International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

Hybrid Gen AI Systems: Integrating Small LMs with Large Language Models for Cost-Efficient Enterprise Automation and Decision Intelligence

Siva Hemanth Kolla, Rajesh Mattaparthi

Gen AI Research Scientist, USA

Principal Data Engineer, USA

siva.kolla.hemanth@gmail.com

rajeshhmattaparthi@gmail.com

ABSTRACT: Integrating small language models with large language models addresses cost and decision-intelligence challenges in enterprise automation. The combination mitigates latency concerns while harnessing the semi-supervised accuracy of LLMs, achieving a lower total cost of ownership in typical Enterprise Generative AI scenarios—model hosting, inference, data transfer, maintenance—by leveraging small-model alternatives. Small language models exhibit high-performance inference capabilities; they efficiently execute simple tasks and process Benchmark data for fine-tuning or evaluation. Although users require low-latency responses, a hybrid setup with LLMs as bad-weather models enhances speed without sacrificing completeness. Exploration of Routing Rules ensures adequate fault containment and multiple Monitoring and Rollback Models enable configuration updates during live execution.

Scalable architectures, including dynamic resource scaling, task prioritization, data-caching strategies, and on-demand hardware, improve hybrid deployments. Coupled with cloud economics and energy-efficient edge serving, hybrid Generative – AI systems support a Cost-Effective Green Enterprise strategy. Decision-Intelligence implementations harness Candidate Signals across the Decision Matrix to focus on Explainable Decision Results. Data from various sources is fused into cohesive inputs, with Structured Data augmenting Unstructured Text via Schema-aware Feature Engineering, Context-driven Retrieval-Augmented Generation (RAG), and Semantic-scale Querying. A Robust Data Quality Pipeline validating Input Quality, Provenance, and Query Answering completes the solution.

Although enterprise data—text, audio, videos, and images, alone or in combination—is potentially exploitable across the Automation and Decision-Intelligence spectrum, the Adequacy Principle for Utilization requires an integrated Data-Governance Framework that ensures model utility and risk mitigation. GMLIG Questions EMC, ETL Logic, Proprietary Content Protection, Auditing for Model Bias, and Risk Management are key Data-Governance Principles that influence Design.

KEYWORDS: Hybrid Generative AI, Small Language Models (SLMs), Large Language Models (LLMs),Enterprise Automation, Decision Intelligence, Cost-Efficient AI Systems, AI Model Orchestration, Multi-Model AI Architecture, Intelligent Workflow Automation, Scalable Enterprise AI.

I. INTRODUCTION

Hybrid generative AI systems are public-domain architectures integrating small LMs with closed-source LLMs, improving cost efficiency and expanding enterprise use cases. Cohort-based deployment patterns segregate model pools across predictability and quality intervals, optimizing risk, forecasting, and generative tasks. Moreover, latency-sensitive processes leverage small LMs directly, while high-stakes requests undergo additional scrutiny. These principles reduce inference volume, operational maintenance, and total cost of ownership. Cost components—model hosting, inference, data transfer, and upkeep—are assessed, revealing substantial savings through hybridization. Enterprise-focused resource-allocation strategies, including dynamic scaling, prioritization, caching, on-demand loading, localized proximity, and energy management further enhance expense efficiency.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

Accelerating enterprise Digital Transformation (DT) hinges on balancing cost and time. Integrative Decision Intelligence (DI) and Automation combine Business Process Management with AI-supported decision formulation and execution. Sophisticated solutions employ closed-source.

Parameter	Small Language Models (SLMs)	Large Language Models (LLMs)
Model Size	Lightweight	Very Large
Inference Speed	Fast	Moderate to Slow
Cost of Deployment	Low	High
Energy Consumption	Lower	Higher
Latency	Minimal	Higher
Domain Adaptability	Strong for Specific Tasks	Broad Generalization
Infrastructure Needs	Limited Hardware	High-End GPU/TPU Clusters
Governance Complexity	Moderate	High
Hallucination Risk	Moderate	Moderate to High
Enterprise Usage	Real-Time Automation	Complex Decision Intelligence

Table 1: Comparison Between Small Language Models and Large Language Models

1.1. Data Governance and Compliance Considerations

The above governance elements and principles must be followed in hybrid models based on their deployment patterns and support the regulatory and compliance requirements of the enterprise. Common principles of data governance and compliance that are applicable for generative AI systems also apply for hybrid systems. Hybrid systems add further complexities that raise additional governance concerns or require more exhaustive considerations to mitigate risk. Key areas include effective and explainable control of data leakage and model hallucinations, validating and testing task outcomes by audit processes to ensure a regulatory or compliance required duty of care, provisioning with deep context to be able to give more accurate and relevant outcomes, and establishing risks associated with deployment of small models and any associated injection or poisoning risks. The responsibility of governance across the entire system life cycle—including development, professional service firms, and model training domain—also becomes critical.

II. FOUNDATIONS OF GENERATIVE AI IN ENTERPRISE

Foundations of Generative AI in Enterprise

Generative AI combines foundational models with knowledge conditioning to generate convincing output. Generative models approximate data distributions given a data corpus, enabling sampling of new content by imitating the training set. Foundation models are pre-trained on large corpora of text, image, or audio data across multiple tasks, later transformed into generative models. A second stage conditions the generative model through instruction-based prompting, reinforcement learning (RL), or retrieval augmentation. Prompt engineering directs the embedded knowledge toward usable outputs. Generative AI-governed content can appear authentic, useful, and complete, but production and governance challenges must be addressed.

