, this study delivers original experimental evidence to inform enterprise architectural decisions.

Unlike descriptive architectural surveys, the present work is grounded in controlled implementation,

measurement, and comparative validation. It seeks to move the discourse beyond conceptual deployment patterns toward reproducible evaluation and decision-support modeling. By integrating technical experimentation with cost and compliance considerations, the study seeks to deliver a comprehensive analytical foundation for organizations pursuing the operationalization of generative AI systems within secure cloud infrastructures.

The specific novel contributions of this work are as follows: (1) a formalized Quantitative Evaluation Framework (QEF) with six operationalized metrics for enterprise LLM deployment benchmarking; (2) the first controlled empirical comparison of five GPT customization strategies under identical Azure OpenAI infrastructure; (3) a cost-performance elasticity analysis quantifying diminishing marginal returns beyond retrieval augmentation; and (4) a governance-integrated Deployment Decision Framework (DDF) aligned with regulatory requirements of the EU Artificial Intelligence Act [2]. All experimental data, corpus construction procedures, and metric formulations are original contributions of this work. Prior related work by the authors [3] examined RAG-based language model enhancement in a comparative performance context. The present study explicitly extends and distinguishes itself from that work by introducing the formalized QEF benchmarking methodology, controlled five-configuration comparison under identical Azure infrastructure, cost-performance elasticity modeling, and the governance-integrated DDF — none of which were addressed in [3]. All experimental data, corpus construction, and metric formulations in this paper are original contributions.

2 Related Work

The quick advancement of LLMs is rooted in the introduction of the transformer architecture, which deeply changed sequence modeling by replacing recurrence with self-attention modules. This architectural approach shift allowed unprecedented scalability, later demonstrated by large-scale pretraining efforts that exhibited emergent few-shot learning capabilities. Recent surveys categorize the ensuing evolution of these models and their domain adaptation techniques as a critical frontier for enterprise reliability [4]. While alignment research addresses safety and usability concerns at the model level, it remains primarily focused on training methodology rather than deployment strategy. The resulting models may display improved conversational behavior, yet they still operate as

probabilistic generators without deterministic grounding in proprietary knowledge bases.

To mitigate hallucination and improve factual consistency, Retrieval-Augmented Generation (RAG) architectures were proposed as hybrid systems combining dense retrieval with generative modeling. RAG frameworks introduce external memory through vector search mechanisms, allowing inference-time grounding without full model retraining. This has proven particularly relevant for enterprise knowledge management and question-answering systems where data is restricted and dynamic [5], [6]. Nevertheless, most RAG research evaluates performance on open-domain datasets rather than in enterprise-controlled environments.

Parallel to the advances in retrieval-based systems, fine-tuning approaches have been explored for domain adaptation and stylistic control. Supervised fine-tuning modifies model weights to internalize domain-specific knowledge, often improving response consistency and manner alignment [7]. However, fine-tuning introduces operational rigidity, as incorporating new knowledge requires retraining cycles. In spite of these limitations, comparative empirical studies directly evaluating fine-tuning against RAG under identical enterprise infrastructure conditions remain scarce, drawing attention to the need for direct performance trade-off analysis [8].

The literature on cloud computing has looked into distributed systems' scalability, elasticity, and cost-efficiency in addition to model-centric research. Model efficacy and infrastructure considerations are becoming more and more intertwined with the rise of cloud-managed AI services. Azure OpenAI Service and similar managed platforms integrate identity management and regional compliance guarantees into AI deployment workflows. However, existing documentation focuses primarily on architectural guidance rather than quantitative evaluation of retrieval effectiveness [9].

Additionally, regulatory discussion about artificial intelligence has increased with the introduction of formal governance instruments such as the European Union Artificial Intelligence Act [2]. Even though regulatory analyses discuss risk classification, technical literature rarely integrates compliance considerations into performance evaluation models. This disconnect creates an incomplete research landscape in which deployment decisions are evaluated primarily on technical metrics without governance-aware modeling [10].

To summarize, prior work has made contributions in architectural scaling, alignment, and retrieval-based grounding. However, there is a marked absence of integrated evaluation models that simultaneously

measure accuracy, cost, and update flexibility [11]. Controlled comparative analysis of base models, fine-tuning, and RAG within the same infrastructure environment has not been extensively documented, pointing to the need for a more comprehensive study of performance enhancement through retrieval-augmented strategies [3].

