

The efficiency paradox: A critical review of domain-specific vs. general-purpose AI in enterprise information systems

Gelareh Towhidi, *University of West Georgia*, gtowhidi@westga.edu

Saeed Tabar, *University of West Georgia*, stabar@westga.edu

Mariana Sanchez, *University of West Georgia*, msanchez@westga.edu

Abstract

This paper presents a critical review of the emerging trend toward domain-specific artificial intelligence (AI) and small language models (SLMs) within enterprise information systems (EIS). In contrast to the prevailing narrative favoring scale and generality, this study identifies and examines an apparent efficiency paradox: specialized, often smaller models can deliver superior performance on specific enterprise tasks while requiring significantly fewer computational, financial, and data resources in well-defined domains. Through a systematic review of literature (2020-2025), the authors synthesize the drivers of this trend, including cost pressures, data privacy imperatives, and the need for task-specific reliability. A conceptual framework is proposed categorizing domain-specific AI along three dimensions: (1) Specialization Depth, (2) Data Efficiency & Privacy, and (3) Architectural Integration. The paper analyzes implementation patterns across key domains and rigorously examines the mechanisms and boundary conditions underlying the efficiency paradox. Significant organizational and technical challenges are delineated, such as integration complexity, talent gaps, and the shift from operational to capital expenditure in total cost of ownership. Finally, a multifaceted research agenda for Information Systems (IS) scholars is proposed. This work contributes to IS theory by providing a structured lens to analyze AI specialization and offers pragmatic insights for EIS strategy and governance.

Keywords: domain-specific AI, enterprise AI, AI efficiency, small language models, AI strategy, information systems architecture

Introduction

The ascendancy of large language models (LLMs) and foundation models has catalyzed a wave of adoption of general-purpose artificial intelligence (AI) in enterprises. These models demonstrate remarkable versatility across diverse tasks, from content generation to basic reasoning. However, their integration into enterprise information systems (EIS) has unveiled significant impediments: prohibitive computational and application programming interface (API) costs, intractable data privacy and security risks, insufficient domain expertise leading to "hallucinations" and inaccuracies, and operational inefficiencies in latency-sensitive environments (Bender et al., 2021; Wei et al., 2022). This has precipitated a countervailing trend toward domain-specific AI - systems intentionally designed, trained, and optimized for constrained application domains, industries, or enterprise functions (Raghu & Schmidt, 2020).

This paper addresses a core strategic tension in EIS architecture and governance: the trade-off between the versatility of general-purpose AI and the precision, efficiency, and control of domain-specific approaches. While industry discourse often equates capability with scale, empirical evidence from enterprise

deployments reveals a more nuanced reality. This study conceptualizes this as the "efficiency paradox": the phenomenon where seemingly less capable, specialized models outperform their larger, general counterparts on targeted tasks while consuming a fraction of the resources. For instance, Bolton et al. (2024) demonstrated that BiomedLM, a 2.7-billion parameter model fine-tuned on biomedical text, outperformed general models ten times its size on clinical natural language processing (NLP) tasks while enabling on-premise deployment that kept sensitive patient data secure. Similarly, in the legal domain, domain-adapted models have achieved superior contract review accuracy at a fraction of the inference cost of general-purpose LLMs (Zhao et al., 2024). This paradox challenges prevailing assumptions and warrants systematic scholarly investigation within the information systems discipline.

While prior work has addressed model compression (Schwartz et al., 2020), task-technology fit in AI (Goodhue & Thompson, 1995), and enterprise system design (Xu et al., 2024), no study has systematically synthesized these streams to examine the strategic tension between generality and specialization in enterprise AI. The contribution of this paper is threefold: (1) to synthesize the theoretical foundations and empirical evidence motivating the shift to domain-specific AI, identifying and theorizing the "efficiency paradox" as a distinct phenomenon requiring IS attention; (2) to develop a multidimensional framework that integrates technical, organizational, and strategic dimensions of AI specialization for analyzing and classifying implementations within organizations; and (3) to articulate a forward-looking research agenda that moves beyond technical performance to address implementation, value, and governance. This synthesis, and the framework that emerges from it, represents a novel contribution to the IS literature on enterprise AI.

Methodology

This paper employs a systematic conceptual review methodology following established guidelines for integrative reviews in information systems research (Webster & Watson, 2002; vom Brocke et al., 2015). The review focuses on synthesizing conceptual and empirical literature to develop theoretical insights and frameworks rather than conducting meta-analysis of quantitative results.

Literature Search Strategy: Searches were conducted across Google Scholar, IEEE Xplore, ACM Digital Library, and AIS eLibrary for publications between January 2020 and August 2025. Primary search terms included: "domain-specific AI," "enterprise AI," "small language models," "specialized AI," "efficient AI," "AI for business," combined with "information systems," "enterprise systems," and "organizational implementation." Peer-reviewed journal articles and conference proceedings were prioritized, supplemented with influential preprints and industry white papers when they provided unique insights.

Inclusion Criteria:

- Publications addressing AI specialization or efficiency in enterprise/organizational contexts
- Theoretical or empirical papers comparing general vs. domain-specific AI approaches
- Literature discussing implementation challenges, ROI, or strategic implications
- Papers published between 2020-2025, with emphasis on 2023-2025 for emerging trends

Exclusion Criteria:

- Technical papers focused exclusively on algorithmic details without organizational implications
- Papers addressing AI in consumer applications without enterprise relevance
- Publications before 2020 unless they established foundational concepts

Sample and Screening Process: The initial search yielded 847 results across all databases. After removing duplicates (n=189) and screening abstracts against inclusion/exclusion criteria, 234 full-text articles were assessed. Of these, 97 met all inclusion criteria and were included in the final synthesis. Articles were assessed for quality based on publication venue, methodological rigor (for empirical studies), and relevance to the research questions.

Data Extraction and Analysis: Data was extracted from included articles using a standardized form capturing: (1) publication details, (2) research focus and questions, (3) methodology, (4) key findings relevant to AI specialization, (5) implications for enterprise implementation, and (6) limitations. Thematic

analysis was employed to identify recurring patterns, contradictions, and gaps in the literature. The analysis followed an iterative process of coding, categorization, and synthesis, leading to the development of the conceptual framework presented in Section 4.

Literature Review Process Overview:

The literature review process is summarized in Figure 1 below, following established guidelines for systematic literature reviews in information systems research (Webster & Watson, 2002; vom Brocke et al., 2015).