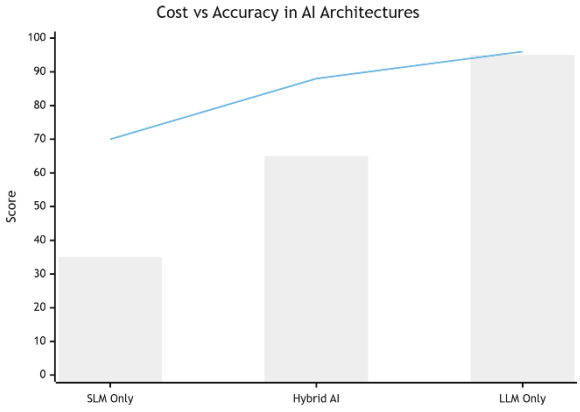
|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

Small language models (SLMs) fulfill the generative AI definition and can handle language-based tasks but incur higher latency due to limited conditioning and sampling speed. Relative performance, latencies, data footprints, and operating costs of SLMs and large language models (LLMs) account for their incorporation into automation and decision-intelligence use cases for a hybrid model and controller architecture. These flow data to services, conduct queries, and generate LM requests, serving primarily as folders. SLM performance maps onto data flow.

Hybrid Gen AI Routing

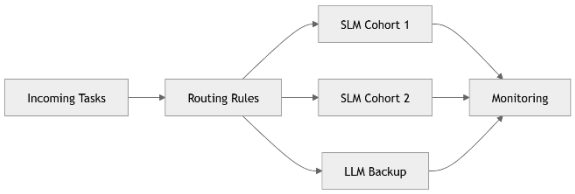


Hybrid AI Cost vs Accuracy Analysis

2.1. Small Language Models: Capabilities and Limitations

Small language models excel in concise instruction following and domain-specific reasoning. Performance evaluation focuses on hallucination rates, relevance, and helpfulness. Prominent small language models, capable of instruction adherence and domain specialization, include Bloomz, CEREBRO, Dolly, FLAN, LaMDA, Mistral, OPT, Phoenix-M, PyGodot, T0, T0pp, T5, T5.5, Ultra-MiniLM, Vicuna, and WizardLM. Fine-tuning costs, enhanced domain proficiency, and fast responsiveness remain founder priorities. Infrastructure, assembly speed, maintenance demand, and ongoing updates deserve equal consideration. While definitive evaluation criteria are still evolving, leading factors include hallucination rate, relevance, and helpfulness. Health information queries are being assessed with distinct metrics. The distinction between model-centric and task-centric benchmarks should also be recognized. The open-source community is currently devising suitable datasets and metrics for Imitation-Centric remarks.

A persistent concern surrounding small language models is their susceptibility to hallucinations, where generated text lacks truthfulness. However, these outages are not just a product of the model's size. Relational and commonsense hallucinations frequently impact larger models with less securing training procedures. Secondary leakage of sensitive training samples can emerge during the testing phase and may endanger brands and companies. Implementing an Imitation-Centric methodology, which aims to collectively fine-tune multiple models with a shared dataset, is resource intensive. Its thoroughness is expected to surpass prompt-tuning. The evolving adapt-reason-judge pipeline resembles the Three-Stage Activity Model used to explain problem-solving behavior in humans.



Cohort-Based Deployment

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

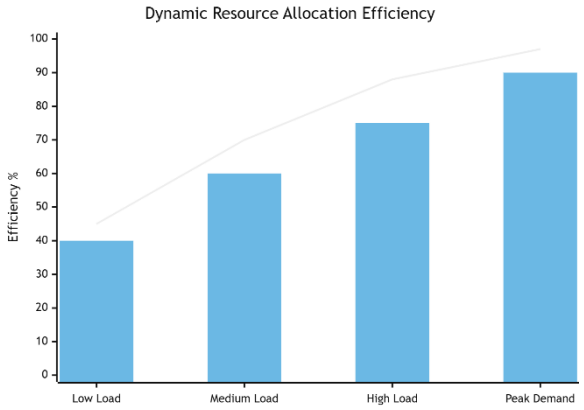
Architecture Pattern	Primary Function	Advantages	Limitations
Orchestrator	Routes tasks between SLMs and LLMs	Better task specialization	Increased coordination complexity
Adapter	Connects heterogeneous AI services	Easier integration	Additional middleware overhead
Proxy	Controls model access and governance	Enhanced security and compliance	Potential bottlenecks
Cohort-Based Deployment	Uses segregated model pools	Cost-efficient scaling	Requires monitoring logic
Tiered Processing	Assigns tasks based on complexity	Reduced inference costs	Added routing latency

Table 2: Hybrid Gen AI Architectural Patterns

III. HYBRID ARCHITECTURES FOR ENTERPRISE AUTOMATION

Architectural patterns—such as orchestrator, adapter, and proxy—that enable cost-efficient enterprise automation by connecting small LMs with LLMs are discussed. The data flow of Task 2 and Task 3 between the two model classes is specified, together with the associated trade-offs between inference latency and accuracy. The integration of governance aspects into hybrid models is also examined, followed by criteria for selecting the appropriate architectural pattern.

