Review

Not peer-reviewed version

An Extensive Review of Organizational AI Adoption Challenges and Consequent Integrated AI Appliance Proposal for Adoption Facilitation and Impact Studies

Pius Onobhayedo *, Peter Cardon , Paul Osemudiame Oamen

Posted Date: 4 December 2025

doi: 10.20944/preprints202512.0428.v1

Keywords: artificial intelligence; adoption; organization; adoption experimentation; organizational AI appliance

 doi:10.20944/preprints202512.0428.v1

Review

An Extensive Review of Organizational AI Adoption Challenges and Consequent Integrated AI Appliance Proposal for Adoption Facilitation and Impact Studies

Pius Onobhayedo [1,*], Peter Cardon [1] and Paul Osemudiame Oamen [2]

[1] University of Southern California
[2] University of Aberdeen
* Correspondence: pius.onobhayedo@usc.edu

Abstract

Although AI is widely believed to have transformative potential in organizations, recent reports reveal that many organizations are grappling with value derivation therefrom and the ability to take ownership of due ethical and regulatory demands, among other responsible uses of technology. Our goal is to examine these challenges with a view to proposing an approach to effective AI adoption by organizations and pave the way for further impact studies. As a first step, we reviewed and clarified these challenges, categorizing them into Weak or Non-Existent Strategy, Poor Data Readiness and Privacy Concerns, Inadequate Integration with Existing Technology Stack, Inadequate Human Knowledge Skills and Attitudes/Abilities, Scalable and Secure Infrastructure Challenges, Ethical Governance Concerns, Regulatory Framework Lag, Responsibility and Accountability Concerns as well as Reliability Concerns. Next, we carried out a thematic review of constituent AI technology innovation concepts and tools that have adoption potential in organizations vis-à-vis Enterprise Resource Planning (ERP). In the light of these reviews, we used inductive reasoning to propose an approach to AI adoption and create a tool (OAAD) that exemplifies our recommendations, and which could facilitate well-informed adoption and real-life impact research. To set a compass for our effective adoption approach proposal, we expanded on Yang et al. (2024) and defined organizational AI readiness as the organization's capacity and disposition to deploy and use AI technology tools in ethical, responsible and accountable ways that add value to the organization. Finally, we make some recommendations for progressive impact studies in line with our proposed adoption experimentation.

Keywords: artificial intelligence; adoption; organization; adoption experimentation; organizational AI appliance

1. Introduction

Artificial Intelligence (AI) is widely believed to have a potential transformative effect on organizations. As far back as about ten years ago, Gartner's top 10 technology trends prediction for 2017 gave a prominent place to AI among other emerging transformative technologies like blockchain, augmented and virtual reality (Panetta 2016). The release of ChatGPT in November 2022, founded on OpenAI's Large Language Model (LLM) and subsequent advancements in reasoning algorithms, further accentuate the attention to AI and its transformative potential (Marr 2023). Ideas on how this can be actualized have also been evolving. Gartner had proposed a Mesh app and service architecture (MASA) that requires significant changes to development tooling and best practices, to meet the challenges of digital business (Panetta 2016). The debut of ChatGPT shifted attention to Generative AI which may be defined as "computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data." (Feuerriegel et al. 2024, p. 111). More recently, Agentic AI which we parsimoniously refer to as goal-oriented, autonomous and

 doi:10.20944/preprints202512.0428.v1

adaptive AI task execution systems, seems to now be the strongly touted, as a path to realizing this transformative aspiration (Acharya et al. 2025; Huang 2025; Reddy 2025).

Perceived enthusiasm about AI potential notwithstanding, many organizations seem to be grappling with how to strategically or profitably incorporate AI into their digital solutions framework (Challapally et al. 2025; Hradecky et al. 2022; Khandabattu 2025; McKinsey 2025; Neumann et al. 2024; Potluri and Serikbay 2025; Selten and Klievink 2024; Shah et al. 2024; Sharma et al. 2022; Vial et al. 2023; World Economic Forum 2024; Yang et al. 2024). For example, in their study involving a systematic review of over 300 publicly disclosed AI initiatives, structured interviews with representatives from 52 organizations, and survey responses from 153 senior leaders collected across four major industry conferences, Challapally et al. (2025) of MIT's Project NANDA (Network Agent AI in a Decentralized Architecture) reported a significant disconnect between investment in AI and positive impact on profit and loss, a disconnect which they referred to as the AI Divide. In their findings, only about 5% of organizations are extracting value from AI investment while the rest are getting zero return. Similarly, McKinsey (2025) reported from their global survey, a climb in AI usage in organizations, yet more than 80 percent say that they have not seen tangible impact on enterprise-level EBIT (Earnings before interest and taxes) from their use of Generative AI. In their 2025 hype cycle artificial intelligence report, Gartner asserted that Gen AI has entered the trough of disillusionment with low-maturity organizations finding it difficult to identify suitable use cases and mature organizations battling with literacy (Khandabattu 2025). Even where there are experiences of success, challenges have also been reported. For example, Dinculeana (2024), in a study of AI adoption in some European banks, concluded that the role of AI in banking customer service is multifaceted and laden with both promises and challenges like regulatory compliance, ethical considerations, data management, technological integration, and organizational change which are all critical factors influencing the successful deployment of AI technologies. Echoing some industry players, Hornyak (2025, p. 1) similarly asserts that agentic AI still faces major challenges. For example, he quotes the Databricks CEO Ali Ghosdsi as having cautioned that "AI agents will compound errors as task complexity grows, making human supervision vital, at least in the near term".

The background we present here suggests the need for more clarity about path to effective adoption by organizations, especially where strategic alignment is crucial. Although no technology (including AI) provided as means for people to carry out their organization's activities may be considered inherently strategic, we hold the assertion of Mendes et al. (2024) that any asset or component that may jeopardize a critical mission may be considered strategic. This level of risk in our opinion requires a more proactive approach to research engagement with organizations, in a way that simultaneously provides expertise to the participating organizations in their adoption process.

1.1. Defining Objectives

Our overarching goal in this work is to inductively propose an approach to effective AI adoption by organizations and pave the way for further impact studies. To achieve this goal, our specific objectives include to

- examine and categorize known barriers to effective adoption,
- carry out thematic review of emerging AI innovation systems with potential for enterprise adoption vis-à-vis alignment, enhancement and/or strategic replacement of features of Enterprise Resource Planning (ERP), a mainstream organizational technology integration pattern,
- propose an approach to effective adoption
- create a tool that exemplifies our adoption proposal which could facilitate well-informed adoption decisions and further impact research.

The relationship with ERP expressed in the second specific objective above is our attempt to facilitate clarity about how emerging AI innovations could fit into existing organizations' technology platforms. Besides applying lessons from reported attempts to adopt AI, we see the need to additionally draw lessons from one or more widely adopted technology system and explore effective AI adoption in the light of alignment, enhancement and/or strategic replacement of the status quo; AI adoption

 doi:10.20944/preprints202512.0428.v1

should not destroy past gains. Our choice of ERP is informed by the fact that it is a well-established and largely standardized technology type that is typically implemented in any given organization type, for integrated use by a wide range of organizational stakeholders, across value chains (supply chain, demand chain, back office, etc.). The term ERP was coined by T. Lee WyLie of Gartner Group as criteria for evaluating the extent to which the software actually integrated both across and within the various functional silos (Nazemi et al. 2012; Robert Jacobs and Ted Weston 2007). From the 1990s ERP system practically became the standard for replacing legacy systems (Nazemi et al. 2012) and it is still considered essential for today's business (Reuters, Thomson 2025).

A detailed literature review of ERP is beyond the scope of this article. However, for clarity, we itemize here the core constituent ERP features to be 1) Rule-based business logic for organizational functional units e.g., Customer Relationship Management (CRM), Asset and Financial Management, Supply Relationship Management, Inventory Management, Sales, Facilities Management, Warehouse Management, Project Management, Human Resource Management (McGaughey and Gunasekaran 2007); 2) Database Management System (DBMS) for integrated data availability, integrity, security and independence (Sumathi 2007) across the functional units; and 3) Authentication and Authorization for access control. Features closely associated with ERP include 1) Management Information System (MIS)/Business Intelligence (BI) which traditionally serves as Decision Support System (DSS) involving descriptive analytics, that provide information for management to take decisions (Demigha 2021); 2) Workflow Engines for automating tasks execution steps in a given business process; 3) Document Management System/Enterprise Content Management Systems for managing all types of documents and digital content across the organization; 4) Knowledge Base/Knowledge Management System which serves the purpose of user self-service for knowledge contribution and retrieval (i.e., knowledge sharing); 5) Identity Management Systems/Single Sign-On (SSO) across functional units; and 6) Enterprise Integration Patterns (EIP) for standardized integration with disparate in-house systems or even third-party solutions, as proposed by Hohpe (2003).

