
# Hybrid AI: Small and Large Models for Enterprise Automation

**Sophie Dubois**
Asst Professor

## 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.

## 1. 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

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

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.

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.

## 2. Foundations of Generative AI in Enterprise

Foundations of Generative AI in Enterprise

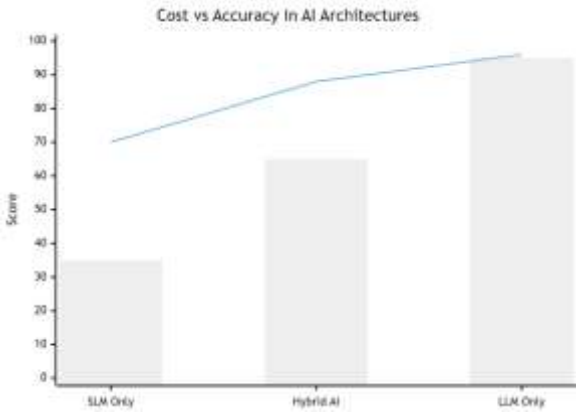
Generative AI combines foundational models with knowledge conditioning to generate convincing output. Generative models approximate data distributions given a data corpus, enabling sampling of new content by imitating the training set. Foundation models are pre-trained on large corpora of text, image, or audio data across multiple tasks, later transformed into generative models. A second stage conditions the generative model through instruction-based

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

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.

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.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

**Cohort-Based Deployment**

| 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 |

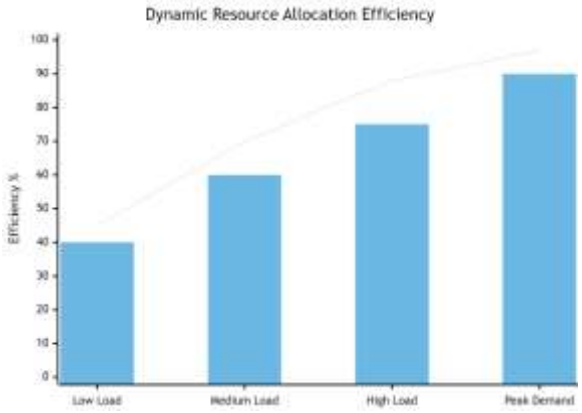
**Table 2: Hybrid Gen AI Architectural Patterns**

## 3. 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.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09




**Enterprise Resource Scaling Efficiency**

### 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**

## 4. Cost-Efficiency through Hybridization

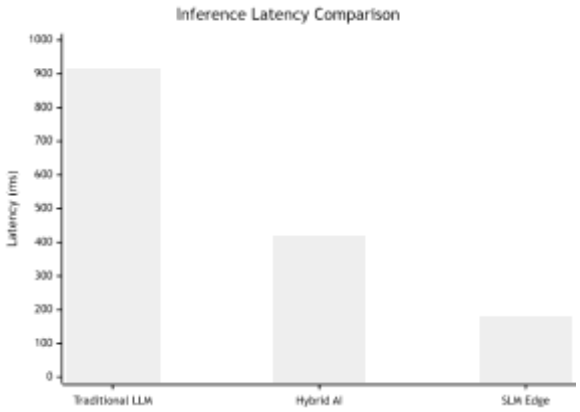
Cost-efficiency is a primary driver for hybridizing generative AI architectures. Successful adoption of hybrids requires careful selection of operating patterns, resource allocation and

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

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-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

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

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 |

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

| Strategy | Objective | Enterprise Benefit |
|---|---|---|
| Dedicated Cohorts | Specialized model pools | Better performance efficiency |
| Cloud Bursting | Shift workloads to cloud | Improved scalability |
| TPU-Based Acceleration | Hardware optimization | Lower latency |

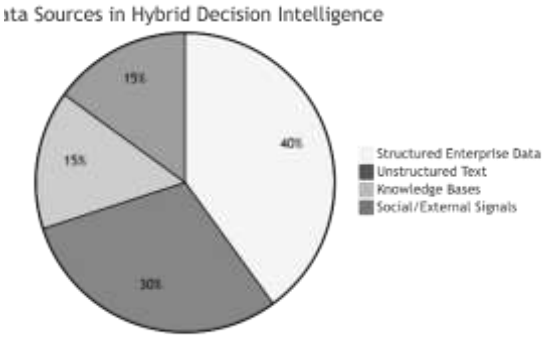
**Table 4: Resource Allocation and Scaling Strategies**


**Decision Intelligence Flow**

## 5. 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.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

**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.

**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**

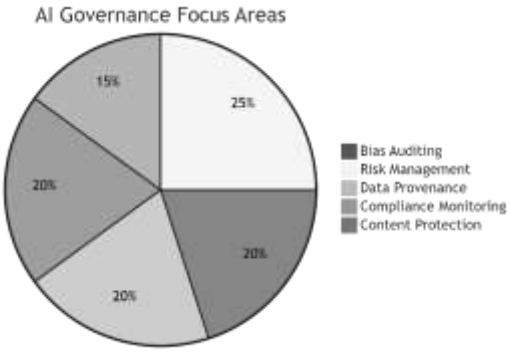
## 6. Conclusion

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

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

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.