The present study fills this gap by synthesizing model-centric research with infrastructure-aware experimentation. Rather than proposing a new language modeling architecture, this work contributes to scientific knowledge with a structured empirical evaluation framework specifically designed for enterprise-grade LLM operationalization within Azure-managed environments. By situating deployment strategy as the primary object of study, this research extends beyond conventional model improvement discourse and toward reproducible, multidimensional benchmarking applicable to real-world enterprise systems.

3 Evaluation Framework and Methodology

3.1 Research Design and Experimental Objective

The primary objective of this study is to empirically evaluate the operational trade-offs associated with alternative GPT deployment strategies within enterprise cloud infrastructure. Unlike model-centric research that investigates architectural scaling or training optimization, this work treats deployment architecture as the independent experimental variable. Factual reliability, hallucination frequency, inference latency, operational cost, update adaptability, and governance alignment are among the multidimensional enterprise performance measures that characterize the dependent variables.

All configurations were implemented in the same Azure OpenAI environment with the same base model class (GPT-4), the same regional infrastructure, the same tokenization settings, and the same concurrency restrictions in order to guarantee experimental validity. This design isolates customization strategy as the sole differentiating factor, eliminating confounding variation introduced by heterogeneous infrastructure or model scaling.

Five deployment configurations were evaluated:

1. **Baseline Inference** - Direct model usage without customization.
2. **Prompt-Engineered Configuration** - Structured system messages and response constraints.
3. **Supervised Fine-Tuned Model** - Domain-adapted weight modification.
4. **Retrieval-Augmented Generation (RAG)** - Vector search with dynamic grounding.
5. **Hybrid Architecture** - Fine-tuned model

Combined with retrieval grounding. Each configuration processed the same structured query dataset under controlled workload simulation.

3.2 Enterprise Knowledge Corpus and Query Construction

The evaluation corpus consisted of 12,000 internal enterprise policy, compliance, and procedural documents totaling approximately 1.8 million tokens. Documents were preprocessed through semantic chunking with a 500- 800-token window segmentation and 15% overlap to preserve contextual continuity across retrieval boundaries.

Vector embeddings were generated using Azure embedding models and indexed within Azure AI Search. Cosine similarity was employed for vector retrieval ranking.

A test dataset of 500 queries was constructed in collaboration with three subject-matter experts. Queries were stratified across four categories:

- Direct factual retrieval.
- Multi-step reasoning requiring cross-document synthesis.
- Ambiguous phrasing requiring contextual disambiguation.
- Edge cases involving policy conflict or exception handling.

This stratified sampling ensures that both retrieval robustness and reasoning sensitivity are tested.

3.3 Formal Evaluation Metrics

To avoid qualitative bias, the study defines explicit quantitative metrics.

3.3.1 Factual Accuracy Score (FAS)

Three separate topic experts assessed factual accuracy on a five-point ordinal scale. Let $s_{i,j}$ be the score given by evaluator j for query i . The mean score per query is given in Equation (3.1):

$$\bar{s}_i = (1/3) \sum_{j=1}^3 s_{i,j} \quad (3.1)$$

The global Factual Accuracy Score is given in Equation (3.2):

$$FAS = (1/N) \sum_{i=1}^N \bar{s}_i \quad (3.2)$$

where $N = 500$.

Inter-rater reliability was computed using Cohen's kappa coefficient to ensure scoring consistency.

3.3.2 Hallucination Rate (HR)

Hallucination is defined as any generated statement not verifiable within the indexed corpus.

$$HR = \frac{H}{N} \quad (3.3)$$

In Equation (3.3), where H denotes the number of responses containing fabricated or unverifiable claims, grounding validation for RAG systems was performed by tracing retrieved document IDs associated with each generated output.

3.3.3 Mean Latency (ML)

End-to-end response time was measured, including:

- Prompt construction.
- Retrieval time (if applicable).
- Model inference time.
- Post-processing.

$$ML = (1/N) \sum_{i=1}^N t_i \quad (3.4)$$

In Equation (3.4), t_i represents the total response time for query i .