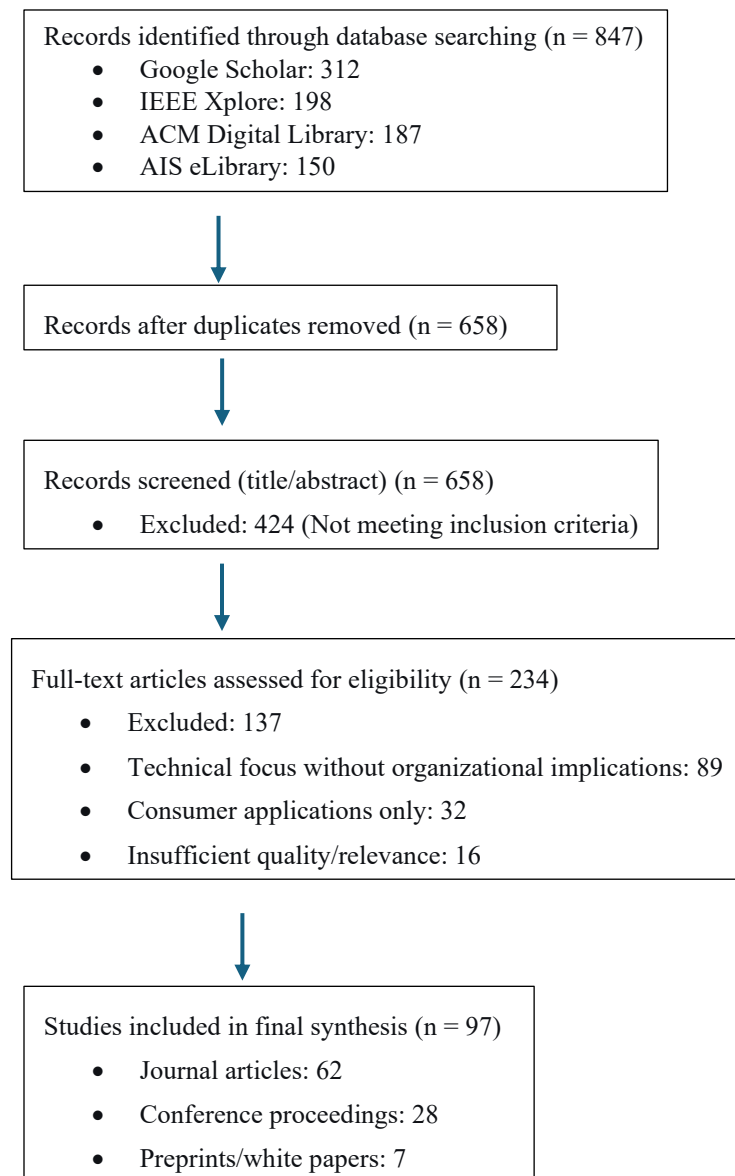


Figure 1. Literature Review Process Flow Diagram

Note: The diagram illustrates the systematic screening and selection process, showing the number of records at each stage from initial identification through final inclusion in the synthesis.

The initial search across all databases (Google Scholar, IEEE Xplore, ACM Digital Library, and AIS eLibrary) yielded 847 results. After removing duplicates ($n=189$), 658 unique records were screened based on titles and abstracts against the inclusion and exclusion criteria. Of these, 424 records were excluded, leaving 234 full-text articles for detailed assessment. Following full-text review, 137 articles were excluded for reasons including: exclusive focus on technical algorithms without organizational implications ($n=89$), consumer applications without enterprise relevance ($n=32$), and insufficient quality or relevance to the research questions ($n=16$). The final synthesis included 97 studies (62 journal articles, 28 conference proceedings, and 7 influential preprints/industry white papers).

Quality Assessment:

Included studies were assessed for quality based on: (1) publication venue (peer-reviewed journals and conferences prioritized), (2) methodological rigor for empirical studies (sample size, research design, validity measures), (3) relevance to the research questions, and (4) contribution to understanding domain-specific AI in enterprise contexts. Studies that met at least three of four criteria were included in the final synthesis.

Theoretical Foundations: Distinguishing Domain-Specific from General-Purpose AI

The Evolution of AI in Enterprise Contexts

Enterprise adoption of AI has evolved through distinct phases. Early expert systems (1980s-1990s) were inherently domain-specific but limited in scope and adaptability (Liao, 2005). The machine learning era (2000s-2010s) introduced more flexible approaches but still required substantial domain customization (Jordan & Mitchell, 2015). The current era of foundation models represents a paradigm shift toward general-purpose capabilities that can be adapted to various domains through techniques like fine-tuning and prompt engineering (Wiggins & Tejani, 2022).

However, this generality comes at significant cost. Large foundation models like GPT-4, Claude, and Gemini require massive computational resources for training and inference, raising concerns about sustainability and accessibility (Patterson et al., 2021; Wu et al., 2023). Additionally, their broad training on internet-scale data introduces risks of misinformation, bias amplification, and privacy violations in enterprise settings (Weidinger et al., 2021; Kasirzadeh & Gabriel, 2023). Critically, the "black box" nature of these models poses significant challenges for enterprise governance, as organizations may struggle to explain, audit, or control their behavior in high-stakes contexts (Rudin, 2019).

Defining Domain-Specific AI

Domain-specific AI refers to artificial intelligence systems designed, trained, and optimized for particular application domains, industries, or enterprise functions. These systems prioritize precision, efficiency, and relevance within their target domain over general capabilities (Raghu & Schmidt, 2020). Key characteristics include:

- **Specialized Training Data:** Utilization of domain-specific corpora, ontologies, and knowledge bases rather than general internet data (Gururangan et al., 2020).
- **Architectural Optimization:** Model architectures tailored to domain requirements, such as incorporating domain knowledge through symbolic components or specialized attention mechanisms (Addepalli et al., 2022).
- **Task-Specific Evaluation:** Performance metrics aligned with domain-specific outcomes rather than general benchmarks (Raji et al., 2021).
- **Integration with Domain Tools:** Designed to interface with existing enterprise systems, workflows, and domain-specific software (Xu et al., 2024).