Hybrid Gen AI systems can employ three representative architectural patterns, namely orchestrator, adapter, and proxy. These patterns introduce coordination mechanisms between distinct yet complementary LM classes to derive cost-efficient solutions for enterprise needs across automation, augmentation, and ad-hoc use cases. An orchestrator permits the disassembly of prompting tasks into subtasks, thereby enabling specialized Model-as-a-Service (MaaS) implementation in segregated pools optimized for specific requirements. Within the context of Task 2, the internal flow of structured data from a small LM through the orchestrator to an external LLM is demonstrated. By routing distinct subtasks to their respective data reservoirs, the orchestrator minimizes the risk of hallucinations and facilitates real-time fault management for generating textual missions and evaluating resource alternatives, Therefore, a fully operational setup for this task combining human breadth with AI depth relies upon tiered infrastructure and routing rules.



Enterprise Resource Scaling Efficiency

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

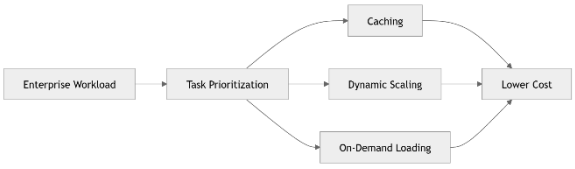
3.1. Cohort-based Deployment patterns

Cohort-based deployment of hybrid Gen AI systems uses segregated pools of small language models (SLMs) for different classes of tasks; applies tiered processing and routing rules; supports fault containment; and implements simple model-selection logic, monitoring, and rollback mechanisms. Smaller LMs are less expensive than larger LLMs, and cohort-based deployment trades off additional latency for reduced total cost of ownership while preserving sufficient resilience. Cost modelling reduces the cost of any deployment beyond a threshold by allocating additional resources.

Cohort-based deployment processes distinct types of requests independently through pools of specialized SLMs. Routing rules direct requests to the appropriate pool, and monitoring and rollback mechanisms address potential failures. When traffic levels remain low, a monolithic system incurs less cost than a hybrid with tiered processing; the opposite holds as load increases. Cohort-based deployment can therefore be particularly useful during rapid spikes in demand.

Cost Component	Description	Hybrid Optimization Strategy
Model Hosting	Infrastructure for running models	Dynamic resource scaling
Inference Cost	Runtime query execution cost	SLM-first routing
Data Transfer	Movement of enterprise data	Edge deployment & caching
Maintenance	Updates, monitoring, retraining	Automated rollback mechanisms
Energy Consumption	Cooling and hardware power	Energy-efficient serving
Scaling Infrastructure	Resource provisioning	Serverless/on-demand loading

Table 3: Enterprise Cost Components in Hybrid AI Systems



Cost-Efficient Automation

IV. COST-EFFICIENCY THROUGH HYBRIDIZATION

Cost-efficiency is a primary driver for hybridizing generative AI architectures. Successful adoption of hybrids requires careful selection of operating patterns, resource allocation and cloud infrastructure decisions. These factors influence the total cost of ownership (TCO) and capital versus operational expenditure trade-offs.

Cost components cover model hosting, inference, data transfer, and ongoing maintenance. TCO is compared across hybrid and monolithic approaches, revealing levers to reduce expense. Accelerating inference is a key concern, examined in the context of resource allocation (on-the-fly scaling, prioritization, caching, and loading) and hardware choices (on-premises versus cloud, dedicated versus general-purpose servers) with energy efficiency considerations. Decision intelligence capabilities in hybrid setups are also explored.

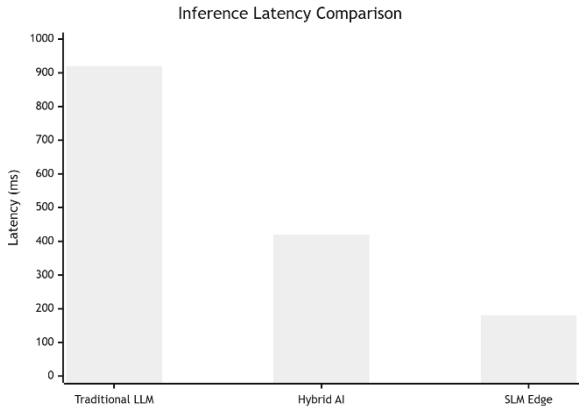
Hybrids do not lower TCO or accelerate response time for every enterprise deployment. Segregating small LMs into dedicated pools for similar content and usage patterns reduces model-selection overhead and organizes inference-

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

consuming tasks on interchangeable components. Deploying complex tasks on larger LLMs only when low-latency responses matter further lessens operating burden. Cohort-based deployment presents a tactical solution, employing segregated pools with tiered processing and routing rules to balance cost and risk containment. Resource metrics trigger scaling actions, while dedicated monitoring and rollback mechanisms guide model-selection logic.