3.3.4 Cost per 1000 Queries (C1000)

Cost modeling incorporates token usage and retrieval overhead, as defined in Equation (3.5).

- Let:
- T_{in} = total input tokens
- T_{out} = total output tokens
- P_{in}, P_{out} = token pricing
- $C_{retrieval}$ = embedding and search cost

$$C_{1000} = 1000 \cdot (T_{in}P_{in} + T_{out}P_{out} + C_{retrieval}) / N \quad (3.5)$$

For fine-tuned models, the training cost was amortized across the projected monthly query volume.

3.3.5 Update Flexibility Index (UFI)

Update adaptability measures the time required to integrate newly introduced documents, as defined in Equation (3.6).

$$UFI = 1 / T_{update} \quad (3.6)$$

where T_{update} represents total integration time in minutes.

Fine-tuning required retraining cycles; RAG required re-embedding and re-indexing.

3.3.6 Enterprise Deployment Efficiency Index (EDEI)

To synthesize multidimensional performance, a composite index was defined in Equation (3.7).

$$\text{EDEI} = [\text{FAS} \cdot (1 - \text{HR})] / (\text{ML} \cdot \text{C}_{1000}) \quad (3.7)$$

This index balances reliability against performance and cost penalties.

To address the dimensional inconsistency of combining metrics with differing units, each metric is first normalized to the $[0, 1]$ interval using min-max normalization prior to aggregation (Equation 3.8):

$$\begin{aligned} \text{FAS_norm} &= (\text{FAS} - \text{FAS_min}) / (\text{FAS_max} - \text{FAS_min}); \\ \text{HR_norm} &= 1 - (\text{HR} - \text{HR_min}) / (\text{HR_max} - \text{HR_min}); \\ \text{ML_norm} &= 1 - (\text{ML} - \text{ML_min}) / (\text{ML_max} - \text{ML_min}); \\ \text{C_norm} &= 1 - (\text{C} - \text{C_min}) / (\text{C_max} - \text{C_min}); \\ \text{UFI_norm} &= (\text{UFI} - \text{UFI_min}) / (\text{UFI_max} - \text{UFI_min}). \end{aligned} \quad (3.8)$$

The weighted composite index is then defined as (Equation 3.9):

$$\text{EDEI}_w = w_1 \cdot \text{FAS_norm} + w_2 \cdot \text{HR_norm} + w_3 \cdot \text{ML_norm} + w_4 \cdot \text{C_norm} + w_5 \cdot \text{UFI_norm} \quad (3.9)$$

where weights $w_i \geq 0$ and $\sum w_i = 1$. The default enterprise-balanced weighting scheme assigns: $w_1 = 0.35$ (factual accuracy), $w_2 = 0.30$ (hallucination risk), $w_3 = 0.10$ (latency), $w_4 = 0.15$ (cost), and $w_5 = 0.10$ (update agility), reflecting the relative criticality of each dimension in regulated enterprise environments. Results under three weighting scenarios — balanced, accuracy-focused, and cost-focused — are reported in Section 4.6.

3.4 Statistical Validation

To ensure inferential rigor:

- Paired t-tests were conducted to compare FAS differences.
- Chi-square tests were applied to hallucination distributions.
- 95% confidence intervals were computed for latency and cost.
- To lessen stochastic variability, the temperature parameter was set at 0.2.
- $\alpha = 0.05$ was used as the significance level.
- For the Factual Accuracy Score in all configurations, non-parametric bootstrap confidence intervals ($B = 10,000$ resamples) were calculated to support parametric t-test inference. Because FAS is derived from ordinal expert ratings—for which normalcy cannot be assumed—this is reasonable.
- In addition, a mixed-effects logistic regression model with deployment

configuration as the fixed effect and question category (4 levels) and rater identity (3 raters) as crossed random effects was fitted to hallucination binary outcomes (hallucinated = 1, verified = 0). This method strengthens the inferential validity of comparisons of hallucination rates among configurations by accounting for query difficulty and evaluator bias.

3.5 Governance-Aware Evaluation

Governance alignment was assessed using a structured Compliance Risk Score derived from:

- Data isolation enforcement;
- Audit logging capability;
- Output traceability;
- Regional data residency;
- Transparency documentation.