**Governance & Risk Mitigation Metrics**

## References

1. Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems: A bibliographic review and open research issues. Electronic Markets, 31(3), 643–670.
2. Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine learning for industrial applications: A comprehensive literature review. Expert Systems with Applications, 175, 114820.
3. Reddy, V. A. R. (2025). Journal of Rare Cardiovascular Diseases. Health, 5(3), 402-422.
4. He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems, 212, 106622.
5. John, M. M., Holmström Olsson, H., & Bosch, J. (2021). Architecting AI deployment: A systematic review of state-of-the-art and state-of-practice literature. In Software Business (Lecture Notes in Business Information Processing, Vol. 421, pp. 14–29). Springer.
6. Min, B., Ross, H., Sulem, E., Ben Veyseh, A. P., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys.
7. Radha, S., Gottimukkala, V. R. R., Thottara, S., Vandhana, K., & J, Gokulraj. (2025). Adaptive Video Streaming Over 5G Networks Using Deep Reinforcement Learning with Closed-Loop Feedback Mechanism for Bitrate Control. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). https://doi.org/10.1109/ic3it66137.2025.11341184
8. Ng, K. K. H., Chen, C.-H., Lee, C. K. M., Jiao, J., & Yang, Z.-X. (2021). A systematic literature review on intelligent automation: Aligning concepts from theory, practice, and future perspectives. Advanced Engineering Informatics, 47, 101246.
9. Price, R., Mehrabani, M., Gupta, N., Kim, Y.-J., Jalalvand, S., Chen, M., & Black, A. W. (2021). A hybrid approach to scalable and robust spoken language understanding in enterprise virtual agents. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers (pp. 63–71).
10. Mangalampalli, B. M., Bandi, V. D. V. K., Kolla, S. K., & Kumar, M. V. K. (2025). Towards Self-Evolving Healthcare Intelligence: Integrating Advanced Learning Systems with Real-Time Clinical Data Pipelines.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH

ISSN : 3049-2122

VOLUME: 03 ISSUE: 04

RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

Cultura: International Journal of Philosophy of Culture and Axiology, 22(12s), 464-486.

11. van de Wetering, R., Mikalef, P., & Helms, R. (2022). Artificial intelligence and business value: A literature review. Information Systems Frontiers, 24(5), 1709–1734.
12. Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., Dennehy, D., Metri, B., Buhalis, D., Cheung, C. M. K., Conboy, K., Doyle, R., Dubey, R., Dutot, V., Felix, R., Goyal, D. P., Gustafsson, A., Hinsch, C., Jebabli, I., … Wamba, S. F. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.
13. Sudhakar, A. V. V., Inala, R., Verma, A. K., Nag, K., Pandey, V., & Anand, P. S. (2025, September). Hybrid Rule-Based and Machine Learning Framework for Embedding Anti-Discrimination Law in Automated Decision Systems. In 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT) (pp. 1-6). IEEE.
14. Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business & Information Systems Engineering, 66(1), 111–126.
15. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, P., … Wen, J.-R. (2023). A survey of large language models. arXiv.
16. Mangalampalli, Bindu Madhavi, Sasi Kumar Kolla, Velangani Divya Vardhan Kumar Bandi, Uday Surendra Yandamuri, and PR Sudha Rani. "Designing intelligent healthcare ecosystems through adaptive data integration and autonomous learning systems." Vascular and Endovascular Review 8, no. 20s (2025): 330-347.
17. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Yin, Z., … Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv.
18. Fan, A., Gokkaya, B., Harman, M., Lyubarskiy, M., Sengupta, S., Yoo, S., & Zhang, J. M. (2023). Large language models for software engineering: Survey and open problems. arXiv.
19. Inala, R., Kaulwar, P. K., Nagabhyru, K. C., Adusupalli, B., & Arun Raj, S. R. (2025, October). Leveraging IEC 61850 for Interoperable and Resilient Smart Grid Communication Architecture. In International Conference on Microelectronics, Electromagnetics and Telecommunication (pp. 549-566). Cham: Springer Nature Switzerland.
20. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., & Liang, P. (2023). On the opportunities and risks of foundation models. ACM Computing Surveys.
21. Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., Scialom, T., & others. (2023). Augmented language models: A survey. Transactions of the Association for Computational Linguistics.
22. Nabende, P., & Wanyama, T. (2008). An expert system for diagnosing heavy-duty diesel engine faults. In Advances in computer and information sciences and engineering (pp. 384-389). Dordrecht: Springer Netherlands.
23. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv.
24. Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI in organizations: Opportunities and challenges. Business & Information Systems Engineering, 66(1), 111–126.
25. 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.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