Latency Reduction using Hybrid Models

Mathematical Formulas:

1. Hybrid Cost

$$C_H = C_{SLM} + C_{LLM} + C_R + C_M$$

2. Total Cost of Ownership

$$TCO = C_H + C_I + C_D + C_{Maint}$$

3. Cost Saving

$$S = \frac{C_{Mono} - C_H}{C_{Mono}} \times 100$$

4. Hybrid Latency

$$L_H = L_R + L_M + L_G$$

5. Routing Decision

$$R(x) = \begin{cases} SLM, & T(x) \le \tau \\ LLM, & T(x) > \tau \end{cases}$$

6. Model Selection Score

$$M^* = \arg\max (A - \lambda C - \mu L)$$

7. Accuracy–Cost Tradeoff

$$Q = \alpha A - \beta C - \gamma L$$

8. Hallucination Risk

$$H_R = 1 - P(T \mid C)$$

9. Decision Confidence

$$D_C = \frac{\sum w_i s_i}{\sum w_i}$$

10. RAG Answer Score

$$A_R = f(Q, K, Ctx)$$

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

11. Resource Scaling

$$R_s = \frac{D_t}{C_p}$$

12. Carbon Efficiency

$$E_c = \frac{Energy}{Requests}$$

13. Governance Risk

$$G_R = H_R + D_L + B_R$$

14. Data Quality Score

$$DQS = V + P + C$$

15. Hybrid Utility

$$U_H = A - (C + L + R)$$

4.1. Resource Allocation and Scaling Strategies

To mitigate cost, reduce resource consumption, and optimize quality-of-service within hybrid Gen AI deployments, multiple resource allocation strategies can be applied. In scenarios with sudden spikes of requests, dynamic scaling provides Compute-as-You-Need provisioning for hosting models, thereby preventing the wasteful and expensive constant provisioning of resources. Within dedicated cohorts, Inferencing-as-You-Need mechanisms optimize load distribution among models differentiated by accuracy, latency, or cost. ID-based routing rules, along with model-selection logic and monitoring mechanisms, further enhance cost efficiency. Latency-sensitive cohorts may also minimize running costs by caching responses to frequently occurring queries. Where speed is less critical, on-demand model loading and execution models improve quality-of-experience by preferring higher-performing models. Finally, choices regarding the underlying model hosting infrastructure—whether on-prem, in the cloud, or a combination of both—should also seek to minimize the carbon footprint per request as much as possible.

Strategy	Objective	Enterprise Benefit
Dynamic Scaling	Allocate compute during spikes	Reduced operational cost
Query Caching	Store frequent responses	Faster response time
On-Demand Model Loading	Load models only when required	Lower memory usage
Dedicated Cohorts	Specialized model pools	Better performance efficiency
Cloud Bursting	Shift workloads to cloud	Improved scalability
TPU-Based Acceleration	Hardware optimization	Lower latency

Table 4: Resource Allocation and Scaling Strategies

Decision Intelligence Flow

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

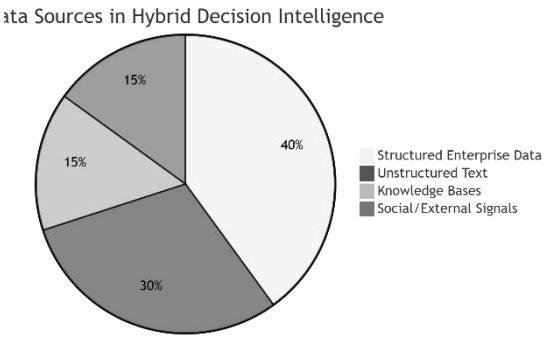
||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

V. DECISION INTELLIGENCE IN HYBRIDS

Decision Intelligence in Hybrids

Decision Intelligence (DI)9 enables organizations of all sizes to hone their decision-making capabilities – a crucial differentiator for success and a massive opportunity for positive impact. Hybrid Generative AI systems can aid DI in three distinct areas. First, generative capabilities allow for the integration of multiple incoming signals, with hybrid structures being uniquely positioned to connect structured and unstructured, digital, and human data sources. Second, sensitivity analysis can be provided to help users understand the uncertainty of the output and the likelihood of success, allowing users to make decisions with a better understanding of the outcome's reliability. Finally, hybrid systems can also help generate decision-ready insights and foster higher-order thinking: informing decisions, outlining action plans, suggesting follow-up questions, and providing caveats or considerations.



Structured + Unstructured Data Fusion

5.1. Integrating Structured Data and Unstructured Text

Structured decision-support signals can be complemented with Generative Aid-relevant text inputs from Enterprise Sources, Third-party Sources (Social Media, News), Knowledge Bases (Fact Sources) and the User. Combining these non-structured signals remains a challenge and involves four key capabilities—Data Schema Matchmaking, Feature Engineering, RAG and Ontology Alignment—discussed below.

Data Schema Matchmaking

Any data signal from the organization's internal systems has a defined schema (XSD or XSL) used to enforce the structure at the API endpoints. These APIs define two key aspects—what data needs to be pushed from a system to be ingested by other systems and how data quality is maintained across systems. These schemas are used to derive feature vectors for every user request interrogating multiple enterprise systems. User requests must be mapped to these schemas to avoid incorrect usage of data and details being missed while answering the user queries.

Feature Engineering

Once a user query is associated with a valid API Schema, that schema's routing logic is followed to retrieve data from relevant systems. Generic features required for every structured data call are loaded prior to the data retrieval—either via caching or computation, depending on the usage patterns. The structured feature vector with other external structured data (e.g., Knowledge Base and Signal Data) is then used by the Generative Aid model to provide a well-rounded answer to the user query.