The Rise of Small Language Models (SLMs)

A particularly significant development within domain-specific AI is the emergence of small language models, typically under 10 billion parameters, that rival or surpass larger models on specific tasks (Bolton et al., 2024; Grattafiori et al., 2024). These models benefit from:

- **Efficient Training:** Reduced computational requirements and carbon footprint (Schwartz et al., 2020).
- **Privacy Preservation:** Ability to train on sensitive enterprise data without external exposure (Li et al., 2024).
- **Customization Flexibility:** Easier fine-tuning and adaptation to evolving domain needs (Ding et al., 2023).
- **Operational Efficiency:** Lower inference costs and latency, enabling real-time applications (Mendez et al., 2022).

Notable examples include Microsoft's Phi series (2.7B parameters), Google's Gemma models (2B-7B parameters), and various enterprise-focused models like BloombergGPT (50B parameters, but finance-specific) (Wu et al., 2023; Xie et al., 2024).

The Distinctiveness of the Efficiency Paradox

The efficiency paradox (defined in Section 6.1) must be distinguished from related but distinct phenomena. While model compression research focuses on technical efficiency, making models smaller and faster without significant performance loss (Schwartz et al., 2020; Hinton et al., 2015), the efficiency paradox addresses a different dynamic: the counterintuitive superiority of smaller, specialized models on domain-specific tasks, driven by the interaction of technical, organizational, and economic factors. Similarly, while task-technology fit theory (Goodhue & Thompson, 1995) addresses general alignment between tasks and technology, the efficiency paradox specifically addresses the trade-off between generality and specialization in the context of AI, where specialization can yield superior performance despite smaller scale. This phenomenon is not reducible to model compression alone, nor to task-technology fit alone, but emerges from their intersection in the context of enterprise AI adoption—a context characterized by unique constraints around data privacy, regulatory compliance, integration, and total cost of ownership.

A Conceptual Framework for Domain-Specific AI in Enterprises

Based on the literature synthesis, this paper proposes a multi-dimensional framework for understanding, evaluating, and designing domain-specific AI implementations in enterprise settings. This framework serves as both an analytical lens for researchers and a decision-making tool for practitioners, addressing the core strategic trade-offs identified in the efficiency paradox.

The framework (Figure 2) organizes domain-specific AI implementations along three interconnected dimensions: (1) **Specialization Depth**, representing the extent to which AI systems incorporate domain expertise; (2) **Data Efficiency & Privacy**, addressing how models balance performance with data governance and resource constraints; and (3) **Architectural Integration**, examining the technical and organizational choices that enable domain specificity. These dimensions are not independent; decisions along one dimension affect options and trade-offs along others. For example, deeper specialization may enable greater data efficiency but may require more complex architectural integration with existing systems.

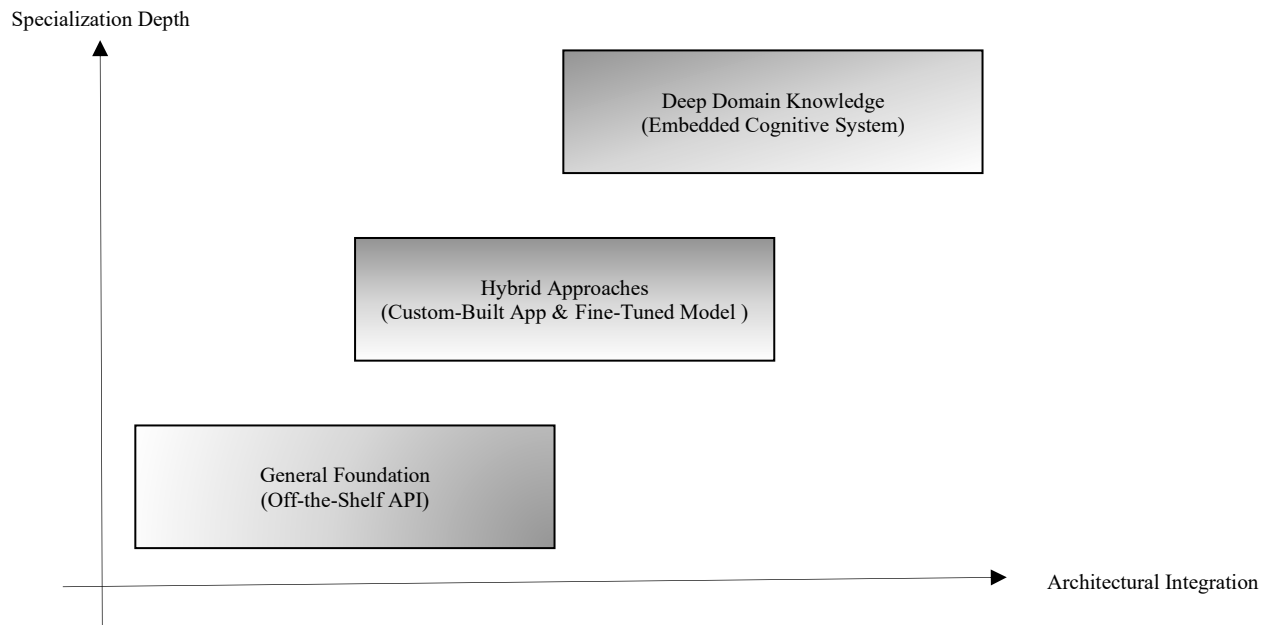


Figure 2. Conceptual Framework for Domain-Specific AI in Enterprises

Note: The framework illustrates three interrelated dimensions- Specialization Depth (vertical axis), Data Efficiency & Privacy (depth axis), and Architectural Integration (horizontal axis)- that define the design space for domain-specific AI implementations.

Dimension 1: Specialization Depth

This dimension captures the extent and manner to which an AI system incorporates deep domain knowledge versus maintaining general capabilities. It represents a continuum from generic tool to embedded expertise.

- **Level 1: General-Purpose AI (Baseline).** Pre-trained on broad internet data, adapted for enterprise use primarily through prompting or lightweight configuration (e.g., GPT-4 with enterprise plugins, off-the-shelf Claude API).
- **Level 2: Domain-Adapted Models.** General foundation models that undergo significant continued pre-training or fine-tuning on curated domain-specific data corpora (e.g., Legal-BERT fine-tuned on case law, a fine-tuned version of Llama 3 on financial news/earnings call transcripts).
- **Level 3: Domain-Designed Models.** Model architectures specifically designed from inception for domain requirements, incorporating structural priors or symbolic components (e.g., AlphaFold for protein folding, BloombergGPT trained on financial data, specialized architectures for time-series forecasting in manufacturing).
- **Level 4: Integrated Domain Cognitive Systems.** AI deeply embedded within domain-specific software applications, with specialized interfaces, workflows, and continuous feedback loops (e.g., AI-powered diagnostic support integrated into a hospital's Electronic Health Record (EHR)/Picture Archiving and Communication System (PACS), predictive maintenance modules within a Computerized Maintenance Management System (CMMS)).