1.2. Methodology

In the first place, in line with the definitions of Grant and Booth (2009) we carried out a state-of-the-art literature review of recent academic and industry publications that speak to the challenges of AI adoption by organizations, to identify and categorize the published potential barriers to AI adoption. We primarily focused on publications after the release of ChatGPT in line with our goal of proposing an adoption approach that addresses contemporary barriers. We only accommodated publication earlier than the year 2022 where it gives perspective to the state-of-the-art. We then carried out a thematic review and selected for critical analysis, AI technology innovations which by their inherent features could be considered milestones in the advancement of AI technology applicability to organizations vis-à-vis ERP, as explained in our objectives above. Finally, using inductive reasoning, we synthesized an approach to AI adoption in organizations that address the observed barriers and implemented a tool that exemplifies our proposal.

2. Observed Barriers to Effective Adoption

Yang et al. (2024) defined AI readiness as a firm's capacity to deploy and use AI in ways that add value to the firm. This definition to a large extent captures what we refer to in this work as effective adoption, which is inclusive of both successful adoption mechanism and value derivation therefrom. In this section, we present our findings in literature on observed obstacles to effective adoption which often speaks to a firm's AI readiness. For clarity of identification, we have used subheadings in this section to categorize the barriers we encountered in our review.

2.1. Weak or Non-Existent Strategy

Prominent among the observed barriers is an adoption-strategy gap which in our opinion merits research attention as it could significantly affect organizations bottom-line. AI needs to be adopted for the right reasons say Neumann et al. (2024), to solve previously unsolved problems, and not for

 doi:10.20944/preprints202512.0428.v1

solving problems you first must create. Adoption without strategy alignment is evident in the findings by Shah et al. (2024) in a study across Europe, Asia-Pacific and the US, tagged Arm AI Readiness Index. While over 80% of organizational global leaders believe that it is urgent for their organizations to embrace AI and a significant number aligning their budget to match their belief, less than 40% have measurable key performance indicators (KPIs), underscoring the need for strategic roadmaps. Prior to the ChatGPT milestone, Hradecky et al. (2022) already observed among Western European organizations in the exhibition industry, that despite their belief that AI will increase efficiency, reduce costs and enhance customer experience, most organizations do not have a future strategy to implement AI. They encountered in their study a lack of vision and progressivity from CEOs who are primarily responsible for the strategic direction of the organization, as a key obstacle to AI adoption. Without an adequately clear sense of direction, AI in public organizations also remains limited both in terms of the number of applications used and in terms of the depth of their integration (Selten and Klievink 2024).

2.2. Poor Data Readiness and Privacy Concerns

Data readiness has been a common source of concern. Shah et al. (2024) observed that 60% of organizations they studied face challenges with data readiness. Reported concerns include data accuracy, privacy and trust (Gurjar et al. 2024; Madanchian and Taherdoost 2025; Praveen et al. 2025; Romeo and Lacko 2025; Williams 2024); robustness of the firm's data infrastructure for quality data availability, which may not be unconnected with available financial resources and management vision, especially among small firms (Yang et al. 2024); as well as data in silos (Shah et al. 2024) as opposed to data integration for adequate data pipelining.

2.3. Inadequate Integration with Existing Technology Stack

Weak integration of AI technology stack itself constitutes another form of barrier reported (Challapally et al. 2025; Madanchian and Taherdoost 2025; Williams 2024). Challapally et al. (2025) asserted that complexity of AI solutions integration and lack of fit with existing workflows lead to stalling of organizational adoption of custom AI solutions. Madanchian and Taherdoost (2025) identified expenses associated with such integration as a deterrent. Khanna and Bhusri (2025) similarly observed what they considered a key gap, that companies only use AI application for a very specific business function and do not adopt cross-disciplinary integration with the remaining operational functions, which results in inefficient AI implementation. In their study, up to 50% of respondents cited integration with existing IT systems as the most significant challenge.

2.4. Inadequate Human Knowledge Skills and Attitudes/Abilities (KSA)

Even though AI advancement has been touted to possibly result in significant employee disengagement, experience suggests that adequate employee training on the flip side, is required for AI to be adopted effectively (Leoni et al. 2024; Shah et al. 2024; Williams 2024; Yang et al. 2024). For example, Yang et al. (2024) observed that employees' digital skills and management's AI literacy impacts the AI readiness of the professional firms they studied. Likewise, Leoni et al. (2024) identified the presence of unqualified personnel to be one of the most significant barriers to implementing the appropriate AI tools. Esmaeilzadeh (2024) similarly emphasized the importance of human capital, continuous learning and a supportive environment for AI integration to thrive. For Shah et al. (2024), the People pillar faces a critical skills gap, with only 30% of the organizations they studied, offering comprehensive AI training. Gartner similarly asserted in their 2025 hype cycle report that mature organizations are battling to find skilled professionals (Khandabattu 2025).

2.5. Scalable and Secure Infrastructure Challenges

Security and scalability concerns occupy a place in AI implementation challenges (Khanna and Bhusri 2025; Williams 2024). This challenge may be even more pronounced as organizations see the need for AI within their network boundaries as an edge deployment, for security and privacy reasons, rather than send private data to cloud LLMs. Inadequate scalable and secure infrastructure at the edge

 doi:10.20944/preprints202512.0428.v1

thus exacerbates barrier to AI adoption (Shah et al. 2024). Sánchez et al. (2025) indicated that small and medium enterprises are particularly vulnerable to challenges in this respect, due to technological complexities in AI adoption, among others.

2.6. Ethical Governance Concerns

Even prior to the release of ChatGPT, Taeihagh (2021) drew attention to the importance of AI governance stating that understanding and managing risks posed by AI is crucial to realize the benefits of the technology. Shah et al. (2024) observed that ethical governance is underdeveloped, with only 35% of organizations studied addressing ethical concerns. Marocco et al. (2024) pointed out three kinds of ethical concerns - the implications of AI making life-or-death decisions, the risk of discrimination due to biased data or programming, and the ethical challenges of replacing humans with AI. They also asserted that managers with ethical concerns, become less curious about AI systems and consequently reduce their willingness to adopt. Rushing out AI is likely to backfire says Davies (2025) drawing on a survey among 900 professionals responsible for implementing AI across the UK, US and Canada in which a significant number of professionals were skeptical, expressing fear, pressure and potential risk to both employees and customers. For Dinculeana (2024), one of the most pressing ethical challenges is the potential for algorithmic bias as AI models trained on historical data can inadvertently reinforce existing biases in the data.

2.7. Regulatory Framework Lag

The rapid advancement of AI technologies often outpaces the development of regulatory frameworks (Dinculeana 2024). Praveen et al. (2025) identified regulatory complexities as a key barrier in AI adoption. Marocco et al. (2024) on a positive note observed that regulatory guidance helps reduce uncertainty, boosting managers' confidence in AI technologies. On the contrary, regulatory uncertainty hinders confidence in AI adoption (Poon et al. 2025). In this regard, Yang et al. (2024) distinguished between large and small firms. In a study of professional firms, they observed that large firms are more hindered by gaps in regulatory frameworks compared to other AI readiness factors like infrastructure.

2.8. Responsibility and Accountability Concerns

The term responsible AI emerged even prior to the first ChatGPT release, to refer to demand for a set of principles that ensure ethical, transparent, and accountable use of AI technologies consistent with user expectations, organizational values, and societal laws and norms (Mikalef et al. 2022). Along this line, the black-box nature of AI models is a source of discomfort, especially among managers who feel responsible for their assigned role in the organization. Such managers may be apprehensive about delegating authority to AI which led Marocco et al. (2024) to infer from their study, the importance of designing AI-based systems that enable managers interact with, modify and oversee AI-generated recommendations. To build trust in automated decision-making, managers may demand more explainability and interpretability in AI models (Procter et al. 2023; Romeo and Lacko 2025). AI should be made accountable, not only to the individual, but also to the organization says Procter et al. (2023). Unfortunately, when AI agents fail, responsibility is unclear (Kundaliya 2025).

2.9. Reliability Concerns

Reliability of AI-based systems has sometimes been called into question especially when it comes to taking decisions in organizations. While not overlooking the fact that AI systems are normally designed to improve, as learning systems, it is important to rise beyond hype and understand what the current level of AI is when planning adoption. Succinctly put by Stief (2025, p. 1), "we are currently between basic and strong AI". The often-assertive nature of output from Generative AI (even when it is hallucinating) seems to sometimes result in a situation where AI becomes crutch instead of tool for collaboration, a phenomenon which BetterUp Labs (2025) in collaboration with the Stanford social media lab refers to as *workslop*. They defined *workslop* as AI-generated content that appears good but lacks substance. Based on insights from their online survey of 1150 US desk workers in September 2025,

 doi:10.20944/preprints202512.0428.v1

they concluded that *workslop* makes individuals feel frustrated, confused and disengaged; teams waste cycles, duplicate efforts, and lose trust; and organizations lose time, are misled by false productivity and experience stalled AI adoption. They recommended that organizational leaders set guardrails to curb it, modelling thoughtful use of AI, and fostering the use of AI collaboratively rather than seeking to use it to avoid work. Similarly, in a study primarily among British workers, employees raised alarm over the rise of AI agents in the workplace, warning that they are unreliable, unresponsive to feedback, and in some cases creating more work instead of reducing it (Kundaliya 2025). In the same vein, Anderson et al. (2025, para. 8), while recognizing the benefits of thoughtful use of Generative AI in coding, drew attention to the hidden costs. Developers they interviewed indicated a number of problems that come with coding with AI – code duplications, integration problems, dependency conflicts, a lack of context awareness, etc. which can lead to the compounding of technical debt. Alluding to the seriousness of the situation, they further cautioned organizations to treat AI tool's tendency to increase technical debt as a strategic risk, not just an operational nuisance.