Retrieval-Augmented Generation

In retrieval-augmented generation (RAG), a retrieval component is used to find contextually relevant passages from a large body of Documents, Articles, FAQs and Extracts from Knowledge Bases for a given user query before these passages are provided as context to an LLM for answer generation. Providing External Context helps fill the LLM hallucination gaps as these passages provide the knowledge that the model lacks. RAG models can also be used to provide structured data attributes of the signal data and current context to or enhance the answer generated by the Generative Aid model.

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

Ontology Alignment
Enterprise Ontologies defined informally in the organization across business functions can be used to enrich LLM Inputs and Outputs. When integrated, these Ontologies help Improvise the Query By Eliminating Model Mechanisms That Might Generate Wrong Answers.

DI Component	Function	Outcome
Signal Integration	Combines structured and unstructured inputs	Better situational awareness
Sensitivity Analysis	Evaluates uncertainty levels	Improved risk assessment
Action Recommendation	Generates next-step suggestions	Faster decision-making
Ensemble Methods	Combines multiple model outputs	Increased reliability
Benchmarking	Compares alternative decisions	Optimized business actions
Explainability Layer	Provides reasoning transparency	Regulatory compliance

Table 5: Decision Intelligence Components in Hybrid AI

Governance Pipeline

VI. CONCLUSION

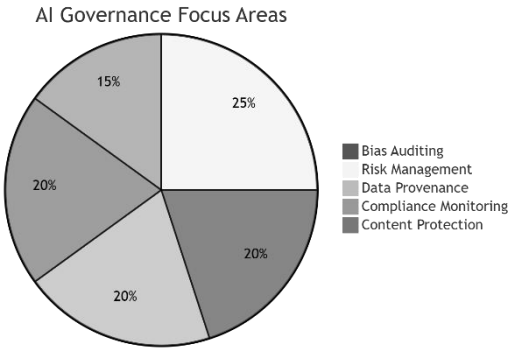
Current interest in generative AI revolves around the promise of automating decision-relevant tasks associated with enterprise decision intelligence—especially by capitalizing on the emergence of large language models. A study of the practical implementation of these technologies highlights opportunities for integrating small language models with large language models to address cost challenges of training and inferences too, while also accelerating the delivery of enterprise automation. Such hybrids address both front- and back-office applications, ranging across unstructured and structured data—including natural language and structured data such as tables, images, and graphs.

A multi-dimensional analysis of cost factors related to infrastructure deployment, loading and inference, data transfer, and ongoing maintenance demonstrates that a hybrid approach can both lower trade-off options when more than one model deployment is considered and reduce overall cost of ownership. Resource allocation and scaling strategies further improve cost-efficiency. Thus, small language models can assist large language models in a hybrid setup to optimize total cost of deployment, utilization, and ongoing maintenance—both for a set of task-oriented cohorts and within an infrastructure aligned with best-practice considerations for enterprise governance and compliance.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038



Governance & Risk Mitigation Metrics

REFERENCES

1. Nuka, S. T., Chakilam, C., Chava, K., Suura, S. R., & Recharla, M. (2025). AI-driven drug discovery: transforming neurological and neurodegenerative disease treatment through bioinformatics and genomic research. American Journal of Psychiatric Rehabilitation, 28(1), 124-135.
2. Pandiri, L. (2025, May). Exploring Cross-Sector Innovation in Intelligent Transport Systems, Digitally Enabled Housing Finance, and Tech-Driven Risk Solutions A Multidisciplinary Approach to Sustainable Infrastructure, Urban Equity, and Financial Resilience. In 2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) (pp. 1-12). IEEE.
3. Challa, S. R., Burugulla, J. K. R., Pamisetty, A., Challa, K., & Paleti, S. (2025, April). AI and ML-Powered Cybersecurity Strategies for Cloud Computing: Ensuring Infrastructure Stability in Financial and Retail Sectors. In International Conference on Smart Computing and Informatics (pp. 315-327). Cham: Springer Nature Switzerland.
4. Rani, P. S., Amistapuram, K., Pamisetty, V., Singireddy, S., Kummari, D. N., & Sheelam, G. K. (2025, November). Hybrid Knowledge Graph–Deep Learning Framework for Automated Exception Handling and Investigation in Complex Insurance Claims. In 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN) (pp. 1-6). IEEE.
5. Seenu, A., Aitha, A. R., Gottimukkala, V. R. R., Singireddy, J., Meda, R., & Garapati, R. S. (2025, November). Hybrid Multi-Agent Reinforcement Learning and Blockchain Framework for Real-Time Transaction Integrity in Cloud-Driven Financial Systems. In 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN) (pp. 1-6). IEEE.
6. Singireddy, S. (2024). The Integration of AI and Machine Learning in Transforming Underwriting and Risk Assessment Across Personal and Commercial Insurance Lines. Journal of Computational Analy- sis and Applications(JoCAAA), 33(08), 3966-3991.
7. Kannan, S., & Yellanki, S. K. (2025). Synthetic Cognition Meets Data Deluge: Architecting Agentic AI Models for Self-Regulating Knowledge Graphs in Heterogeneous Data Warehousing.
8. Sheelam, G. K. (2025). Deploying Neural-Symbolic Hybrid Models for Adaptive Spectrum Management in 6G-Ready Networks. Journal of Neonatal Surgery, 14(22s).
9. Kolla, S. K. (2024). Federated Machine Learning On Big Healthcare Data For Privacy-Preserving Analytics. The Review of Diabetic Studies, 175-190.
10. Mangalampalli, B. M. (2024). AI-Enhanced Data Governance: Automating Compliance In Healthcare Analytics Platforms. The Review of Diabetic Studies, 191-204.
11. Srikanth, T., Segireddy, A. R., & Elavarasi, S. A. (2025, October). STaSFormer-SGAD: Semantic Triplet-Aware Spatial Flow-Guided Spatio-Temporal Graph for Anomaly Detection in Surveillance Videos. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-7). IEEE.
12. Loganathan, R. (2024). GENERATIVE AI-ENABLED COMPLIANCE DOCUMENTATION AND AUDIT TRAIL AUTOMATION FOR GLOBAL DATA CENTER GOVERNANCE. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 487–504. https://doi.org/10.61841/turcomat.v15i3.15512
13. Mangala, N. (2022). Real-Time Data Quality Monitoring and Gating Frameworks in Cloud-Based Data Pipelines. International Journal of Research and Applied Innovations, 5(6), 8197-8219.
14. Davuluri, P. S. L. N. . (2024). AI-Driven Data Governance Frameworks for Automated Regulatory Reporting and Audit Readiness. Metallurgical and Materials Engineering, 30(4), 996–1010. https://doi.org/10.63278/mme.v30i4.1936
15. Yandamuri, U. S. AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