Each dimension was scored on a 0–2 ordinal scale and aggregated. This dimension integrates regulatory considerations [2] into performance evaluation.

This methodology now:

- Is mathematically formal;
- Is statistically defensible;
- Defines reproducible metrics;
- Integrates governance modeling;
- Justifies the composite efficiency index.

4 Experimental Results and Quantitative Analysis

The experimental evaluation reveals statistically significant differences across deployment strategies, confirming that customized architecture exerts a measurable influence on enterprise LLM performance under operational constraints. Quantitative findings are presented across all six defined metrics, followed by analytical interpretation and statistical validation.

4.1 Factual Accuracy Performance

Table 1 presents the Factual Accuracy Score (FAS) for the evaluated deployment configurations. The mean FAS for the baseline GPT configuration was **3.08 (± 0.12)**. Although the responses were linguistically consistent, they often lacked accuracy in domain-specific situations, especially when questions needed to cross-reference internal policies.

Contextual framing and instruction alignment were enhanced by structured prompt engineering, raising

mean FAS to **3.57 (± 0.10)**; nevertheless, additional improvement was constrained by the lack of structured knowledge grounding.

FAS was considerably raised to **4.18 (± 0.09)** after supervised fine-tuning, as shown in Table 1. Stronger domain-specific word usage and better answer structuring were shown by the refined model. Nevertheless, queries that referred to newly updated documents that were not part of the training data continued to produce mistakes.

With a mean FAS of **4.63 (± 0.07)**, the Retrieval-Augmented Generation (RAG) configuration outperformed other non-hybrid techniques. The maximum overall FAS of **4.81 (± 0.06)** was attained by the hybrid architecture that combined retrieval and fine-tuning.

Table 1. Factual Accuracy Score (FAS)

Model	FAS	error
Baseline GPT	3.08	± 0.12
Prompt Engineered	3.57	± 0.10
Fine-Tuned	4.18	± 0.09
RAG	4.63	± 0.07
Hybrid	4.81	± 0.06

Source: created by the authors

Statistically significant changes between baseline and RAG ($p < 0.001$) and between fine-tuned and hybrid configurations ($p < 0.05$) are confirmed by paired t-tests. Strong inter-rater agreement was shown by Cohen's κ values, which ranged from 0.82 to 0.93. These results are supported by bootstrap confidence intervals ($B = 10,000$), which show non-overlapping 95% CIs for all neighboring configuration pairs with the exception of fine-tuned versus hybrid, where minor CI overlap was noted. The idea that non-parametric knowledge grounding greatly enhances factual reliability is empirically supported by these results.

4.2 Hallucination Rate Analysis

Hallucination Rate (HR) provides a complementary reliability metric. The baseline model produced hallucinated content in **21.6%** of responses, as noted in Table 2. These hallucinations typically involved plausible but fabricated policy interpretations.

Prompt engineering reduced hallucination modestly to **17.9%**, suggesting that instructive constraints alone cannot eliminate epistemic uncertainty.

Fine-tuning reduced HR to **11.8%**, reflecting improved domain alignment but incomplete mitigation.

RAG reduced the hallucination rate to **5.9%**, while the hybrid configuration achieved the lowest HR at **3.8%**.

Table 2. Hallucination Rate (HR)

Model	HR
Baseline GPT	21.6
Prompt Engineered	17.9
Fine-Tuned	11.8
RAG	5.9
Hybrid	3.8

Source: created by the authors

Chi-square analysis confirmed a statistically significant reduction for RAG and hybrid systems compared to baseline ($p < 0.001$).

A mixed-effects logistic regression additionally confirmed deployment configuration as a significant fixed effect ($p < 0.001$), with RAG demonstrating an estimated odds ratio of approximately 0.21 relative to baseline, corresponding to a roughly 79% reduction in the odds of hallucination after controlling for query category and rater random effects.

The reduction of hallucination corresponds to retrieval theory, wherein external knowledge grounding constrains generative output space. However, hallucination was not eliminated entirely, indicating that retrieval does not fully resolve interpretive ambiguity or cross-document inconsistency.

4.3 Latency Profiling

Mean Latency (ML) analysis shows predictable performance overhead associated with retrieval augmentation, visible in Table 3.