3. Milestone AI Innovations with Organizational Adoption Potential Vis-à-Vis ERP Integration

Here, we present our thematic review of AI innovations that we consider to be milestones based on our perception of their relevance to organizational usefulness and the clarity of problems that they address. The AI innovations we present in no particular order are not necessarily product-specific but are more of conceptual foundations, approaches or algorithms and in some cases associated tools, that have increasingly become more prominent since the advent of LLMs like ChatGPT and that we consider to have the potential to functionally and technically align, enhance or replace aspects of classical ERP, our chosen mature technology reference point.

3.1. Retrieval-Augmented Generation (RAG)

Before the first release of ChatGPT, Lewis et al. (2021) used the term Retrieval-Augmented Generation (RAG) to refer to the phenomenon in which they enhanced factual accuracy of large pre-trained language models in knowledge-intensive tasks, by combining pre-trained parametric memory (weights intrinsic to LLMs' neural network) with non-parametric memory (numerical representation of domain-specific data stored as indexed vector embeddings), for language generation. They successfully illustrated how retrieval from pre-trained parametric memory can be hot-swapped to update the model output, without requiring any retraining. This demonstrated RAG step quickly paved the way for application to LLMs like ChatGPT as they emerged. In organization context, the central idea is the potential to augment LLMs (public or private) with organization's internal data for more specialized and up-to-date responses. RAG has the potential to enhance the natural language interaction experience with ERP's Knowledge Base/Knowledge Management System.

Advances in embedding and retrieval models as well as vector databases have been instrumental in making RAG implementation more feasible hence, we present further information about them in the subsections below, for more informed adoption choices.

3.1.1. Embedding Models

Embedding models are machine learning algorithms that can be used to convert complex real-life digital objects into numeric vector representation known as dense vectors or embeddings, in a way that preserves context sensitive meaningful relationships between parts of the encoded object. Complex objects may be words, images, audio, video. The advances in embedding models have been a gamechanger that paved the way for modern Generative AI.

The breakthrough algorithm which popularized embeddings is word2vec, created in Google in 2013 by Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean (Bianchini 2025; Mikolov et al. 2013). It was adopted by some commercial organizations to power their recommendation engines. Shortly after word2vec was released, some researchers at Stanford created GloVes (Global Vectors for Word Representation) in 2014 and released a newer version in 2024 (Carlson et al. 2025), which

 doi:10.20944/preprints202512.0428.v1

could capture semantics broader than word2vec. Embedding algorithms needed to mature more to power major Generative LLMs like ChatGPT. Better contextual sensitivity was needed, among other improvements. Again, in Google, a team of researchers proposed Transformer in 2017, in a research paper titled "Attention is All You Need" (Vaswani et al. 2023) which paved the way for the creation of generative pre-trained transformers (GPT). Following this paper, Google researchers released bidirectional encoder representations from transformers (BERT) in 2018 for word embeddings (Devlin et al. 2019). Shortly afterwards, Liu et al. (2019) of Facebook released RoBERTa as an improved version of BERT. However, both were designed for single word embeddings. Reimers and Gurevych (2019) released Sentence-BERT (SBERT) with capability to generate embeddings for full sentences or phrases, an improvement over the single word limitation for sentence-level tasks.

Besides word and sentence-level tasks, embedding algorithms have also been created for images, audio and video. Dosovitskiy et al. (2020) of Google created Vision Transformer (ViT) architecture that applies transformer architecture to images split into 16x16 pixels as tokens, for more efficient pre-training. Radford et al. (2021) of OpenAI then applied ViT to create Contrastive Language–Image Pre-training (CLIP) that connects text and images. Similar to CLIP, Radford et al. (2022) still of OpenAI created Whisper that applies transformer architecture to speech recognition which can listen to and transcribe spoken language. Tong et al. (2022) of Nanjing University and Tencent AI Lab further extended the transformer architecture concept to videos, creating VideoMAE (Video Masked Encoders).

These transformer architecture-based model algorithms for text (SBERT), image (CLIP), speech (Whisper) and video (VideoMAE) constitute significant milestones in Generative AI. Pre-trained SBERT and CLIP models are available for use from the Hugging Face hub as part of python sentence_transformers package. Similarly, pre-trained models of both Whisper and VideoMAE are available from Hugging Face but from a different python package named transformers. Both packages are downloadable from python package repository, https://pypi.org/. NVIDIA NeMo Framework, a development platform for building custom generative AI models, supports these architectures in the cloud.

3.1.2. Retrieval Models

As AI adoption advances, retrieval is shifting from relying on statistical algorithms to neural models with the implied superior semantic capability (Zhu et al. 2025). This shift is important as retrieval efficiency can be quite dependent on understanding user intent which can be sometimes complicated. Short and ambiguous user queries make it difficult to precisely understand the user' intent (Zhu et al. 2025). Retriever models seek to address more effectively these challenges through query rewriting. An advanced way to carry out query rewriting is chain-of-thought (CoT) prompting which was first introduced by Wei et al. (2022). CoT breaks down prompting into a series of intermediate reasoning steps thus improving the ability of the model to perform accurately, especially with complex prompts that are best dealt with in steps. There are models in Hugging Face hub that are available for use via the transformers package and that support CoT. An "explain step by step" kind of statement accompanying the query makes such models perform best.

Although CoT is a marked improvement in the path towards retrieval accuracy, there is at least one more challenge that needs attention – context rot – which refers to increasingly unreliable performance of models as input length grows (Hong et al. 2025). In other words, "as the number of tokens in the context window increases, the model's ability to accurately recall information from that context decreases" (Anthropic 2025, para. 7). The term "context rot" appears to have been first used by a user named Workaccount2 in a comment thread on an article captured on Hacker News online (YCombinator Hacker News 2025). Researchers at MIT have proposed Recursive Language Models (RLMs) as an approach to addressing the context rot problem (Zhang 2025). RLM is new and is yet to be incorporated into a python package in https://pypi.org as at this writing but there is a basic implementation at https://github.com/alexzhang13/rlm. The more recently released DeepSeek-OCR

 doi:10.20944/preprints202512.0428.v1

has the potential to mitigate against context rot, with its demonstrated capability to compress long contexts through optical 2D mapping (Wei et al. 2025).

We also note that context rot is expectedly more associated with transformers that have to struggle with inefficiencies due to their quadratic attention complexities. This draws our attention to two other architectures which have emerged, that use different techniques for sequence modeling (e.g., conversion of input tokens to vector embeddings) namely RWKV (Receptance Weighted Key Value) and Mamba-2. Both to some extent replace the attention mechanism used by transformers with other more efficient mechanisms which operate with linear complexity 0(n) rather than quadratic complexity 0(n2) as obtainable with transformer architecture. RWKV takes advantage of the linear scaling of Recurrent Neural Network (RNN), combining it with the parallelization and scalability of transformers (Peng et al. 2023). Models based on the version RWKV-7, are available for download from Hugging Face repository e.g., https://huggingface.co/RWKV. Mamba-2 architecture is designed using state space duality (SSD) framework, a faster derivative of state space model (SSM) combined with transformer capabilities (Dao and Gu 2024). Hugging Face also has a collection of downloadable Mamba-2 models e.g., https://huggingface.co/collections/nvidia/nvidia-nemotron-v2 with pre-trained weights and quantization support. Mamba-2 also has the added advantage of being programmatically usable via the python package named transformers. As both may trade some bidirectional precision compared to transformers, such models may be best used in retrieval phase in hybrid with transformers. IBMs Granite-4 also can come in handy as an intrinsic Mamba-2/transformer architecture hybrid (Soule and Bergmann 2025) and are available in Hugging Face for enterprise use and with opensource license.

3.1.3. Vector Databases

Vector databases which store data as vectors date back to the 1980s and 1990s for spatial data and have become more popular as general purpose vector databases with the rise of RAG (Fitz 2023). Recent vector databases include those built purely as such, e.g., Milvus and those developed as extensions of existing SQL databases like PGVector and the better performing PGVectorscale as PostgreSQL extensions. The dense vector indexing for search using approximate nearest neighbors (ANN) algorithm and its variants has contributed significantly to the scalability of semantic search. Researchers in Facebook building on ANN created FAISS (Facebook AI Similarity Search) for fast in-memory vector search with GPU support (Douze et al. 2024; Johnson et al. 2017). Google researchers followed the improvement path with the release of ScaNN (Scalable NN) based on their developed algorithm, named Anisotropic Vector Quantization (Guo et al. 2019). Similar to FAISS, ScaNN is in-memory but unlike FAISS, it does not use GPU. Researchers at Microsoft created DiskANN which can work with SSDs at massive scale with reasonable performance, thus beating down cost (Subramanya et al. 2019). PGVectorscale supports DiskANN indexing which makes it superior to the basic PGVector extension of PostgreSQL.