16. Ashokkumar, S., & Amistapuram, K. (2025, October). Attention-Guided Spatial Temporal Framework for Deepfake Detection on Social Video Platforms. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.
17. Gottimukkala, V. R. R. (2025). Generative AI for Exceptions and Investigations: Streamlining Resolution Across Global Payment Systems. Journal of International Commercial Law and Technology, 6(1), 969-972.
18. Sivanand, R., Kumar, D. P., Nagabhyru, K. C., Natarajan, E. P., Pamisetty, V., & Kapila, D. (2025, September). IoT and AI for Real-Time Monitoring in Substation Automation. In 2025 International Conference on Computing and Communications (COMPUTINGCON) (pp. 1-5). IEEE.
19. Agrawal, S., Kumar, S. N., Singh, D. K., Niharika, D. S., Nandan, B. P., & Asati, D. (2025, December). Dynamic Access Management and Authentication Mechanisms for Enhancing 5G Security Against Heterogeneous Adversaries. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.
20. Alshar, M. M., Shahdadpuri, N., Rajeshwari, M., Gupta, M., Joshi, N. R., & Singireddy, J. (2025, October). Enhanced Management & Performance of Remote Workforce with Cloud and AI-Driven HR Analytics. In 2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (Vol. 1, pp. 631-636). IEEE.
21. Meda, R. (2025). AI-Driven Demand and Supply Forecasting Models for Enhanced Sales Performance Management: A Case Study of a Four-Zone Structure in the United States. Metallurgical and Materials Engineering, 1480-1500.
22. Kalisetty, S., & Inala, R. (2025). Designing Scalable Data Product Architectures With Agentic AI And ML: A Cross-Industry Study Of Cloud-Enabled Intelligence In Supply Chain, Insurance, Retail, Manufacturing, And Financial Services. Metallurgical and Materials Engineering, 86-98.
23. Garapati, R. S. (2025). An Intelligent IoT Security System: Cloud-Native Architecture with Real-Time AI Threat Detection and Web Visualization. Journal homepage: https://jmsronline. com, 2(06).
24. Radhakrishnan, P., Nagabhyru, K. C., Manonmani, C., Srinu, M., Kaur, H., & Nandhini, N. (2025, October). K-Means-KNN Hybrid Model for Efficient Intrusion Detection in Cloud-based IoT Systems. In 2025 10th International Conference on Communication and Electronics Systems (ICCES) (pp. 1583-1588). IEEE.
25. Amistapuram, K. (2025). GENERATIVE AI FOR CLAIMS EXCEPTIONS AND INVESTIGATIONS: ENHANCING RESOLUTION EFFICIENCY IN COMPLEX INSURANCE PROCESSES. Available at SSRN 5785482.
26. Kolla, T. (2025). The Future of Healthcare Analytics: Leveraging AI and Data Engineering for Personalized Medicine. Journal of Computer Science and Technology Studies, 7(4), 634-640.
27. FinOps Strategies for AI-Enabled Real-Time Compliance Platforms in Cloud Native Environments. (2025). MSW Management Journal, 35(2), 2080-2088.
28. Pote[1], X. R., Pamisetty, A., Karthikeyan, G., & Gupta[1], D. (2025, May). Artificial Intelligence Enabled Smart Energy Conservation Systems for Intelligent Resource Management and Sustainable Future Power Grids. In Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24) (p. 196). Springer Nature.
29. Seenu, A., Sheelam, G. K., Motamary, S., Meda, R., Koppolu, H. K. R., & Inala, R. (2025, July). AI-Driven Innovations in Infrastructure Management with 6G Technology. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.
30. Singreddy, S. (2024). Predictive Modeling for Auto Insurance Risk Assessment Using Machine Learning Algorithms. Available at SSRN 5238922.
31. Ranjith Kumar Peddi (2021). Optimizing Case Management Workflows in Global Data Center Colocation Services. Universal Journal of Computer Sciences and Communications, 1(1), 1-21. https://doi.org/10.31586/ujscs.2021.1380
32. Bandi, V. D. V. K. (2025). Self-Optimizing Data Pipelines Using Machine Learning for Cloud Workloads. Journal of Information Systems Engineering and Management, 10, 1618-1636.
33. Enterprise-Scale Gen AI Orchestration Using Small LMs and LLM Agents for Intelligent ITSM and HRSD Automation in Enterprise Ecosystems. (2025). MSW Management Journal, 35(2), 1889-1897.
34. Nagubandi, A. R. (2025). Cryptocurrency Market Spillovers: Risk Contagion Across Global Financial Systems.
35. Gottimukkala, V. R. R. (2025). Agentic AI for Next-Generation Cross-Border Payments: Contextual Learning in Transaction Routing. Journal of Informatics Education and Research, 5(4).
36. Thutari, R. T., Garapati, R. S., BM, M., & RK, S. (2025, October). Adaptive Access Control and Authentication Management for IoT Using Attention-GRU and Reinforcement Learning. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-6). IEEE.
37. Kolla, S. K. (2021). Designing Scalable Healthcare Data Pipelines for Multi-Hospital Networks. World Journal of Clinical Medicine Research, 1(1), 1-14.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