Managerial Decision Variables for Specialization Depth:

- **Task Precision Requirement:** Ranges from Low (general guidance, brainstorming) to High (mission-critical accuracy, legal/financial decisions). Higher precision requirements favor deeper specialization.

- **Domain Knowledge Stability:** Ranges from Dynamic (rapidly evolving fields like cybersecurity) to Stable (well-established domains like legal precedent or accounting standards). More stable domains enable deeper specialization with lower maintenance costs.
- **Error Cost Level:** Ranges from Low (minor inconvenience, internal use) to High (legal liability, patient safety, financial loss). Higher error costs justify greater investment in specialization.
- **Development Cost Tolerance:** Ranges from Low (limited budget, rapid deployment needed) to High (significant investment capacity, long-term horizon). Higher tolerance enables deeper specialization.

Dimension 2: Data Efficiency & Privacy

This dimension addresses how models balance performance with critical enterprise constraints: data privacy, security, compliance, and resource consumption.

- **Data Strategy:** Source (general internet vs. proprietary/sensitive enterprise data vs. synthetic domain data), governance, and lifecycle management.
- **Privacy Approach:** Implementation approach reflecting risk tolerance (centralized cloud training vs. federated learning vs. differential privacy techniques vs. strict on-premise/air-gapped deployment).
- **Efficiency Metrics:** Total computational footprint (training and inference), latency, scalability, and direct/indirect costs (energy, carbon, vendor API fees).

Managerial Decision Variables for Data Efficiency & Privacy:

- **Data Sensitivity Level:** Ranges from Low (publicly available data) to High (proprietary, Personally Identifiable Information (PII), regulated data such as Health Insurance Portability and Accountability Act (HIPAA) or General Data Protection Regulation (GDPR)). Higher sensitivity favors specialized, on-premise solutions.
- **Regulatory Compliance Requirement:** Ranges from Low (no specific regulatory constraints) to High (strict industry regulations like HIPAA, GDPR, Financial Industry Regulatory Authority (FINRA)). Higher compliance requirements favor specialized solutions with audit capabilities.
- **Latency Requirement:** Ranges from Lenient (batch processing, non-real-time) to Stringent (millisecond-level response, real-time). Higher latency requirements favor smaller, specialized models.
- **Cost Model Preference:** Ranges from Variable (operational expense, pay-per-use API) to Fixed (capital expense, owned infrastructure). Fixed cost preferences favor investment in specialized models.

Dimension 3: Architectural Integration

This dimension examines technical and organizational implementation choices that enable and sustain domain specificity.

- **Knowledge Integration Mechanism:** How domain knowledge is encoded and accessed (prompt engineering, Retrieval-Augmented Generation (RAG), integration with enterprise knowledge graphs, neuro-symbolic components).
- **System Integration Pattern:** Manner of integration with existing EIS (loose coupling via API, tight embedding within Enterprise Resource Planning (ERP)/Customer Relationship Management (CRM) workflows, event-driven microservices).
- **Adaptability & Evolution:** Mechanisms for continuous learning, versioning, and adaptation to domain evolution without catastrophic forgetting or performance degradation.

Managerial Decision Variables for Architectural Integration:

- **Workflow Maturity:** Ranges from Ad hoc/Unstructured (creative tasks, exploration) to Structured/Standardized (routine processes, clear rules). More mature workflows enable tighter integration.
- **IT Infrastructure Readiness:** Ranges from Legacy/On-premise (limited API capabilities) to Cloud-native/API-ready (modern microservices architecture). Higher readiness enables more sophisticated integration patterns.
- **Domain Evolution Rate:** Ranges from Slow (years, stable domains like legal) to Rapid (weeks/months, dynamic domains like technology). Higher evolution rates require more sophisticated adaptation mechanisms.
- **Model Adaptation Requirement:** Ranges from Static (one-time deploy with minimal updates) to Continuous (frequent retraining, online learning). Higher adaptation requirements favor architectures with robust versioning and monitoring.

Framework Synthesis and Implications

The three dimensions of the framework are interconnected; decisions made along one dimension have implications for the others. For example, deeper specialization (Dimension 1) may enable greater data efficiency (Dimension 2) by filtering out irrelevant information, but may require more sophisticated architectural integration (Dimension 3) to maintain domain currency. Similarly, stringent privacy requirements (Dimension 2) may constrain architectural choices (Dimension 3) by limiting cloud-based integration options.

Table 1 synthesizes the managerial decision variables across all three dimensions, providing practitioners with a structured approach to evaluating AI specialization decisions. The strategic implications column highlights how each variable influences the choice between general-purpose and domain-specific AI approaches. Organizations can use this table as a decision-support tool, assessing their position on each variable to determine the optimal AI strategy for specific tasks or functions.

The framework also illuminates the efficiency paradox: smaller, specialized models can achieve superior performance when the interactions among dimensions are optimized. Specifically, when an organization faces high task precision requirements, high data sensitivity, and low latency tolerance, the integrated solution of a deeply specialized, privacy-preserving, tightly integrated model yields the efficiency paradox outcome. Conversely, when tasks require broad cross-domain reasoning, involve low-stakes variety, or operate in rapidly shifting domains with limited training data, general-purpose models retain their advantage.

For practitioners, this framework provides a structured approach to AI strategy, guiding investment decisions and architecture choices based on organizational context. For researchers, it offers a theoretical lens for analyzing AI specialization patterns and their organizational implications.