3.2. GraphRAG

GraphRAG is essentially RAG across knowledge sources (Bianchini 2025) which was introduced by Larson and Truitt (2024), researchers in Microsoft, to enhance RAG accuracy performance. As explained by the Microsoft team, "GraphRAG, uses the LLM to create a knowledge graph based on private dataset. This graph is then used alongside graph machine learning to perform prompt augmentation at query time" (Larson and Truitt 2024, para. 3). In other words, GraphRAG enhances the knowledge base by complementing semantic context representation in basic RAG embeddings with relational context representation, based on entities and their relationships identified in the dataset. This enriches the accuracy of augmented responses.

The representation of the network of interrelated entities also known as knowledge graph is made up of nodes and edges. Each node represents an identified entity, and edges represent the relationships between the entities. Knowledge base involving GraphRAG shares chunking and embedding steps with standard RAG. For the former however, the chunks are subjected to a graph building stage which consists of extracting entities and their relations in the form of triplets (head, relation, tail) which are

 doi:10.20944/preprints202512.0428.v1

then stored in the graph database. For queries, this is particularly useful for multi-step knowledge retrieval (Han et al. 2025).

Python packages like Relik (https://pypi.org/project/relik/) can be used for knowledge graph extractions part of GraphRAG workflow. Tools like LightRAG (https://pypi.org/project/lightrag-hku/) and GraphRAG (https://pypi.org/project/graphrag/) cater for the whole workflow in association with suitable LLMs for embedding and for reranking at query time. If using LangChain as agentic AI framework, LangChain's inbuilt python modules LLMGraphTransformer and GraphCypherQAChain provide a native way to achieve knowledge graph generation and query, in association with suitable LLM or SLM, with stronger integration with the agents created.

3.3. Small Language Models Creation

Recently, some researchers from Nvidia strongly opined that small language models are the future of Agentic AI (Belcak et al. 2025). In their own words, "small language models (SLMs) are sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems and are therefore the future of agentic AI" (Belcak et al. 2025, p. 1). They consider most models with below 10 billion parameters in size to be SLMs. They further posit that concretely, SLMs have higher inference efficiency, fine-tuning agility, edge deployment suitability and parameter utilization. Zhang et al. (2025) similarly asserted that when sufficiently trained or fine-tuned on domain-specific data, SLMs have proven to be efficient solutions for target tasks. Mitigation of privacy and security risks by keeping sensitive data within secure network boundaries is yet another reason for SLM advocacy (Criddle and Murgia 2024; Zhang et al. 2025).

SLMs may be created directly or by shrinking (or distilling) existing LLMs. A number of SLMs created through shrinking of larger models are available for free online for download, from sites like https://huggingface.co/. It is more efficient to take off from such resources and further fine-tune as necessary for target specialization. The common shrinking techniques which include quantization, knowledge distillation and pruning have differing characteristics and advantages. Quantization involves reducing the number of bits required for each model's weights, resulting in models that consume less memory and storage space and capable of faster inference, thus favoring operations in a wider range of devices, including embedded devices (Gou et al. 2021; Lang et al. 2024). Knowledge distillation shrinks models by distilling the knowledge in a larger model (or an ensemble of models) onto a smaller (or single) model mimicking a teacher-student relationship (Gou et al. 2021; Hinton et al. 2015). Pruning focuses on the reduction of number of parameters to achieve model shrinking (Qian and Klabjan 2021). Unlike quantization which keeps all the parameters but simply reduces their numerical precision (e.g., from 32-bit floats to 8-bit integers), pruning drops entire parameters to achieve size reduction.

SLMs produced through knowledge distillation are generally most suitable where the goal is to further fine-tune the model for a given domain expertise because it has the highest probability of retaining the reasoning capability of the parent (teacher) model. The fine-tuned model can still be further subjected to quantization to improve inference efficiency. Quantization should be more effective than pruning because pruning has a higher risk of degrading reasoning as it removes entire parameters whereas quantization keeps them all and only reduces numerical precision.

3.4. Parameter-Efficient Fine-Tuning (PEFT)

In the preceding section on SLMs, we alluded to the place of model fine-tuning in the creation of specialized SLMs. Similar to distillation methods, fine-tuning techniques have also evolved. Parameter-efficient fine-tuning (PEFT) has evolved as a more efficient alternative to full fine-tuning which involves resource intensive updating of all weights (Abdullah et al. 2025; Wang et al. 2025). Essentially, PEFT involves the use of only some selected parameters for fine-tuning and categories of PEFT techniques differ based on parameter selection criteria. Wang et al. (2025) in their review of PEFT methodologies, identified three main categories – Addictive PEFT, Reparameterized PEFT and Selective PEFT – and other derived categories namely Hybrid PEFT, Quantization PEFT and Multi-task PEFT. Additive

 doi:10.20944/preprints202512.0428.v1

PEFT adds and architecturally integrates a small set of trainable parameters to the target pre-trained model. During fine-tuning, only these added parameters are adjusted. Specific additive methods include adapter, soft prompt (e.g. prefix-tuning and prompt-tuning), scale and shift (e.g., IA^3 which has become the conventional acronym for Infused Adapter by Inhibiting and Amplifying Inner Activations). Reparameterized PEFT involves the construction of a low-rank parameter matrix for fine-tuning. At inference time, the fine-tuned matrix is combined with the pre-trained parameters. Specific methods here include LoRA (Low-rank Adaptation) and other LoRA derivatives like DyLoRA, AdaLoRA, DoRA and IncreLoRA. Selective PEFT does not involve parameter addition but involves taking a small set of pre-trained parameters, for fine-tuning. Regarding the derived PEFTs, Hybrid PEFTs attempt to combine advantages of various PEFT methods. Quantization PEFT includes attempts to apply quantization to the PEFT methods for more computational and memory efficiency. For example, QLoRA (Quantized LoRA) involves working with quantized weights instead of the full pre-trained weights. Multi-task PEFT is essentially PEFT for multi-task learning. Examples include AdapterFusion and AdaMix (Adapter-based); SPoT and ATTEMPT (Soft Prompt-based); LoRAHub, L-LoRA and MOELoRA (LoRA-based).

Wang et al. (2025) in their survey of methodologies identifies LoRA and its derivatives as leading PEFT methodologies that consistently deliver high performance. The choice of the specific variant depends on the target use. AdaLoRA for example seems particularly suitable for domain fine-tuning for small to midsize models with strong accuracy and parameter efficiency. Adaptive rank allocation lets it focus more capacity on layers critical to the target domain semantics. In conclusion, we recommend taking off from an SLM generated from LLM via knowledge distillation (preserves reasoning capability of the LLM) and then carrying out fine-tuning using AdaLoRA. If additional shrinking is required, the fine-tuned model can be further subjected to quantization which will trade a little numerical precision for inference speed.

3.5. Document Understanding

Documents are substantive part of any organization where a lot of information pertaining to an organization's knowledge base resides. AI has the potential to move document information extraction beyond basic OCR to semantic knowledge discovery, applying machine learning at different points in the workflow including effective computer vision, even for partially legible physical documents.

Many documents exist in multimedia forms (e.g. combined text and images, audiovisual electronic documents, etc.). Knowledge extraction from such documents logically stands to benefit from AI model algorithms that are multimodal. Among the transformer-based algorithms we have discussed so far, only CLIP supports more than one content mode – text and images. The rest are unimodal. Multimodal language models can be used for unified document tokenization – printed or handwritten text, image, layouts, etc. An example is Qwen3-VL with distilled variants like Qwen3-VL-2B-Instruct, Qwen3-VL-2B-Thinking which can recognize broad range of entities in documents, carry out OCR with multiple language recognition, etc. They are available for free download from Hugging Face.

3.6. Agentic AI

The buzz around Generative AI paved the way for yet another buzz – Agentic AI – in the quest for AI based autonomous task execution. AI wave has evolved from basic predictive AI wave to Generative AI wave and then to Agentic AI wave (Falconer 2025b). The term Agentic AI seems to have been coined by Andrew Ng and used in a Deeplearning.ai online magazine, *The Batch*. In the episode dated April 10, 2024, he used the title "agentic AI design pattern" to label a link to a previous article in which he started to describe agentic design patterns (Ng 2024a,b). The term Agentic AI has since become conventional among industry players and academics. Acharya et al. (2025) refer to it as qualitative leap in the development of Artificial Intelligence, with capability to set complex intermediate goals, autonomously adapting to a changing and uncontrolled situation and autonomously managing their resources as the agent steers itself towards the goal set for it. Core to the technical capability of AI agents are reinforcement learning (RL), goal-oriented architecture and adaptive control mechanisms

 doi:10.20944/preprints202512.0428.v1

(Acharya et al. 2025). The key components in writing an agent include defining the environment for agent interaction (e.g. events), implementing reasoning logic (e.g. as in a model), enable actions and incorporating adaptation.