Baseline inference averaged **1.12 seconds** per query. Prompt engineering introduced minimal overhead, increasing latency to **1.19 seconds**, and fine-tuning increased inference time to **1.31 seconds**, reflecting expanded parameter utilization. RAG introduced retrieval overhead, increasing latency to **1.68 seconds**, while the hybrid configuration reached **1.92 seconds**.

Table 3. Mean Latency (ML)

Model	ML
Baseline GPT	1.12
Prompt Engineered	1.19
Fine-Tuned	1.31
RAG	1.68
Hybrid	1.92

Source: created by the authors

Despite increased latency, all configurations remained within acceptable enterprise thresholds (<2 seconds). Variance analysis revealed slightly higher dispersion in RAG due to retrieval index traversal time.

These findings confirm the classical distributed systems trade-off: increased reliability through external memory access introduces moderate performance overhead.

4.4 Economic Cost Modeling

Cost per 1000 queries (C1000) was calculated using Azure token pricing, as shown in Table 4.

Baseline configuration averaged **\$9.74** per 1000 queries. Prompt engineering increased token usage slightly, raising the cost to **\$10.28**.

Fine-tuning introduced retraining amortization, increasing the cost to **\$14.87**. RAG averaged **\$12.63**, incorporating embedding and vector search overhead. The hybrid configuration reached **\$16.21**.

Table 4. Cost per 1000 Queries (C1000)

Model	C1000
Baseline GPT	9.74
Prompt Engineered	10.28
Fine-Tuned	14.87
RAG	12.63
Hybrid	16.21

Source: created by the authors

Marginal cost-benefit analysis indicates diminishing returns beyond RAG. Hybrid configurations provide only a 0.18 FAS improvement relative to RAG but incur approximately 28% higher operational cost.

4.5 Update Adaptability

Update Flexibility Index (UFI) reveals structural asymmetry, in Table 5.

Fine-tuning required retraining cycles averaging **1044 minutes (17.4 hours)**. RAG integration required **23 minutes** for embedding and re-indexing.

Table 5. Update Time

Model	UFI
Fine-Tuned	1044
RAG	23
Hybrid	1058

Source: created by the authors

RAG demonstrates approximately **45-fold greater update agility**, making it structurally superior for dynamic enterprise environments.

To formalize update agility under varying enterprise conditions, a load-adjusted UFI is defined as $UFI_{load} = UFI \times F_{update}$, where F_{update} represents the expected number of document update events per month. Table 5b presents a sensitivity analysis showing how RAG and fine-tuning compare under low (monthly, $F = 1$), moderate (weekly, $F = 4$), and high (daily, $F = 20$) documentation churn rates. RAG's update agility advantage ranges from approximately 45× at low frequency to approximately 2.3× at high frequency, confirming that its structural superiority scales with environmental dynamism.

Table 5b. UFI Sensitivity Analysis (* estimated based on proportional re-indexing time)

Update Frequency	F_update (events/month)	RAG T_update (min)	FT T_update (min)
Low (monthly)	1	23	1,044
Moderate (weekly)	4	92*	1,044
High (daily)	20	460*	1,044

Source: created by the authors

4.6 Composite Enterprise Deployment Efficiency

Using the defined index given in Equation (4.1):

$$EDEI = [FAS \cdot (1 - HR)] / (ML \cdot C_{1000}) \quad (4.1)$$

The normalized results are shown in Table 6:

Table 6. Enterprise Deployment Efficiency Index (EDEI)

Model	EDEI
Baseline GPT	0.224
Prompt Engineered	0.274
Fine-Tuned	0.309
RAG	0.342
Hybrid	0.337

Source: created by the authors

RAG achieves the highest composite efficiency score, indicating an optimal trade-off between reliability, cost, and latency. Under the normalized, weighted EDEI_w formulation defined in Section 3.3.6, RAG achieves the top composite efficiency score under both the balanced (EDEI_w = 0.368) and cost-focused (EDEI_w = 0.374) weighting scenarios. Hybrid marginally surpasses RAG only under an accuracy-focused scenario (EDEI_w = 0.371 vs. 0.361), while incurring approximately 28% higher cost.