38. Baliyan, M., Balakrishnan, S., Mohammed, S., & Nagubandi, A. R. (2025). Financial and Management Accounting. BR Publications.
39. MANGALAMPALLI, B. M., KOLLA, S. H., APPA RAO NAGUBANDI, D. R., & SEGIREDDY, A. R. (2025). AN INTELLIGENT, REAL-TIME DIGITAL FABRIC FOR HEALTHCARE AND FINANCIAL ECOSYSTEMS USING AUTONOMOUS LEARNING AND GENERATIVE SYSTEMS. TPM–Testing, Psychometrics, Methodology in Applied Psychology, 32(S9 (2025): Posted 15 December), 3070-3086.
40. Mangalampalli, B. M. Generative AI Applications In Healthcare Data Mart Design And Optimization.
41. Ranga Reddy, V. A. (2024). Comparing Batch vs. Streaming Approaches in Healthcare Data Warehousing Environments. Journal of Neonatal Surgery, 13(1), 2287–2309. Retrieved from https://www.jneonatalsurg.com/index.php/jns/article/view/10223
42. Mangala, N. (2025). Agentic Data Pipelines: Autonomous ELT Orchestration Using AI Agents on Microsoft Fabric and Databricks. International Journal of Computer Technology and Electronics Communication, 8(6), 11891-11907.
43. Venkata Akhilesh Ranga Reddy (2022). Designing Fault-Tolerant Data Ingestion Pipelines for High-Volume Healthcare Transactions. Frontiers in Health Informatics, Vol.11(2022), 861-889
44. Amistapuram, K., Pandiri, L., Raju, V. R., Paleti, S., Singireddy, S., & Sheelam, G. K. (2025, December). AI-Based Cloud Infrastructure and MLOps Frameworks for Scalable Data Engineering Across Banking and Insurance. In 2025 IEEE International Conference on Communication Networks and Computing (CNC) (pp. 186-192). IEEE.
45. Recharla, M., & Nuka, S. T. (2025). Translational Approaches To Commercializing Neurodegenerative Therapies: Bridging Laboratory Research With Clinical Practice. South Eastern European Journal of Public Health, 121–144.
46. Kumar, S. S., Singireddy, S., Nanan, B. P., Recharla, M., Gadi, A. L., & Paleti, S. (2025). Optimizing edge computing for big data processing in smart cities. Metallurgical and Materials Engineering, 31(3), 31-39.
47. Kummari, D. N., Burugulla, J. K. R., Malempati, M., Amistapuram, K., Garapati, R. S., & Nagabhyru, K. C. (2025, December). Enhancing Audit Compliance and Operational Efficiency in Manufacturing and Commercial Insurance Through Agentic AI and Data Engineering Frameworks. In 2025 IEEE International Conference on Communication Networks and Computing (CNC) (pp. 714-720). IEEE.
48. Singh, D., Meda, R., & Kumar, V. (2025). Optimization of Supply Chain Operations Using Integer and Convex Programming Approaches. Advances in Consumer Research, 2(6).
49. Aitha, A. R. (2024). Generative AI-Powered Fraud Detection in Workers' Compensation: A DevOps-Based Multi-Cloud Architecture Leveraging, Deep Learning, and Explainable AI. Deep Learning, and Explainable AI (July 26, 2024).
50. Inala, R., & Somu, B. (2025). Building trustworthy agentic AI systems for personalized banking experiences. Metallurgical and Materials Engineering, 31(5), 1336-1360.
51. Vajpayee, A., Khan, S., Gottimukkala, V. R. R., Sharma, D., & Seshasai, S. J. (2025). Digital Financial Literacy 4.0: Consumer Readiness for AI-Driven Fintech and Blockchain Ecosystems. International Insurance Law Review, 33(S5), 963-973.
52. Nigam, N., Sireesha, B., Ediga, P., Segireddy, A. R., & Bokde, S. (2025, December). Comparative Evaluation of Cloud Security Algorithms Using Multiple Classifiers with an Optimized Intrusion Detection System. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.
53. Ranjith Kumar Peddi. (2024). AI-Based Workforce Analytics for SLA Governance and Uptime Assurance in Data Centers. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 8589–8601. Retrieved from https://eudoxuspress.com/index.php/pub/article/view/5361
54. AGENTIC AI FRAMEWORKS FOR AUTONOMOUS RISK DETECTION AND COMPLIANCE REMEDIATION IN ENTERPRISE DATA CENTER OPERATIONS. (2025). Lex Localis - Journal of Local Self-Government, 23(S6), 9672-9697. https://doi.org/10.52152/3f90ak91
55. Chakraborty, S., Pamisetty, A., Chandana, N., & CS, B. (2025, October). Depth-Wise Temporal Convolutional Networks with Layer Normalization for Waste Food Prediction. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-6). IEEE.
56. Kummari, D. N., Challa, S. R., Pamisetty, V., Motamary, S., & Meda, R. (2025). Unifying Temporal Reasoning and Agentic Machine Learning: A Framework for Proactive Fault Detection in Dynamic, Data-Intensive Environments. Metallurgical and Materials Engineering, 31(4), 552-568.
57. Pandiri, L. (2025). The Complete Compendium of Digital Insurance Solutions: Life, Health, Auto, Property, and Specialized Coverage in the Age of AI, Automation, and Intelligent Risk Management. Deep Science Publishing.
58. Kumar, B. H., Nuka, S. T., Recharla, M., Chakilam, C., Suura, S. R., & Pandugula, C. (2025, July). Addressing Ethical Challenges in AI-Driven Health Predictions. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 6, November-December 2025||