We advocate organizational capacity building or expert partnerships for agile creation of agents in line with business demand dynamics, in order to facilitate due control. Many frameworks have emerged to assist in the development of single or multi-agents that can collaborate. Examples include PydanticAI (https://ai.pydantic.dev/), Haystack (https://haystack.deepset.ai/), AutoGen (https://microsoft.github.io/autogen/stable/index.html) and Strands Agents (https://strandsagents.com), LangChain (https://github.com/langchain-ai/langchain) and CrewAI (https://github.com/crewAIInc/crewAI), all of which are opensource projects. Even though a careful choice of any of these frameworks could go a long way to facilitating Agent development, we believe it is important to be clear about the essence of any Agent, and be as flexible as possible in the development, as the need arises. Agents should fit into the organization and not the other way round.

Besides frameworks, some new protocols have emerged for standardization of Agent operations. Model Context Protocol (MCP) was created by Anthropic for the standardization of access to tools, data, contexts, outside the model itself (Anthropic 2024; Hou et al. 2025). A2A Communication for agent-to-agent communication was created by Google and launched by The Linux Foundation (2025) The Linux Foundation (2025), to pave the way for interoperability across differing technology platforms.

There seems to be growing advocacy for event-driven AI agents for better scalability, avoiding potential bottlenecks associated with point-to-point, albeit decentralized architectures, as may be experienced in microservices architecture. Manditereza (2025) for example sees such a scalability challenge with A2A protocol which relies on HTTP and gRPC's direct point-to-point architectures and advocates an event-driven architecture for agent-to-agent communication. Similarly, the future of Agentic AI is event-driven where EDA (event-driven architecture) acts as "central nervous system" for dynamic information flow, says Falconer (2025a,b) of Confluent. He posits this architecture as a natural fit for MCP. Along this line, Song (2025) of Alibaba categorized Agentic AI applications into two types based on their triggering mechanisms namely, user-triggered agents and system-triggered agents. He foresees a future where majority of agents operate in industrial setting as system-triggered agents which must run continuously, act autonomously and recover from failure without manual intervention. In this light, he proposes leverage on Apache Flink's capabilities in real-time distributed event processing, state management and exact-once consistent fault tolerance as framework for building such system-triggered event-driven AI agents. A project named Apache Flink Agents (https://github.com/apache/flink-agents) has been initiated in the Apache community for this purpose but still at an early stage.

Agents may also be broadly classified into two types – ReAct agent and Workflow agent – depending on the level of autonomy of the inherent workflow decisions. ReAct (short for Reasoning and Action) was first introduced by Yao et al. (2022) to describe the use of LLMs to generate in an interleaved manner, both reasoning traces and task-specific actions, the latter including access to external resources and tools as needed. In workflow terms, a ReAct type agent does not require a predetermined workflow path but autonomously figures out through a reasoning process the next steps (which may include iteration) based on previous step's outcome in relation to user-defined goals. The primary understanding of what an autonomous agent is aligns with this ReAct concept. However, from an organization's perspective, it is relatively easy to identify where discomfort could come from for managers who are particular about being in control of decisions as core to their responsibilities. Adopting ReAct thus calls for some basis for trustworthiness. Workflow agents on the other hand may be more connatural with the way organizations classically predetermine their workflow steps in handling data-driven tasks but with the added leverage on agentic autonomy in some predefined task execution steps. Apache Flink Agent framework supports both ReAct and Workflow agent types. The official documentation explicitly states that the Workflow agent paradigm is inspired by the need to

 doi:10.20944/preprints202512.0428.v1

orchestrate complex, multi-stage tasks in a transparent, extensible, and data-centric way, leveraging Apache Flink's streaming architecture.

3.7. Authentication, Authorization and Guardrails

We group these three concepts together because they all have in common the need for the organization to exercise some form of control on the use and operations of AI technology. On the one hand, authentication (the process of establishing that users are who they claim to be) and authorization (the process of giving users their due access rights only) are well established concepts in technology use in organizations. On the other hand, the use of guardrails is more recently being explored as a way to exercise control over the operations of autonomous agents.

Although authentication and authorization approaches are well-established, attempts to standardize them for AI agents are still at an early stage. Efforts have been made to extend existing authentication and authorization systems but "OAuth, mTLS (mutual transport layer security), RBAC/ABAC (role-based access control/attribute-based access control), and cloud-native IAM (identity and access management) solutions merely provide an entry point to securing AI agents as they are incapable of coping with the changing agent behaviors, ongoing agent execution, and delegation complexities" (Chinni 2025, p. 4). AI agents come with a new set of security risks like prompt-based manipulation or prompt injection (exploiting user input for unintended behavior), model pollution (targeting the corruption of model foundational knowledge e.g., with bias), institutional identity confusion and privacy leaks (Chinni, 2025; He et al., 2024). Model pollution and privacy leaks, for example, could occur when models are fine-tuned with user data if care is not taken. (Chinni 2025, p. 6) further asserted that such problems "require the next generation of guardrails: contextual filtering on prompts, dynamic privileges, agent-specific IAM, and effective human-in-the-loop controls". He et al. (2024, para. 7) proposed an authenticated delegation approach which keeps humans in the loop. They described their approach as a "process of instructing an AI system to perform a task that requires access to tools, the web, or computer environments in such a way that third parties can verify that (a) the interacting entity is an AI agent, (b) that the AI agent is acting on behalf of a specific human user, and (c) that the AI agent has been granted the necessary permissions to perform specific actions". As specific solutions emerge for applying classical authentication and authorization to Agents, it is clear that there needs to be a way that is dynamic enough to meet the dynamic nature of security threats associated with Agentic AI. Any successful initial authentication and authorization notwithstanding, the guardrail concept seems to provide a way for mitigating risks in a dynamic way.

In practical terms, guardrails are AI operations technology stack that help to "ensure that organization's AI tools, and their application in the business, reflect the organization's standards, policies, and values" (McKinsey 2024, para. 1). McKinsey further classified guardrails into five types – appropriateness, hallucination, regulatory-compliance, alignment and validation. Use of guardrails is thus a way for organizations to take responsibility for the governance of their AI operations, setting and implementing safe boundaries, when faced with risks associated with the black-box nature typical of language models. Well-implemented guardrails by definition should go a long way to boosting confidence in AI system, especially with respect to privacy, security and trustworthiness concerns. With particular focus on security, Dev et al. (2025) propose a threat modelling approach to building guardrails, to mitigate risks like confidentiality attacks to ML training data, sensitive data leakage from output, AI auto-hack/reward hack, model stealing/proxy ML models and reverse engineering, among others, as these threats are discovered.

Different innovators have created tools that could lead to the standardization of guardrail implementations. Notable among these is Guardrails-ai with an opensource codebase at https://github.com/guardrails-ai/guardrails. It is a Python framework for building applications made up of input/output guards that detect, quantify and mitigate the presence of specified risks.

 doi:10.20944/preprints202512.0428.v1

3.8. AI Operations Transparency: Observability, Explainability and Evaluability

With the clamor for responsible AI, the discomfort associated with the black-box nature of models along with the autonomy of agents, calls for a design-for-transparency framework which embraces observability and explainability. Closely related are reliability concerns that stretch transparency demand to the ability to actually evaluate the operations of AI. When the stakes are high, organizations may wish for visibility (observability) to enable it to steer the language models (LM) towards reliability (Carter 2023) and also appreciate the rationale for a given outcome (explainability) which could facilitate trustworthiness (Kastner et al. 2021). Explainable AI (XAI as it is commonly called) is about having explanation for why the model predicted something and why it is trustworthy (Mishra 2021). In the following subsections, we throw more light on each of these three aspects of AI operations transparency.

3.8.1. Observability

As black boxes that produce nondeterministic outputs based on natural language inputs, the vast range of possible outputs makes it impossible to exhaustively test all inputs and scenarios for quality assurance, as may be done with rule-based client applications (Carter 2023). The only way to monitor quality is to embed telemetry into the interaction environment and observe how users interact with LM at production time. The foundation for observability is thus data from usage footprints (Agrawal 2023). Traditionally, software, network and systems engineering rely on observability using a set of telemetry data types that has become conventionally referred to as MELT, an acronym for metrics (numeric measurements like CPU, GPU usage over time), events (discrete actions performed that cause state changes like user login), logs (detailed records of discrete events) and traces (request journeys including the successes, failures and responses) data types. Analogically, this can be applied to AI observability (Carter 2023; Koc et al. 2025). An IBM staff writer Badman suggests additional metrics which could be used to monitor AI language model's specific interaction quality, namely, token usage (operational expense watching), model drift (observe accuracy deterioration), response quality (tracking hallucination, factual accuracy, consistency of output for similar input, relevance of response to user input, latency) and responsible AI monitoring (monitor bias occurrence, Personally Identifiable Information – PII – in generated content, compliance with ethical guidelines, content appropriateness).