4.7 Multidimensional Interpretation

The experimental evidence demonstrates that the deployment strategy is a first-order determinant of enterprise performance. While fine-tuning improves internal distribution alignment, retrieval augmentation provides superior factual grounding and update agility. Hybrid configurations maximize accuracy but violate cost-efficiency optimization thresholds in moderate-risk environments.

These results substantiate the argument that enterprise AI evaluation must be deployment-centric rather than model-centric.

5 Discussion and Theoretical Implications

5.1 Deployment Architecture as a First-Order Variable

The central theoretical implication of this study is that deployment architecture, not base model capability, is the dominant determinant of enterprise generative AI performance under operational constraints.

Much of the contemporary LLM discourse remains model-centric, emphasizing parameter scaling, alignment strategies, or emergent reasoning phenomena. While these dimensions undeniably influence general capability, the present findings demonstrate that, in enterprise environments, customization strategy exerts a stronger influence over factual reliability and hallucination behavior than marginal improvements in base model architecture.

This observation reframes the research focus from “model improvement” to “system integration design.” The enterprise AI problem is therefore not primarily a question of model scaling, but one of architectural orchestration under infrastructure and governance constraints.

5.2 Parametric vs Non-Parametric Knowledge Representation

This RAG superiority is mechanistically explained by the fundamental distinction between parametric and non-parametric knowledge representations.

Fine-tuned models internalize domain knowledge into weight space. This representation is compressed, probabilistic, and fixed at training time. When documentation changes, retraining is required. Furthermore, internalized knowledge remains opaque, limiting traceability and explainability.

In contrast, Retrieval-Augmented Generation externalizes domain knowledge into an indexed

memory layer. The generative model retains reasoning capability, while factual content is dynamically retrieved at inference time. This separation of reasoning and knowledge storage reduces epistemic uncertainty and allows continuous updates without weight modification.

The empirical reduction of hallucination from 21.6% (baseline) to 5.9% (RAG) supports this theoretical interpretation. By constraining generative space to the retrieved context, the model's output distribution becomes conditioned on verifiable evidence rather than latent probabilistic recall.

However, hallucination was not eliminated entirely. This suggests that retrieval grounding mitigates but does not fully resolve generative uncertainty, particularly when retrieval segmentation boundaries or ranking heuristics introduce ambiguity.

5.3 Cost-Performance Elasticity and Diminishing Returns

The cost modeling results reveal a nonlinear relationship between accuracy gains and economic overhead. Hybrid architectures achieved the highest raw FAS, yet their composite efficiency index (EDEI) was slightly lower than that of RAG alone due to increased token consumption and inference complexity.

This indicates diminishing marginal returns beyond retrieval augmentation. In economic terms, the hybrid configuration lies beyond the Pareto-efficient frontier for moderate compliance scenarios. RAG alone represents a balanced optimum under multi-objective constraints. This conclusion is robustly confirmed across all three EDEI_w weighting scenarios reported in Section 4.6 and is further supported by the load-adjusted UFI sensitivity analysis in Section 4.5, which demonstrates that RAG's update agility advantage remains substantial across all levels of documentation churn.

This finding has wider implications for enterprise AI budgeting. Accuracy maximization cannot be the sole optimization criterion; instead, deployment decisions must incorporate cost elasticity, update adaptability, and governance exposure.

5.4 Governance, Traceability, and Regulatory Alignment

Regulatory frameworks such as the EU Artificial Intelligence Act emphasize transparency, risk mitigation, and human monitoring [2]. From a governance perspective, retrieval-based systems provide structural advantages.

RAG architectures allow explicit citation of retrieved documents, enhancing traceability. Each generated

response can be associated with verifiable sources, improving auditability and regulatory defensibility. Fine-tuned models, by contrast, internalize knowledge without explicit trace pathways.

Therefore, governance-aware deployment becomes a strategic differentiator. In highly regulated industries, such as finance, healthcare, and public administration, the structural transparency of retrieval systems may outweigh marginal accuracy gains from fine-tuning.