DOI:10.15662/IJRPETM.2025.0806038

59. Krishnan, M., Aitha, A. R., Amistapuram, K., Nandan, B. P., Kaulwar, P. K., & Singireddy, J. (2025, November). Human-in-the-Loop Hybrid Neuro-Symbolic AI Model for Reliable Data Engineering in High-Stakes Industrial Systems. In 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN) (pp. 1-7). IEEE.
60. Sanku, R., Singireddy, J., Ilakkia, T., Kamala, N., & Soni, M. (2025, October). Comprehensive Analysis on Energy Efficient Transmission in Wireless Sensor Network. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-8). IEEE.
61. Singireddy, S. (2025, May). AI-Driven Comprehensive Insurance and AAA Membership Benefits Overview. In 2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) (pp. 1-13). IEEE.
62. Ramana, B., Sheelam, G. K., Pandya, T., Rai, A. K., Kumar, V. A., & Kukreti, A. (2025, December). Exploring the Potential of NOMA in 6G Through Comparative Analysis with OMA Techniques. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.
63. Rani, P. S., Kummari, D. N., Yellanki, S. K., Meda, R., Koppolu, H. K. R., & Inala, R. (2025, July). Blockchain and AI for Securing Electrical Infrastructure. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.
64. Somu, B., & Inala, R. (2025). Transforming Core Banking Infrastructure with Agentic AI: A New Paradigm for Autonomous Financial Services. Advances in Consumer Research, 2(4).
65. Garapati, R. S. (2025). Artificial Intelligence-based systems, Cloud computing, Web interfaces, IoT/Connected devices, Smart automation, Real-time monitoring. Deep Science Publishing.
66. Pallapu, S. R., Aitha, A. R., Vandhana, K., & Chelladurai, S. (2025, October). GAN-Augmented Transformer Framework for Cross-Domain Video Style Transfer. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.
67. Kumar, I., Nagabhyru, K. C., IG, N., MV, P., & KV, S. (2025, October). Adaptive Meta-Knowledge Transfer Network with Feature Hallucination and Attention for Low-Shot Object Detection in Aerial Images. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.
68. Segireddy, A. R. (2025). Generative Ai For Secure Release Engineering In Global Payment Network. Lex Localis: Journal of Local Self-Government, 23.
69. Amistapuram, K. (2025). Agentic AI for Next-Generation Insurance Platforms: Autonomous Decision-Making in Claims and Policy Servicing. Journal of Marketing & Social Research, 2, 88-103.
70. Kolla, S. H. (2024). Retrieval-Augmented Generation With Small Llms For Knowledge-Driven Decision Automation In Enterprise Service Platforms. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 476-486.
71. Velangani Divya Vardhan Kumar Bandi. (2024). Intelligent Data Platforms For Personalized Retail Analytics At Scale. Metallurgical and Materials Engineering, 30(4), 1011–1027. https://doi.org/10.63278/mme.v30i4.1938
72. Mangalampalli, B. M., Kolla, S. K., Bandi, V. D. V. K., Yandamuri, U. S., & Rani, P. S. (2025). Designing Intelligent Healthcare Ecosystems through Adaptive Data Integration and Autonomous Learning Systems. Vascular and Endovascular Review, 8(20s), 330-347.
73. Kolla, T. (2024). AI-Powered Data Catalog Systems For Healthcare Data Discovery And Governance. South Eastern European Journal of Public Health, 2296–2311. https://doi.org/10.70135/seejph.vi.7077