Telemetry may be implemented in proprietary ways. To prevent vendor locking, an opensource framework named OpenTelemetry was created with vendor-agnostic APIs, as an observability framework that brings all the data types together in one platform (Young and Parker 2024). Carter (2023) asserts that by using traces with OpenTelemetry, you can track information about user inputs, AI model outputs, and the result of any operations you perform on AI model outputs before showing a result to an end user. Specific telemetry design can vary from simple LM call with static prompt to more complex processes like dynamic prompt building, RAGs, agents chaining, etc. Operational workflow has evolved over time. For example, OpenTelemetry SDK compliant metrics can be exported for scraping by Prometheus from the exposed endpoint (typically /metrics in the URL) and visualized using tools like Grafana which can use Prometheus data source. In place of Prometheus, Jaeger, another open-source platform, can be used for traces.

3.8.2. Explainability

Unlike observability that is external to the model, explainability can be intrinsic to the machine learning algorithm itself though not exclusively (Mishra 2021). Depending on the algorithm, models may be inherently explainable or may require post-hoc explanatory methods. A model that is based on a simple linear regression between the input and output can be inherently interpreted as such. In more complex black-box situations, external, post-hoc methods are used to analyze the input and output, for explanations. Different methods with their respective python code libraries have emerged for this purpose e.g., SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations) which appear to be the two most popular (Salih et al. 2024). LIME based explanations are

 doi:10.20944/preprints202512.0428.v1

limited to a particular input instance that gives rise to a single prediction. SHAP covers a wider scope; besides specific instance explanation (i.e., local), it can be used to explain how the model behaves across all data (i.e., global). Both shap and lime libraries are available for use from the python package repository pypi but shap is actively maintained while lime appears not to be maintained anymore.

3.8.3. Evaluating AI Operations

As AI gains widespread adoption attention, an urgent question is how we evaluate the systems for effectiveness, safety, ethical compliance, etc. (Duranton 2024; Japkowicz and Boukouvalas 2024). The complexities involved in AI make human evaluation tedious and quest for efficiency has paved the way for the use of AI itself as AI evaluation judges (Yu 2025; Zhuge et al. 2024) among other ways. Beyond human-as-a-judge, the AI based evaluation paradigm has evolved from traditional metrics (like BLEU, ROUGE for NLP evaluation) to single LLM-as-a-judge and then to multi-agent-judges (e.g., debate and committee-based approaches) and finally to agent-as-a-judge (process-based evaluation approach) (Yu 2025). While LLM-as-a-judge only evaluates the final output, agent-as-a-judge evaluates intermediate steps as well as the final, thus enhancing informativeness (Yu 2025; Zhuge et al. 2024). Chen et al. (2025) also describes multi-agent-as-judge which involves using multiple LLMs for multi-dimensional evaluation, in a way analogous to multi-dimensional human evaluation, which they found to outperform single LLM-as-a-judge and better aligned with human expert's rating.

Innovators have created tools for evaluations which organizations can leverage. Notable examples include DeepEval (https://github.com/confident-ai/deepeval), an open-source tool, which can be used for more than 40 types of evaluations, along with trained judges (i.e., LLMs and also SLMs fine-tuned for evaluation purposes e.g., Selene-1-Mini-Llama-3.1-8B available on Hugging Face). Selene-1-Mini-Llama-3.1-8B is a highly performant multi-purpose evaluator which can be further fine-tuned for more specific domains (Alexandru et al. 2025).

3.8.4. Combined Tooling

Some innovations include all three (observability, explainability and evaluability) in the same tool, to various degrees. Examples include LangFuse (https://github.com/langfuse) and Phoenix (https://github.com/Arize-ai/phoenix/) both of which are open source and actively developed. In a complementary way, they can also integrate with the more specialized tools like shap libraries for explainability as well as DeepEval and models like Selene-1-Mini-Llama-3.1-8B, for more rigorous evaluation.

4. Synthesis of an Adoption Approach

In the foregoing section, we have highlighted AI innovations that could find their way into organizations, with potential to complement, enhance or perhaps replace different aspects of ERP. In this section we apply inductive reasoning to propose an adoption approach in the light of our enumerated adoption challenges and the state of AI innovations.

As a first step and take-off point in our inductive reasoning process, **we expand on Yang et al. (2024) and define organizational AI readiness as the organization's capacity and disposition to deploy and use AI technology tools in ethical, responsible and accountable ways that add value to the organization**. We believe that this definition sets an adequate compass for our goal of proposing effective AI adoption approach. It not only addresses value creation, but it also addresses behavioral contexts for achieving the target values. Our definition takes a human-centric behavioral approach to organizational AI in as much as organizations are human purpose-driven institutions in line with Chester Barnard's formal organization definition as "a system of consciously coordinated activities or forces of two or more persons" Barnard (1968, p. 81). So, the ethical, responsible and accountable ways embodied in our working definition, pertain to judgement about the human actors who may use AI tools as instruments to achieve their purposes. Our goal here is not to argue about whether machines can or cannot be ethical, responsible, or accountable. However, for the avoidance of doubt, it suffices

 doi:10.20944/preprints202512.0428.v1

to state that we take for granted that machines have no free will but are conditioned by human design and have no ontological possibility of deliberately going against such designs.

Next, as shown in Table 1, we map the keywords in our organizational AI readiness definition to our identified adoption challenges. In line with our objective to propose and exemplify an approach to effective AI adoption, we leverage on the same structure to suggest implementation approaches that could mitigate the challenges, in the light of the strengths and limitations of AI innovations that we have identified (see Column 3 of Table 1). Finally, we describe OAAD, which we designed and created to exemplify our suggestions, as well as to proactively facilitate both AI adoption readiness and further impact studies, an approach we have termed adoption experimentation.

4.1. Organizational AI Appliance Deployer (OAAD)

To facilitate our adoption experimentation agenda, we go the extra mile in this work to create an appliance named Organizational AI Appliance (OAA) that could optionally be used as platform for accelerating AI adoption in organizations, based on established principles. We had to decide what existing subsystems and key libraries to incorporate and what agentic AI framework to recommend and use for in-built prototypes. We further created a system for automatic generation and deployment of OAA which we refer to as organizational AI appliance deployer (OAAD) with optional use of Docker or Kubernetes for containerization of subsystems. OAAD securely glues all the subsystems for seamless internal calls. It consists of an environmental variables file named *setup.env* for custom configuration (e.g., cluster sizes, CPU/GPU resources, integration points, etc. with default values) of the OOA to be generated for a given organization; a bash script named *generate.sh* for generating a new OAA based on the settings in *setup.env*; as well as a *start.sh*, *stop.sh* and *upgrade.sh* scripts for administering the cluster. All generated specific environment files for each subsystem are encrypted and can only be administratively decrypted. The OAAD also includes a custom python module named *utils.py* which contains functions for secure interaction with various parts of subsystems, from custom python codes, including agents, created and deployed on the OAA, as required for achieving defined organizational goals. It also includes some organizational case implementations out-of-the-box for illustrative purposes. For the latter, we choose foundation models and agentic AI framework that we considered most adequate for our proposed approach, within the limits of the current stage of AI innovation advancements.

4.1.1. Choosing Subsystems

In selecting the technology stack, we sought to maximize the use of battle-tested subsystems which to a large extent use open standards, thus with higher feasibility of integration with existing ERP infrastructure and also more likely known to the existing organization's workforce. We also gave preference to open-source tools which are matured and production ready, to make OAAD applicable to a wide range of organizational sizes. Table 2 shows our choices of subsystems for different requisite functional roles and Figure 1 shows how they are glued together as OOA.

 doi:10.20944/preprints202512.0428.v1

Table 1. AI Readiness Keywords Mapping to Adoption Challenges and Operational Ideas.