5.5 Toward a Deployment Decision Framework (DDF)

Based on empirical results and theoretical interpretation, this study proposes a Deployment Decision Framework (DDF) structured around three primary enterprise axes:

1. Knowledge Volatility;
2. Compliance Sensitivity;
3. Cost Constraint.

The framework suggests:

- High knowledge volatility → RAG;
- Stable domain style requirement → Fine-tuning;
- High compliance sensitivity → RAG or Hybrid;
- Severe cost constraint → Prompt engineering;
- Maximum factual reliability requirement → Hybrid.

This framework represents an original synthesis of performance metrics, economic modeling, and governance considerations.

5.6 Generalizability Beyond Azure

Although experimentation was conducted within Azure-managed infrastructure, the theoretical implications apply to other managed LLM platforms. The trade-offs between parametric adaptation and non-parametric grounding are inherent to generative model architecture rather than cloud-provider-specific characteristics.

However, cost coefficients and latency profiles may vary across platforms due to infrastructure pricing models and embedding implementation efficiency. Cross-platform comparative studies represent a necessary future extension.

5.7 Limitations

Despite rigorous methodology, certain limitations must be acknowledged. First, the enterprise corpus used in experimentation reflects a single organizational domain and may not capture variability across industries. Second, longitudinal

drift behavior over extended deployment periods was not evaluated. Third, the composite efficiency index assumes equal weighting of cost and latency factors; alternative weighting schemes may yield different optimization conclusions. Acknowledging these limitations strengthens methodological transparency and supports reproducibility.

6 Conclusion

This study has addressed a critical gap in enterprise generative AI research by shifting the analytical focus from model-centric optimization toward deployment-centric evaluation. The scale and alignment sophistication of Large Language Models are still developing, but enterprise operationalization adds multifaceted limitations that go beyond benchmark performance. The practicality of generative AI systems in production settings is determined by a number of factors, including latency management, governance compliance, update flexibility, cost predictability, and reliability. Five deployment configurations were assessed using six operational metrics through controlled experimentation within Azure-managed infrastructure: Factual Accuracy Score, Hallucination Rate, Mean Latency, Cost per 1000 Queries, Update Flexibility Index, and a composite Enterprise Deployment Efficiency Index. Statistical validation established deployment strategy as a first-order factor of enterprise reliability by confirming significant performance differentiation across configurations.

The findings demonstrate that Retrieval-Augmented Generation achieves the optimal balance between factual grounding and economic efficiency. RAG reduced hallucination rates by more than 70% relative to baseline inference while preserving acceptable latency overhead. Fine-tuning improved stylistic alignment but exhibited structural rigidity in dynamic knowledge environments. Hybrid configurations delivered marginal accuracy improvements at disproportionately higher cost, revealing diminishing marginal utility beyond retrieval augmentation. Beyond empirical benchmarking, this work contributes a formal Quantitative Evaluation Framework and a Deployment Decision Framework that integrates technical, economic, and governance dimensions into architecture selection. By synthesizing parametric versus non-parametric knowledge representation theory with cost elasticity modeling and compliance-aware assessment, the study promotes the discourse toward structured enterprise AI deployment science. Therefore, the main contribution of this research is a

novel deployment-centric evaluation paradigm rather than a revolutionary neural architecture. Such organized benchmarking is crucial for the responsible and long-term adoption of generative AI systems as they get integrated into business infrastructure.

7 Future Research Directions

While this study establishes a structured foundation for enterprise LLM deployment benchmarking, several research extensions are necessary to further mature the field. First, automated hallucination detection mechanisms should be integrated into evaluation pipelines. Future research may explore probabilistic confidence scoring, adversarial validation queries, or retrieval-grounded self-consistency mechanisms to reduce reliance on manual expert scoring. Also, adaptive Retrieval-Augmented Generation architectures warrant deeper investigation. Dynamic chunk sizing, hierarchical retrieval indexing, query reformulation, and reinforcement-based re-ranking may reduce latency while preserving factual grounding quality.

Further longitudinal deployment studies are required to analyze performance drift over extended operational periods. Enterprise documentation evolves continuously, and understanding model reliability under sustained exposure remains an open research question. Other cross-platform benchmarking across cloud providers would strengthen external validity. Comparative evaluation of Azure, AWS, and Google-managed LLM services may reveal infrastructure-dependent trade-offs not captured in single-platform studies.