Table 1. Summary of Managerial Decision Variables Across Framework Dimensions

Dimension	Variable Name	Definition	Range	Strategic Implication
Specialization Depth	Task Precision Requirement	Level of accuracy needed for task success	Low → High	Higher precision favors deeper specialization
	Domain Knowledge Stability	How rapidly domain knowledge evolves	Dynamic → Stable	Stable domains enable deeper specialization
	Error Cost Level	Consequences of AI errors	Low → High	Higher error costs justify specialization
	Development Cost Tolerance	Capacity to invest in custom solutions	Low → High	Higher tolerance enables deeper specialization

Dimension	Variable Name	Definition	Range	Strategic Implication
Data Efficiency & Privacy	Data Sensitivity Level	Confidentiality/privacy of data	Low → High	Higher sensitivity favors on-premise specialized models
	Regulatory Compliance Requirement	Stringency of regulatory constraints	Low → High	Higher compliance favors specialized, auditable systems
	Latency Requirement	Speed needed for inference	Lenient → Stringent	Higher latency requirements favor smaller models
	Cost Model Preference	OpEx vs. CapEx preference	Variable → Fixed	Fixed preference favors investment in specialized models
Architectural Integration	Workflow Maturity	Structure of target workflows	Ad hoc → Structured	Mature workflows enable tighter integration
	IT Infrastructure Readiness	Technical capabilities of existing systems	Legacy → Cloud-native	Higher readiness enables sophisticated integration
	Domain Evolution Rate	Speed of domain knowledge change	Slow → Rapid	Faster evolution requires robust adaptation mechanisms
	Model Adaptation Requirement	Need for continuous model updates	Static → Continuous	Higher adaptation needs favor flexible architectures

Implementation Patterns Across Enterprise Domains

The theoretical advantages of domain-specific AI manifest in distinct implementation patterns across different enterprise domains. These patterns are shaped by industry-specific drivers such as regulatory constraints, operational requirements, data characteristics, and the criticality of accuracy. To provide a structured comparison, the literature is synthesized across four key domains—healthcare, finance, manufacturing, and legal, highlighting how the dimensions of the conceptual framework are prioritized and instantiated differently in each context. Table 2 provides a comparative overview, followed by a detailed analysis of each domain.

Table 2. Comparative Analysis of Domain-Specific AI Implementation Patterns

Domain	Primary Drivers	Typical Specialization Depth	Exemplar Model/System	Key Integration Challenge
Healthcare	Regulatory compliance (HIPAA), critical accuracy, specialized terminology	Level 3-4 (Domain-Designed/Integrated Systems)	BiomedLM, EHR-integrated diagnostic support	Interoperability with diverse EHR/PACS systems; validating clinical utility
Finance	Latency, regulatory explainability, data privacy, precision	Level 2-3 (Domain-Adapted/Designed)	BloombergGPT, fraud detection ensembles	Real-time integration with market data feeds; meeting audit trail requirements
Manufacturing	Edge deployment constraints, real-time inference, sensor data fusion	Level 3-4 (Domain-Designed/Integrated)	Edge-deployed predictive maintenance models	Integration with Operational Technology (OT)/SCADA systems;

Issues in Information Systems

Volume 27, Issue 1, pp. 29-47, 2026

Domain	Primary Drivers	Typical Specialization Depth	Exemplar Model/System	Key Integration Challenge
				reliability in harsh environments
Legal	Precision, citation accuracy, confidentiality, adherence to precedent	Level 2-4 (Adapted to Integrated)	Contract review systems with RAG over case law	Mapping to dense, structured legal workflows and document management systems

Healthcare and Life Sciences

Healthcare represents a paradigmatic case for domain-specific AI due to stringent regulatory requirements, specialized terminology, and critical accuracy needs (Rajkomar et al., 2019). Implementation patterns include:

- **Clinical Decision Support:** Models trained specifically on electronic health record (EHR) data structured with medical ontologies (e.g., SNOMED CT, UMLS) and clinical coding systems to predict patient outcomes or recommend interventions (Shickel et al., 2017).
- **Medical Imaging:** Architectures optimized for specific imaging modalities (MRI, CT, X-ray) rather than general computer vision (Esteva et al., 2021).
- **Drug Discovery:** Generative models trained on chemical compound databases with molecular property prediction (Stokes et al., 2020).

Empirical studies demonstrate that domain-specific models with 1-5 billion parameters can outperform 100B+ general models on specialized medical tasks while maintaining patient privacy through federated learning approaches (Li et al., 2024; Bolton et al., 2024). However, challenges remain in validating clinical utility and ensuring interoperability with diverse EHR systems.

Financial Services

The finance industry has rapidly adopted domain-specific AI for tasks requiring precision, compliance, and real-time processing (Goldstein et al., 2021). Notable implementations include:

- **BloombergGPT (50B):** Specifically trained on financial data, outperforming larger general models on financial NLP tasks while using proprietary data securely (Wu et al., 2023).
- **Fraud Detection:** Ensemble models combining traditional rule-based systems with specialized neural networks trained on transaction patterns (Jurgovsky et al., 2018).
- **Algorithmic Trading:** Lightweight models optimized for low-latency inference on market data streams (Xie et al., 2024).

The efficiency advantage is particularly pronounced in finance, where milliseconds matter and regulatory compliance requires explainability—a challenge for black-box general models. However, the dynamic nature of financial markets presents ongoing challenges for model adaptation and maintenance.

Manufacturing and Supply Chain

Manufacturing environments benefit from AI specialized for predictive maintenance, quality control, and logistics optimization (Tao et al., 2018). Implementation patterns include:

- **Predictive Maintenance:** Small models deployed at the edge (factories, warehouses) trained on equipment sensor data (Kumar et al., 2023).
- **Quality Inspection:** Computer vision models optimized for specific product lines and defect types rather than general object recognition.
- **Supply Chain Optimization:** Hybrid models combining operations research with machine learning trained on historical logistics data (Baryannis et al., 2019).

Edge deployment constraints (limited compute, intermittent connectivity) strongly favor specialized, efficient models over general-purpose alternatives (Mendez et al., 2022). Critical challenges include

integration with legacy operational technology (OT) systems and ensuring reliability in harsh industrial environments.

Legal and Compliance

The legal domain requires precise interpretation of statutes, precedents, and contracts—tasks where general LLMs frequently hallucinate or provide inaccurate citations (Choi & Schwarcz, 2023). Domain-specific approaches include:

- **Legal Document Analysis:** Models trained exclusively on legal corpora with citation verification mechanisms (Zhao et al., 2024).
- **Regulatory Compliance:** Systems that track regulatory changes and map them to organizational policies using specialized knowledge graphs (Phillips et al., 2021).
- **Contract Review:** Focused models for specific contract types (NDAs, licensing agreements) with clause extraction and risk assessment capabilities.

The legal domain illustrates a key nuance of the efficiency paradox: while specialized models outperform general models on accuracy, the integration challenge is particularly acute due to the complexity of legal workflows and document management systems.