26. Hou, B., Zhang, J., Wang, H., Li, X., & Chen, Y. (2024). A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. Electronics, 14(18), 3580.
27. Mandadi, A. (2025). Fine-tuning vs. RAG vs. hybrid approaches for enterprise knowledge tasks: A systematic study. Journal of Information Technology and Applications Research, 1(1).
28. Kolla, S. H., & Mattaparthi, R. (2025). Hybrid Gen AI systems: Integrating small LMs with large language models for cost-efficient enterprise automation and decision intelligence. International Journal of Research Publications in Engineering, Technology and Management, 8(6).
29. Wang, S., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. arXiv.
30. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv.
31. Kolla, T. (2025). Generative AI for Intelligent Medical Coding and Healthcare Analytics. International Journal of Advanced Research in Computer Science & Technology (IJARCST), 8(6), 13285-13299.
32. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. L., Bressand, F., Lengyel, P., Lample, G., Saulnier, L., Lavaud, L., Lachaux, M.-A., Stock, P., Scao, T. L., Fan, A., & others. (2023). Mistral 7B. arXiv.
33. Team Gemini. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv.
34. OpenAI. (2025). GPT-4.1 technical report. OpenAI Research.
35. Li, Y., Li, S., Wang, Z., Ding, Y., & Chen, W. (2023). ChatGPT in artificial intelligence: A comprehensive review on technical foundations, applications, limitations, and future directions. ACM Computing Surveys.
36. 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.
37. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeiffer, F., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.
38. Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R., et al. (2023). Challenges and applications of large language models. arXiv.
39. 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.
40. Pan, S. J., Yang, Q., & colleagues. (2022). Transfer learning in the era of foundation models: Recent advances and future directions. IEEE Transactions on Knowledge and Data Engineering.
41. Mialon, G., Dessì, R., Lomeli, M., Scialom, T., & Wolf, T. (2023). Augmented language models: A survey. Transactions of the Association for Computational Linguistics, 11, 1017–1035.
42. Amistapuram, K., Pandiri, L., Raju, V. R., Paleti, S., Singireddy, S., & Sheelam, G. K. (2025). 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.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

186–192). IEEE. 2025 IEEE International Conference on Communication Networks and Computing (CNC). https://doi.org/10.1109/cnc68716.2025.11484532

43. Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open-domain question answering dataset from medical examinations. Applied Sciences, 11(14), 6421.
44. Yang, L., Yu, Z., Lim, H., & others. (2024). Efficient small language models for edge intelligence: A survey. IEEE Access, 12, 45672–45698.
45. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Wang, S., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations.
46. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36.
47. Loganathan, R. (2025). AGENTIC AI FRAMEWORKS FOR AUTONOMOUS RISK DETECTION AND COMPLIANCE REMEDIATION IN ENTERPRISE DATA CENTER OPERATIONS. Lex Localis-Journal of Local Self-Government, 23 (S6), 9672–9697.
48. Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv.
49. Lin, J., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the Association for Computational Linguistics, 10, 3214–3252.
50. 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.
51. Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Grand, G., & others. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36.
52. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations.
53. Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 37.
54. Wang, L., Ma, C., Feng, W., Zhang, Y., Liu, H., & others. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6).
55. 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.
56. Xi, Z., Chen, W., Guo, X., Yu, W., & Gui, T. (2023). Large language model based agents: A survey. arXiv.
57. Qin, Y., Yang, A., Zhu, B., Wang, Z., Li, C., et al. (2023). Tool learning with foundation models: A survey. arXiv.
58. Wang, P., Zhang, T., Liu, Y., & Sun, J. (2024). Small language models for edge computing: Recent advances and future opportunities. IEEE Internet of Things Journal.
59. Amistapuram[1], K., Kolla, T., Bandi, V. D. V. K., Kolla[4], S. K., & Rani, P. S. Journal of Rare Cardiovascular Diseases.
60. Min, S., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 11048–11064.
61. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35.
62. OpenAI. (2023). GPT-4 technical report. arXiv.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