Keywords	Challenges	Recommendations
Capacity to Deploy (Ability to implement a functional system)	Poor data readiness	1. Carry out data readiness assessment on data generated from operations a. Completeness: identify data repositories associated with all relevant operations (ERP, etc.). b. Governance: identify regulatory compliant access control policies and mechanisms associated with organizational data from operations. 2. Fill data capture gaps across operations using open standards for flexibility.
	Inadequate integration across organization	1. Carry out an operational data integration assessment a. Cross-query accessibility across data from all relevant organizational operations using relevant business cases. 2. Fill integration gaps across operations using relevant Enterprise Integration Patterns (EIPs) with open standards, for flexibility. 3. Implement a system for harmonizing authentication and authorization. 4. Implement systems that can simultaneously be used to solve current problems whether AI-enabled or not and also be used for further AI empowerment, for better guarantee of continuous integration. 5. Create and deploy event-driven agents triggered by output events from integrated systems
	Inadequate infrastructure scalability	1. Carry out infrastructure scalability assessment 2. Every choice of technology stack should have clear horizontal scalability architecture
Disposition to Deploy (Willingness to deploy a functional system)	Privacy concerns	1. Carry out data privacy audit for proposed technology. 2. Adopt alternatives that better guarantee due privacy e.g. local models in place of models external to the organization as data boundaries demand. 3. Establish guardrails for due privacy protection.
	Security concerns	1. Carry out threat modelling for proposed technology using emergent frameworks for AI, e.g., MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) for adversarial threats, NIST AI RMF for governance-oriented risks and OWASP Top 10 for LLMs for practical security vulnerabilities like prompt injection. 2. Establish guardrails for mitigating identified potential threats
	Ethical concerns	[Same recommendations as in Ethical Ways keyword below]
	Regulatory gap concerns	1. Review regulatory policies released pre-AI, but that could apply to AI e.g., GDPR and establish how they could apply to your organization. 2. Review AI-specific regulatory policies 3. Suggest regulatory policies to relevant authorities for gaps identified through available channels
Capacity to Use (Ability to use a functional system)	Knowledge/skills and aptitude gap	1. Establish/adopt skills matrix for employees. Categorize skills in accordance with the nature of work of the employees. For example, a. Technical b. Responsible and Ethical c. Non-technical and Cognitive 2. Carry out AI readiness appraisal for employees using established skills matrix 3. Identify skill gaps and recommend relevant training 4. Collaboratively design/recommend training for employees 5. Hire for skills gaps as necessary
Disposition to Use (Willingness to use a functional system)	Perceived reliability of AI tools for tasks execution (employees want assurance of reliability of outcome)	1. Influence perceived reliability by sharing evaluation results.
	Perceived usefulness (managers questioning usefulness to the bottom line)	1. Influence perceived usefulness by sharing evaluation results.
Ethical Ways (Mindful of avoiding negative impact on stakeholders and environment)	Ethical organizations feel responsible for their employees	1. Develop in-house capacity as recommended under "Knowledge/skills and aptitude gap".
	Ethical organizations want to guard against bias	1. Implement bias mitigation guardrails.
	Ethical organizations want to be more energy efficient	1. Observe and decommission inefficient options.

 doi:10.20944/preprints202512.0428.v1

Table 1. *Cont.*

Keywords	Challenges	Recommendations
Responsible Ways (Can take ownership of operations with a commitment mindset)	Explainability and interpretability concerns (managers want to know what to expect from a given operation and why and be in a position to responsibly communicate with other stakeholders)	1. Fine-tune SLMs in line with organizational goals 2. Establish a no-black-box policy for agents and guardrails a. Own both agent and guardrail logic. 3. Build in-house capacity; complement with external expertise if necessary but own the logic a. Prioritize intrinsic agent observability and explainability
	Reliability concerns (managers want to guard against AI errors)	1. Fine-tune SLMs in line with organizational goals 2. Establish a no-black-box policy for agents and guardrails a. Own both agent and guardrail logic i. Build in-house capacity; complement with external expertise if necessary but own the logic b. Prioritize intrinsic agent observability and explainability 3. Implement requisite guardrails 4. Establish evaluation-driven development a. Evaluate pre-production proof-of-concept b. Implement production time observability and continuous evaluation
	Regulatory concerns (managers want to guard against regulatory non-compliance)	1. Fine-tune SLMs in line with known regulatory demands 2. Establish a no-black-box policy for agents and guardrails a. Own both agent and guardrail logic. i. Build in-house capacity; complement with external expertise if necessary but own the logic evaluation b. Implement production time observability and continuous evaluation 3. Implement guardrails for regulatory compliance.
Accountable Ways (Can be held answerable for operations outcome)	Controllability concerns (managers want to interact with, modify and oversee AI operations to justify being accountable)	1. Fine-tune SLMs in line with organizational goals 2. Establish a no-black-box policy for agents and guardrails a. Own both agent and guardrail logic. i. Build in-house capacity; complement with external expertise if necessary but own the logic b. Incorporate human-in-the-loop agentic workflows. c. Implement requisite control guardrails
	Reliability concerns (managers want to be able to validate and approve outcome of AI operations)	1. Fine-tune SLMs in line with organizational goals 2. Establish a no-black-box policy for agents and guardrails a. Own both agent and guardrail logic i. Build in-house capacity; complement with external expertise if necessary but own the logic b. Prioritize intrinsic agent observability and explainability 3. Implement requisite guardrails 4. Establish evaluation-driven development a. Evaluate pre-production proof-of-concept b. Implement production time observability and continuous evaluation
	Regulatory concerns (managers want to validate regulatory compliance)	1. Fine-tune SLMs in line with organizational regulatory demands 2. Establish a no-black-box policy for agents and guardrails a. Own both agent and guardrail logic. i. Build in-house capacity; complement with external expertise if necessary but own the logic evaluation 3. Implement guardrails for regulatory compliance.
Value Creation (Generating identifiable and quantifiable benefits)	Weak or Non-existent Strategic alignment (executives want clearer picture of how AI can impact their bottom line)	1. Define goals 2. Develop an impact evaluation culture and practice 3. Develop in-house quality criteria for use-cases
	Hype-driven adoption risk among early adopters (inflated expectations preceding realistic outcomes)	1. Develop an impact evaluation culture and practice.

 doi:10.20944/preprints202512.0428.v1

Table 2. Constituent Subsystems of OAAD.

Functional Role	Subsystem	Justification
1. Enterprise Integration (data, process, etc.) 2. Data pipelining 3. Event streaming	Apache Kafka (https://kafka.apache.org/)	1. In-built scale-out architecture with consumer partitions 2. High-availability clustering with Kraft 3. Strong EIP support with Apache Camel Integration 4. Mainstream adoption 5. Backed by Apache Foundation 6. Active maintenance development
1. Stateful computations over data streams 2. Machine learning tasks in stateful computations 3. Emerging framework for Agentic AI	Apache Flink (https://flink.apache.org/)	1. In-built scale-out architecture with multiple task managers and job multiple partitioning support 2. High-availability clustering 3. Dynamic extensibility with runtime deployment of new tasks including those based on suitable agentic AI frameworks 4. Mainstream adoption 5. Backed by Apache Foundation 6. Active maintenance development
1. Service Orchestration and Workflow 2. Emerging framework for Agentic AI	Apache Airflow (https://airflow.apache.org/)	1. scale-out architecture with multiple executor clusters like Celery, Kubernetes 2. High-availability clustering with multiple Airflow workers and Redis broker. 3. Plugin architecture for implementing new process automations including those based on suitable agentic AI frameworks 4. Support for Human-in-the-Loop service orchestration workflows 5. Mainstream adoption 6. Backed by Apache Foundation
Object store for 1. metadata 2. local AI models 3. state data 4. files in general	Minio (https://www.min.io/)	1. In-built robust scale-out architecture 2. In-built high-availability clustering 3. Mainstream adoption 4. Backed by MinIO, Inc.
Multiple storage types on same platform 1. Structured data store 2. Vector embeddings store 3. Graph store	PostgreSQL (https://www.postgresql. org/) with PG vector scale extension (https://github.com/ timescale/ pgvectorscale) and with Apache Age extension (https://github.com/ apache/age)	1. Scale-out architecture e.g., with Citus shards 2. High-availability clustering e.g., with Patroni and Etcd 3. Mainstream adoption 4. PostgreSQL is backed by a strong community development (more than 20 years) 5. PG Vector scale backed by Tiger Data 6. Apache Age backed by Apache Foundation
1. Scalable AI compute engine for delegated ML Workloads (optional)	Python Ray (https://www.ray.io/)	1. In-built scale out architecture with head node and multiple worker nodes 2. In-built high-availability cluster of head node and multiple worker nodes 3. Backed by Anyscale Inc.
1. Traces, evaluations, prompt management and metrics to debug and improve Language Model application	LangFuse (https://langfuse.com/)	1. Combined Observability, Explainability and Evaluation 2. Based on OpenTelemetry and can consume OpenTelemetry traces from other providers like DeepEval 3. Supports use of Clickhouse for more performance in handling of traces 4. Backed by Y Combinator
1. Systems monitoring and alerting	Prometheus (https://prometheus.io/)	1. Flexible monitor for applications, systems, and services 2. Backed by Linux Foundation
1. Data visualization for system monitoring	Grafana OSS (https://grafana.com/oss/ grafana/?plcmt=oss-nav)	1. Ease of integration with Prometheus data sources 2. Backed by Grafana Labs
1. Reverse Proxies, Load balancing	Nginx (https://nginx.org/) and HAProxy (https://www.haproxy.org/)	1. High performance load balancing and reverse proxy 2. In active maintenance 3. Nginx is backed by F5, Inc. and a strong and active open-source development community 4. HAProxy is backed by HAProxy Technologies and a strong and active open-source development community