The Efficiency Paradox: Analysis and Evidence

Defining the Paradox

The "efficiency paradox" refers to the counterintuitive finding that smaller, specialized models often achieve superior performance on domain-specific tasks compared to larger general models, while requiring significantly fewer computational resources. This paradox challenges the prevailing assumption that "bigger is better" in AI and has significant implications for enterprise AI strategy. The analysis below identifies several mechanisms driving this paradox, which was first introduced in Section 3.4.

Knowledge Distillation Effects

Specialized training on high-quality domain data effectively distills relevant knowledge while filtering out irrelevant information from general training corpora (Hinton et al., 2015; Xu et al., 2024). This focused learning enables more efficient parameter utilization. Empirical evidence supports this: Gururangan et al. (2020) demonstrated that continued pre-training on domain-specific data significantly improves performance compared to general models, even with smaller parameter counts.

Architectural Optimization

Domain-specific architectures can incorporate structural priors and inductive biases that match domain characteristics. For example:

- Hierarchical structures for document processing
- Temporal patterns for time-series forecasting
- Spatial relationships for geospatial analysis
- Graph structures for molecular property prediction

These optimizations enable smaller models to capture essential patterns that larger general models might learn inefficiently or not at all (Addepalli et al., 2022). The efficiency gains from architectural optimization can be substantial: a specialized architecture may achieve equivalent performance to a general model 10-100 times its size.

Task-Specific Evaluation

General benchmarks (MMLU, HellaSwag, GLUE, etc.) often fail to capture domain-specific requirements like precision, regulatory compliance, or integration with existing workflows (Raji et al., 2021). When evaluated on authentic domain tasks, specialized models frequently outperform despite lower general

benchmark scores. This suggests that general benchmarks may be poor predictors of enterprise performance, particularly in high-stakes domains.

Data Quality over Quantity

Domain-specific models benefit from curated, high-quality training data relevant to their tasks, whereas general models ingest massive but noisy internet data containing contradictions, biases, and irrelevant information (Gururangan et al., 2020). This quality advantage can compensate for smaller training datasets. For instance, in the medical domain, a carefully curated corpus of clinical notes and research papers may provide more useful training signal than the entire internet, despite being several orders of magnitude smaller.

Empirical Evidence: Multiple studies support the efficiency paradox. Bolton et al. (2024) showed BiomedLM (2.7B parameters) outperformed 100B+ parameter general models on clinical NLP tasks. Wu et al. (2023) demonstrated BloombergGPT (50B) surpassed larger general models on financial tasks while using proprietary data securely. Xie et al. (2024) found that specialized financial models achieved superior performance on sentiment analysis and named entity recognition tasks compared to general-purpose alternatives.

Boundary Conditions of the Paradox

It is critical to delineate the boundaries where the efficiency paradox may weaken or reverse, transforming the apparent paradox into a strategic choice. The advantage of domain-specific AI is most pronounced in tasks characterized by:

- **Well-defined knowledge boundaries** and stable ontologies where the domain of discourse is clearly circumscribed
- **High-cost of errors** (financial, legal, medical) where precision and reliability outweigh flexibility
- **Proprietary or privacy-sensitive data** that cannot be exposed to third-party services
- **Stringent real-time performance or latency requirements** where inference speed is critical
- **Standardized and repetitive tasks** where the domain knowledge is stable and well-understood

Conversely, general-purpose models retain a strong, often necessary, advantage in tasks requiring:

- **Broad cross-domain reasoning** that synthesizes information across multiple domains
- **Open-ended creativity** such as brainstorming or content generation
- **Operation in novel environments** with poorly defined rules or rapidly shifting requirements
- **Low-stakes, high-variety tasks** where the cost of error is minimal and flexibility is paramount
- **Situations with limited data availability** where the broad knowledge of general models provides necessary prior knowledge

Therefore, the strategic implication for enterprises is not a binary choice but a hybrid, portfolio approach. Organizations must map their processes and evaluate the cost-benefit trade-off, weighing the higher upfront capital expenditure (CapEx) for developing specialized systems against the ongoing operational expenditure (OpEx) of general model APIs, to leverage each type of AI for its comparative advantage.

The Resource Reallocation Virtuous Cycle

Beyond the technical mechanisms, a crucial managerial mechanism underpins the efficiency paradox. The significant reduction in direct computational and API costs achieved by deploying a smaller, specialized model frees up financial and human resources. These resources can be strategically reallocated to areas that further amplify performance and reliability within the domain: (1) higher-quality data curation and labeling, (2) more rigorous validation and testing against domain-specific edge cases, (3) development of better human-AI interaction interfaces for expert-in-the-loop systems, and (4) investment in monitoring and continuous learning pipelines. This creates a virtuous cycle where efficiency gains enable quality investments that, in turn, deepen the sustainable performance advantage over general-purpose alternatives that remain "black box" services.

Challenges and Limitations of Domain-Specific AI

Despite compelling advantages, domain-specific AI approaches face significant challenges that must be addressed for widespread enterprise adoption.

Development and Maintenance Costs

While inference costs are lower, developing specialized models requires domain expertise, curated datasets, and ongoing maintenance as domains evolve (Raghu & Schmidt, 2020). The total cost of ownership must be evaluated against the alternative of using general models via API. Critically, the shift from operational expense (OpEx, pay-per-use APIs) to capital expense (CapEx, development and infrastructure) represents a significant change in cost structure that organizations must carefully manage. This is not merely a technical trade-off but a financial one, with implications for budgeting, procurement, and organizational strategy.

Integration Complexity

Integrating specialized AI into existing enterprise systems often requires substantial adaptation of workflows, interfaces, and data pipelines (Xu et al., 2024). General models with standardized APIs can offer simpler integration paths despite other limitations. The integration challenge is particularly acute for legacy systems, where API capabilities may be limited or nonexistent. Organizations may need to invest in middleware, data transformation, and workflow redesign to realize the benefits of domain-specific AI.

Talent and Skill Gaps

Developing and maintaining domain-specific AI requires rare combinations of deep domain expertise and AI/machine learning skills (Davenport & Ronanki, 2018). Many organizations struggle to attract or develop this hybrid talent. The shortage of "bilingual" professionals—fluent in both domain knowledge and AI techniques—represents a significant bottleneck to adoption. Organizations may need to develop new training programs, collaborate with universities, or form partnerships to address this gap.

Risk of Over-Specialization

Excessive specialization can create fragile systems that fail to adapt to domain evolution or edge cases not represented in training data (Rudin, 2019). Maintaining appropriate generalization within domain boundaries remains challenging. This risk is particularly acute in dynamic domains where the knowledge base changes rapidly. Organizations must balance the benefits of specialization against the costs of maintaining currency and handling exceptions.