63. Kolla, S. H., & Mangala, N. (2025). DESIGNING AUTONOMOUS LLM AGENT FRAMEWORKS USING GEN AI PIPELINES TO ENHANCE CUSTOMER SERVICE MANAGEMENT AND KNOWLEDGE WORKFLOWS. Lex Localis-Journal of Local Self-Government, 23, 9719-9733.
64. Team, A. (2024). The Llama 3 herd of models. arXiv.
65. Anthropic. (2024). The Claude 3 model family: Technical report. arXiv.
66. Google DeepMind. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv.
67. Vaswani, A., Narasimhan, K., & collaborators. (2024). Scaling transformer-based foundation models for enterprise AI applications. IEEE Intelligent Systems.
68. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. arXiv.
69. Singh, H., Bose, D., Nagubandi, A. R., Prabhu, S., & Naik, S. G. Cryptocurrency Market Spillovers: Risk Contagion Across Global Financial Systems.
70. Guo, D., Shao, Z., Hao, Y., Wang, S., Xu, Y., Wu, A., & Zhang, H. (2024). DeepSeek LLM: Scaling open-source language models for enterprise applications. arXiv.
71. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., et al. (2024). The Llama 3 herd of models. arXiv.
72. Team, G. (2024). Gemma: Open models based on Gemini research and technology. arXiv.
73. Garapati, R. S., & Kanna, S. R. A Digital Twin-Enabled Predictive Maintenance Framework Leveraging Multi-Agent Reinforcement Learning and Industrial IoT Data.
74. Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2023). How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 11, 962–977.
75. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 34, 9459–9474.
76. Mangalampalli, B. M., & Kolla, S. K. (2025). Large Language Models for Automated Healthcare Data Dictionary Generation and Maintenance. Vascular and Endovascular Review, 8(20s), 363-375.
77. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 874–880.
78. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. ACM Computing Surveys, 57(3), 1–38.
79. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Wang, C., & Liu, C. (2023). AutoGen: Enabling next-generation large language model applications via multi-agent conversation. arXiv.
80. Davuluri, P. N. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems.
81. Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain-of-thought reasoning in language models. International Conference on Learning Representations.
82. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.
83. GARAPATI, R. S. SYNERGETIC INTELLIGENCE Converging AI, Cloud, IoT, and Smart Automation for Real-Time Futures. CANEDA GLOBAL JOURNAL GROUP.
84. Yao, S., Yu, D., Zhao, J., Shafran, I., Narasimhan, K., Cao, Y., & Cao, Y. (2023). ReAct: Synergizing

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 03 ISSUE: 04
RECEIVED: OCTOBER 24 REVISED: NOVEMBER 06 ACCEPTED: NOVEMBER 28 PUBLISHED: MARCH 09

reasoning and acting in language models. International Conference on Learning Representations.

85. Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., & Lewis, M. (2023). Measuring and narrowing the compositionality gap in language models. Findings of the Association for Computational Linguistics: EMNLP 2023, 5687–5710.
86. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv.
87. Bandi, V. D. V. K. AI-Based Anomaly Detection Frameworks in Distributed Enterprise Data Systems.
88. Abdin, M., Aneja, J., Awadalla, H. H., Awadallah, A., Bach, N., Bahree, A., Bakhtiari, A., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. arXiv.
89. Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C., et al. (2024). Phi-3 Mini: Technical report. Microsoft Research.
90. Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2024). QLoRA: Efficient finetuning of quantized large language models. Advances in Neural Information Processing Systems, 36.
91. Kirk, H. R., Whitefield, A., Röttger, P., Vidgen, B., & Hale, S. A. (2024). The Prism alignment project: What participatory, representative, and individualised human feedback reveals about the subjective and multicultural alignment of large language models. Advances in Neural Information Processing Systems, 37.
92. Reddy, V. A. R., & Kolla, S. K. (1984). Infrastructure-As-Code Practices For Regulated Healthcare Cloud Environments. Metallurgical and Materials Engineering, 30 (4), 1028–1042.
93. Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI in business and information systems engineering. Business & Information Systems Engineering, 66(1), 111–126.
94. Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., Dennehy, D., Metri, B., Buhalis, D., Cheung, C. M. K., Conboy, K., Doyle, R., Dubey, R., Dutot, V., Felix, R., Goyal, D. P., Gustafsson, A., Hinsch, C., Jebabli, I., et al. (2023). Opinion paper: So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.
95. LEBCIR, I., Shah, C. A., & Appa Rao Nagubandi, D. S. M. D. FinTech and Financial Inclusion: Empirical Evidence from Emerging Markets.