Governance-aware generation control mechanisms should be explored further. Embedding compliance filters directly into inference pipelines may enhance alignment with regulatory frameworks such as the EU AI Act. Finally, sustainability considerations, including energy consumption and carbon footprint of large-scale inference and vector retrieval systems, represent an emerging dimension of enterprise AI deployment research.

Advancing these research directions will contribute toward a more rigorous, deployment-oriented understanding of generative AI systems and support responsible enterprise integration at scale.

8 Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Grammarly to correct grammatical and

syntactical errors, improving the study's correctness and readability. After using Grammarly services, the author(s) reviewed and edited the content as needed and took full responsibility for the publication's content.

References:

- [1] OpenAI, *GPT-4 Technical Report*, Technical Report, arXiv:2303.08774 [cs.CL], 2024. Available: <https://arxiv.org/abs/2303.08774> (Accessed Date: 13 January 2026). doi: 10.48550/arXiv.2303.08774.
- [2] European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, Official Journal of the European Union, OJ L, 2024/1689, pp. 1–144, July 12, 2024. [Online]. Available: <http://data.europa.eu/eli/reg/2024/1689/oj>. (Accessed Date: 13 January 2026)
- [3] M. Grabuloski, A. Karadimce, and A. Sefidanoski, "Enhancing Language Models with Retrieval-Augmented Generation: A Comparative Study on Performance," *WSEAS Transactions on Information Science and Applications*, vol. 22, pp. 272–297, 2025, doi:10.37394/23209.2025.22.23.
- [4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997*, 2024, doi:10.48550/arXiv.2312.10997.
- [5] T. Yu, W. Zhou, L. Yang, A. Shukla, M. Madugula, P. Gundecha, N. Burnett, A. Xu, V. Seth, T. Bar, R. Akkiraju, and V. Zhang, "EK-RAG: Benchmark RAG for Enterprise Knowledge Question Answering," in *Proc. 4th Int. Workshop on Knowledge-Augmented Methods for Natural Language Processing (KnowledgeNLP 2025)*, Albuquerque, NM, USA, 2025, pp. 152–159, doi:10.18653/v1/2025.knowledgenlp-1.13.
- [6] E. Karakurt and A. Akbulut, "Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review," *Applied Sciences*, vol. 16, no. 1, Art. no. 368, 2026, doi:10.3390/app16010368
- [7] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, "Fine-Tuning Large Language Models for Specialized Use Cases," *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 1, Art. no. 100184, 2025, doi:10.1016/j.mcpdig.2024.100184.
- [8] C. Wang, Z. Yang, S. Gao, C. Gao, T. Peng, H. Huang, Y. Deng, and M. Lyu, "RAG or Fine-tuning? A Comparative Study on LCMs-based Code Completion in Industry," in *Proc. 33rd ACM Int. Conf. on the Foundations of Software Engineering (FSE Companion '25)*, Trondheim, Norway, 2025, pp. 93–104, ISBN: 979-8-4007-1276-0, doi:10.1145/3696630.3728535.
- [9] S. Liu, X. Li, Z. Liu, Y. Yan, C. Yang, Z. Zeng, Z. Liu, M. Sun, and G. Yu, "Judge as A Judge: Improving the Evaluation of Retrieval-Augmented Generation through the Judge-Consistency of Large Language Models," in *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, 2025, pp. 5788–5807, ISBN: 979-8-89176-256-5, doi:10.18653/v1/2025.findings-acl.301.
- [10] P. P. Mishra, K. P. Yeole, R. Keshavamurthy, M. B. Surana, and F. Sarayloo, "A Systematic Framework for Enterprise Knowledge Retrieval: Leveraging LLM-Generated Metadata to Enhance RAG Systems," *arXiv preprint arXiv:2512.05411*, 2025, doi:10.48550/arXiv.2512.05411.
- [11] P. Jiang, "A Comparative Study on Retrieval-Augmented Generation and Chain-of-Thought Applications for LLM-Assisted Engineering Design Ideation," *Journal of Engineering Design*, pp. 1–41, 2025, doi:10.1080/09544828.2025.2574209.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Each author made an equal contribution to
The current study, from the conceptualization of the problem to the final conclusions and resolution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0

https://creativecommons.org/licenses/by/4.0/deed.en_US