Measurement and ROI Uncertainty

Quantifying the business value of domain-specific AI implementations presents methodological challenges, particularly for intangible benefits like improved decision quality or risk reduction (Füller et al., 2022). Organizations need better frameworks for AI investment evaluation. Traditional ROI metrics may not capture the full value of specialized AI, particularly when benefits are indirect or long-term.

Research Agenda for Information Systems Scholars

Based on the review and framework, the following multifaceted research agenda is proposed, addressing critical theoretical, design, and managerial gaps at the intersection of domain-specific AI and information systems. This agenda leverages the distinctive methodologies of IS research to move beyond technical performance to address implementation, value, and governance.

Theoretical Development

Domain-specific AI challenges core assumptions in IS theory, requiring extensions and new models that account for systems that are dynamically adaptable, deeply embedded in specialized workflows, and co-evolutionary with organizational knowledge. Research in this area should focus on revising established theories and constructing new frameworks that can explain and predict how specialization impacts technology adoption, organizational capability, and knowledge integration.

- **Adapting IS Theories:** How should established IS theories (Technology Acceptance Model, Task-Technology Fit, Diffusion of Innovation) be extended to account for domain-specific AI characteristics? For example, does the concept of "perceived usefulness" capture the unique value of domain-specific AI, or do new constructs (e.g., "perceived precision" or "perceived privacy") need to be introduced?
- **AI Capability Framework:** Development of comprehensive frameworks for assessing organizational AI capabilities with specialization as a core dimension. How do capabilities differ for organizations pursuing general-purpose vs. domain-specific AI strategies?
- **Knowledge Integration Theories:** Theoretical models explaining how domain knowledge integrates with machine learning in organizational contexts. What mechanisms enable effective integration of human expertise with AI systems?

8.2 Design Science and Implementation

The practical realization of domain-specific AI's potential is hindered by a lack of reusable design knowledge, governance tools, and evaluation methods tailored to specialized systems. A design science research (DSR) agenda should produce and validate artifacts, such as architectural patterns, governance frameworks, and assessment methodologies that lower implementation barriers, ensure responsible management, and demonstrate measurable organizational impact.

- **Architecture Patterns:** Design patterns for integrating domain-specific AI with enterprise systems (ERP, CRM, SCM). What are the common integration architectures, and what are their trade-offs?
- **Governance Artifacts:** Tools and frameworks for governing portfolios of specialized AI systems across organizations. How should organizations manage the proliferation of specialized AI systems and ensure consistency and quality?
- **Evaluation Methodologies:** Novel methods for evaluating domain-specific AI performance, ROI, and organizational impact. What metrics best capture the value of specialized AI in enterprise contexts?

Organizational and Managerial Studies

The shift toward domain-specific AI introduces profound changes in talent requirements, decision-making authority, and organizational learning pathways. Behavioral and managerial studies are needed to explore how organizations can develop hybrid expertise, how decision processes are augmented or transformed by specialized AI, and how learning loops between human experts and AI systems can be institutionalized.

- **Talent Management:** Strategies for developing, acquiring, and retaining hybrid domain/AI talent. How are leading organizations addressing the talent gap?
- **Decision Processes:** How domain-specific AI transforms decision-making processes within specialized functions. Does specialization lead to greater trust in AI systems, or does it raise new concerns?
- **Organizational Learning:** Mechanisms through which organizations learn from and with domain-specific AI systems. How do organizations capture and disseminate insights from specialized AI deployments?

Ethical and Societal Implications

Domain-specific AI does not eliminate ethical concerns but reshapes them, potentially concentrating risks within high-stakes domains and altering competitive landscapes and professional identities. IS research

must critically examine whether specialization amplifies or mitigates bias, how it affects equity of access and competitive advantage, and how it redefines expertise and professional work within organizations.

- **Bias in Specialized Systems:** Whether domain specialization exacerbates or mitigates AI bias compared to general models. Does focusing on a specific domain make bias more detectable and correctable, or does it create new forms of exclusion?
- **Access and Inequality:** Implications of domain-specific AI for organizational digital divides and competitive dynamics. Does specialization advantage large organizations with data resources, or does it democratize access to AI?
- **Workforce Transformation:** How domain-specific AI reshapes specialized professions and expertise development. What happens to domain expertise as AI systems increasingly handle specialized tasks?

Comparative and Longitudinal Studies

Longitudinal and comparative studies are needed to understand the evolution of AI specialization strategies and their organizational impacts over time.

- **Evolution of Specialization:** How do domain-specific AI implementations evolve over time? Do they become more or less specialized as organizations gain experience?
- **Industry Comparisons:** Comparative studies across industries to understand how contextual factors shape specialization strategies and outcomes.
- **Success and Failure Factors:** Identification of critical success factors for domain-specific AI implementations, including both technical and organizational determinants.

Empirical Validation Approaches

While this paper has developed a conceptual framework based on literature synthesis, empirical validation is essential for establishing the framework's utility and predictive power. Several approaches are available:

Case Study Research: In-depth case studies of organizations implementing domain-specific AI across industries could validate the framework's dimensions and identify missing elements. Multiple case designs could compare implementations with different configurations along the three dimensions. For example, researchers could select cases representing high versus low specialization depth, strong versus weak privacy requirements, and tight versus loose architectural integration to examine how these configurations relate to implementation outcomes.

Survey Research: A cross-sectional survey could measure the relationships between the framework's dimensions and outcomes such as implementation success, ROI, and user satisfaction. The managerial decision variables presented in Table 1 could be operationalized as survey scales, enabling statistical testing of hypotheses derived from the framework. This approach would allow for broader generalization across industries and organizational types.

Design Science Research: Building and evaluating artifacts based on the framework could demonstrate its practical utility. This could include developing decision-support tools that help organizations select appropriate AI approaches based on their specific circumstances. Such artifacts would provide proof-of-concept evidence for the framework's applicability in real-world settings.

Archival Research: Analyzing publicly available data on AI implementations in enterprises could test hypotheses derived from the framework, such as the relationship between specialization depth and cost efficiency or between data privacy requirements and privacy-preserving approaches. Public datasets on enterprise technology adoption could be leveraged for this purpose.

Action Research: Researchers could partner with organizations implementing domain-specific AI to iteratively refine and validate the framework based on real-world experience. This approach would provide rich, contextual insights into the practical challenges and success factors associated with domain-specific AI implementation.

Synthesis and Prioritization: These approaches are not mutually exclusive but complementary, each addressing different facets of validation. A phased research program would be most effective: early-stage case studies could refine the framework's dimensions and generate hypotheses; subsequent survey research could test these hypotheses across a broad sample; design science research could demonstrate practical utility through artifact development; and action research could provide real-world validation in organizational contexts. Archival research could complement all phases by testing specific relationships using existing data. This multi-method triangulation would provide robust evidence for the framework's validity, practical value, and generalizability across organizational contexts. Researchers should prioritize case study and design science approaches in the near term, as these are best suited to refining the framework and demonstrating its utility, before proceeding to large-scale survey validation.

Conclusion

This conceptual review has examined the growing trend toward domain-specific AI and smaller language models in enterprise information systems. The literature has been synthesized to develop a framework for understanding domain-specific AI implementations across specialization depth, data efficiency, and architectural dimensions.

The analysis reveals a compelling "efficiency paradox" wherein specialized, smaller models frequently outperform larger general models on domain-specific tasks while requiring significantly fewer computational resources. This paradox arises from knowledge distillation effects, architectural optimization, task-aligned evaluation, and data quality advantages. However, the paradox has clear boundary conditions: it is most pronounced in tasks with well-defined knowledge boundaries, high error costs, proprietary data, stringent latency requirements, and stable domain knowledge. Conversely, general-purpose models retain advantages for broad cross-domain reasoning, open-ended creativity, novel environments, and low-stakes variety tasks.

Empirical evidence supports the paradox in domains such as healthcare (Bolton et al., 2024), finance (Wu et al., 2023; Xie et al., 2024), manufacturing (Mendez et al., 2022), and legal (Zhao et al., 2024). However, the strength of evidence varies across domains, with healthcare and finance having the strongest empirical support to date. Future research should continue to build empirical evidence across all domains.

Domain-specific approaches face substantial challenges including development costs, integration complexity, talent gaps, and measurement difficulties. These challenges create rich opportunities for IS research spanning theoretical development, design science, organizational studies, and ethical implications. Critically, the shift from operational to capital expenditure in total cost of ownership represents a significant organizational change that warrants further investigation.

For practitioners, the review and framework suggest that organizations should move beyond a binary choice and adopt a strategic portfolio approach to AI. This involves mapping enterprise tasks to the framework's dimensions and deliberately combining general-purpose models for broad, low-risk applications with domain-specific systems for mission-critical, specialized functions where the efficiency paradox holds. Success requires investment not only in technology but in hybrid talent (domain expert + data scientist), data curation capabilities, and governance structures capable of managing a diverse AI portfolio.

The contribution of this paper is threefold. First, the "efficiency paradox" is theorized as a distinct phenomenon requiring IS attention, with mechanisms and boundary conditions that explain when and why smaller specialized models outperform larger general ones. This is conceptually derived but supported by emerging empirical evidence. Second, a multidimensional framework is developed that integrates technical, organizational, and strategic dimensions of AI specialization, providing both analytical and practical utility. Third, a research agenda is articulated that moves beyond technical performance to address implementation, value, and governance—areas where IS scholars are uniquely positioned to contribute.

As AI continues to transform enterprise information systems, the tension between generality and specialization will remain a central strategic consideration, echoing historical debates in IT architecture. IS

researchers are uniquely positioned to guide organizations through these complex tradeoffs. By developing theories that explain adoption, designing artifacts that facilitate integration, and studying the organizational impacts, the research community can ensure that the turn toward domain-specific AI delivers sustainable business value, enhances human expertise, and aligns with ethical and governance standards. The ultimate implication is that enterprise information systems may evolve from being consumers of monolithic AI services to becoming orchestrators of a dynamic portfolio of specialized, deeply integrated intelligence modules.

Acknowledgements: The authors declare that no generative AI tools were used in the writing of this manuscript.

References

- Addepalli, S., Bhogale, K., Dey, P., & Babu, R. V. (2022). Towards efficient and effective self-supervised learning of visual representations. In *European Conference on Computer Vision* (pp. 523-538). Springer Nature Switzerland.
- Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: State of the art and future research directions. *International Journal of Production Research*, 57(7), 2179-2202.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).
- Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., ... & Manning, C. D. (2024). BiomedLM: A 2.7B parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.
- Choi, J. H., & Schwarcz, D. (2023). AI assistance in legal analysis: An empirical study. *Journal of Legal Education*, 71(3), 401-432.
- Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108-116.
- Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., ... & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3), 220-235.
- Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., ... & Socher, R. (2021). Deep learning-enabled medical computer vision. *NPJ Digital Medicine*, 4(1), 5.
- Füller, J., Hutter, K., Wahl, J., Bilgram, V., & Tekic, Z. (2022). How AI revolutionizes innovation management—Perceptions and implementation preferences of AI-based innovators. *Technological Forecasting and Social Change*, 178, 121598.
- Goldstein, I., Spatt, C. S., & Ye, M. (2021). Big data in finance. *The Review of Financial Studies*, 34(7), 3213-3225.
- Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, 19(2), 213-236.

-
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Vasic, P. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.
- Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100, 234-245.
- Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, 36(2), 27.
- Kumar, A., Shankar, R., & Aljohani, N. R. (2023). A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management*, 108, 15-31.
- Li, Q., Diao, Y., Chen, Q., & He, B. (2024). Federated learning on non-IID data silos: An experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 36(3), 1053-1067.
- Liao, S. H. (2005). Expert system methodologies and applications—A decade review from 1995 to 2004. *Expert Systems with Applications*, 28(1), 93-103.
- Mendez, J., Bierzynski, K., Cuéllar, M. P., & Morales, D. P. (2022). Edge intelligence: Concepts, architectures, applications, and future directions. *ACM Transactions on Embedded Computing Systems (TECS)*, 21(5), 1-41.
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
- Phillips, P. J., Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., ... & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence. *NIST Interagency Report*.
- Raghu, M., & Schmidt, E. (2020). A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54-63.
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589-1604.
- Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688-702.
- Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, 48, 157-169.
- Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, 37(1), 9.
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), xiii-xxiii.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
- Wiggins, W. F., & Tejani, A. S. (2022). On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4), e220119.
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., ... & Huang, J. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37, 95716-95743.
- Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2024). A survey of large language models. *Communications of the ACM*, 67(3), 98-109.