 doi:10.20944/preprints202512.0428.v1

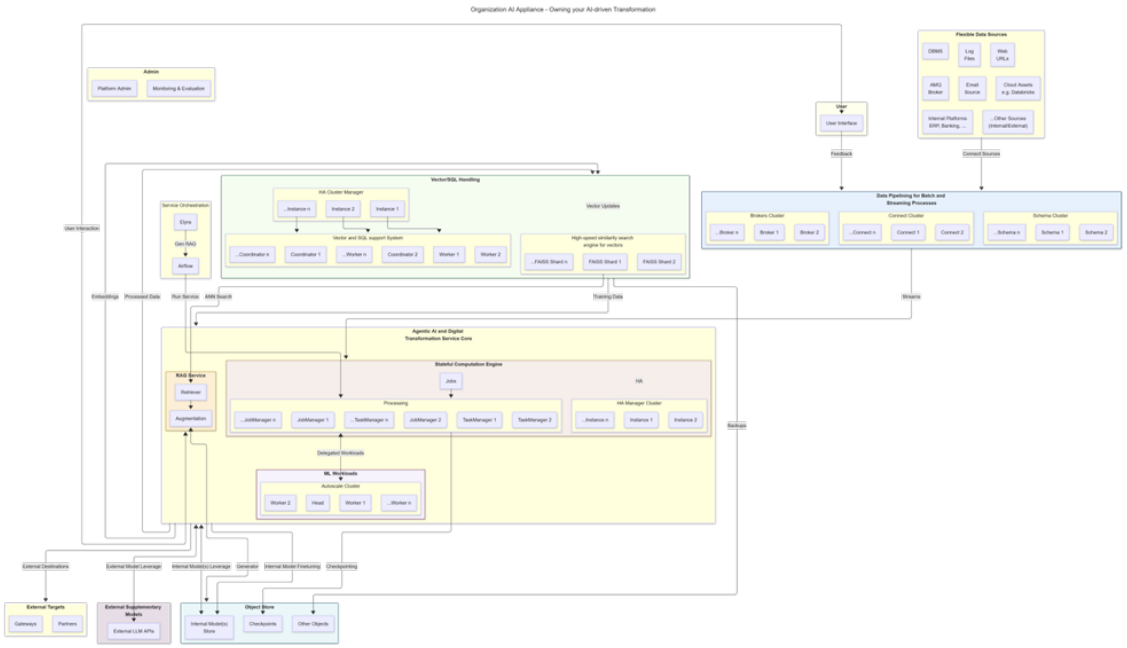

Figure 1. Schematic Diagram of Subsystems Interactions.

4.2. Choosing Foundational Models

By foundational models we refer to those SLMs that are downloaded and kept in the local object store for various functional purposes on OAA, as shown in Table 3. They can be used as is or after further purposive fine-tuning. They can also be replaced with more suitable models as more effective and efficient innovations emerge.

4.3. Choosing Key Python Libraries

By key libraries, we refer here to those python packages that we have chosen to be core to the operations of OAA. OAAD declares them as requirements, out of the box. They can be replaced with newer versions for upgrade. Table 4 shows our choices for various functional requirements.

4.4. Choosing Agentic AI Framework

In choosing agentic AI framework for out-of-the-box illustrations, we deferred to innovations that we consider most compatible with our proposed subsystems and that could also readily facilitate agents' development in line with our recommended no-black-box policy, to boost adoption confidence. To be closer to an objective comparison, we asked Grok4 to rank, with justification, a number of known agentic frameworks in terms of suitability for the following goals:

- Use of local SLMs (external LLMs as supplementary).
- Locally run LangFuse for observability, eval, metrics, etc., as well as use of python libraries like shap and judge models like https://huggingface.co/AtlaAI/Selene-1-Mini-Llama-3.1-8B
- Kafka and Flink-based data pipelines for real-time data
- Agents as airflow plugins or agents as flink UDF as may be necessary for no real-time or real-time data pipelining
- GPU usage support
- Guardrails using https://github.com/guardrails-ai/guardrails
- RAG/GraphRAG using pgvectorscale for vector embeddings and Apache age for graph databases and integrated tools for knowledge graph creation
- Model fine-tuning - batch and real-time - with AdaLoRA or QLoRA
- Locally run Google's embedding_gemma from hugging face for embeddings generation
- Locally run Granite-4 (or Nvidia-nemotron-v2) for Retrieval QA
- Human-in-the-loop workflow support

 doi:10.20944/preprints202512.0428.v1

- Enterprise SSO support
- "No black-box" policy
- Openness
- Battle-tested

LangChain and LangGraph combination had the highest average score of 8.8/10 compared to Haystack (8.0/10), Airflow AI SDK (8.2/10), Flink Agents (7.2/10), Pydantic AI (7.8/10), Crew AI (6.9/10), AutoGen (6.6/10), Strands Agents (6.6/10). We therefore adopted the duo as our agentic AI framework for our OAA. In addition to this, we adopted Apache Flink Agents framework (as an experimental alternative on maturity watchlist), because of its 100

Table 3. Constituent Subsystems of OAAD.

Functional Role	SLM	Justification
Vector embeddings generation	Embedding Gemma (https://huggingface.co/google/embeddinggemma-300m)	1. Highly performant and efficient. 2. Suitable for low resource environment 3. Suited for fine-tuning on specialized domains
QA Retrieval and Graph Extraction	Granite-4 (https://huggingface.co/ibm-granite/ granite-4.0-1b) (Smaller quantized versions are available at https://huggingface.co/models?other=base_model:quantized:ibm-granite/ granite-4.0-1b) or NVIDIA-Nemotron-Nano-9B-v2 (https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-9B-v2) (Smaller quantized versions are available at https://huggingface.co/models?other=base_model: quantized:nvidia/NVIDIA-Nemotron-Nano-9B-v2)	1. Hybrid Mamba-2 and Self-attention for efficiency 2. Suitable for low resource environments 3. Suited for fine-tuning on specialized domains
OCR and knowledge extraction from documents	Qwen3-VL-2B-Instruct (https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct) [Smaller quantized versions are available at https://huggingface.co/models?other=base_model:quantized:Qwen/Qwen3-VL-2B-Instruct]	1. Optimized for document understanding
Evaluation: Small language model-as-a-judge (SLMJ)	Selene-1-Mini-Llama-3.1-8B (https://huggingface.co/AtlaAI/Selene-1-Mini-Llama-3.1-8B) [Smaller quantized versions are available at https://huggingface.co/models?other=base_model:quantized:AtlaAI/Selene-1-Mini-Llama-3.1-8B]	1. Highly performant as can be seen on https://huggingface.co/spaces/allenai/reward-bench

 doi:10.20944/preprints202512.0428.v1

Table 4. Key Python Libraries of OAAD

Functional Role	Library	Justification
Evaluation of model interactions	DeepEval (https://pypi.org/project/deepeval/)	1. Comprehensive evaluation 2. Actively maintained
Explainability of agent operations	Shap (https://pypi.org/project/shap/)	1. Covers a wide scope; besides specific instance explanation (i.e. local), it can be used to explain how the model behaves across all data (i.e. global) Actively maintained
Guardrails creation	Guardrails-ai (https://pypi.org/project/guardrails-ai/)	1. Hardly any significant open competitor 2. Actively maintained
Model fine-tuning	PEFT (https://pypi.org/project/peft/) and Other packages (https://pypi.org/project/transformers/) (https://pypi.org/project/ accelerate/) (https://pypi.org/project/bitsandbytes/) (https://pypi.org/project/ datasets/)	1. For parameter efficient model fine-tuning 2. Library contains target AdaLoRA or QLoRA (if fine-tuning quantized model) 3. Actively maintained
GraphRAG 1. Embeddings generation 2. Graph generation	LangChain Ecosystem (https://pypi.org/project/ langchain/) (https://pypi.org/project/langchain-huggingface/) (https://pypi.org/project/sentence-transformers/) (https://pypi.org/project/langchain-postgres/) (https://pypi.org/project/langchain-community/) and Other packages (https://pypi.org/project/transformers/) (https://pypi.org/project/ accelerate/) (https://pypi.org/project/bitsandbytes/) (https://pypi.org/project/ psycopg/)	1. LangChain ecosystem aligns with our choice of LangChain and 2. LangGraph for agentic AI framework.
Agentic AI framework	[GraphRAG libraries above] + LangGraph (https://pypi.org/project/langgraph/) (https://pypi.org/project/torch/)	1. Primary justification is as shown in the subsection on Choosing Agentic AI Framework 2. Actively maintained
Agentic AI framework (experimental alternative on maturity watchlist)	Flink Agent (https://github.com/apache/flink-agents)	1. Uses LangChain4J under the hood, which follows the philosophy of LangChain but for Java. 2. Best aligned with our recommended Kafka/Flink for data pipelining

5. Conclusion and Recommendations

In this work, we set out to inductively propose an approach to effective AI adoption by organizations in a way that simultaneously facilitates longitudinal impact studies, an approach we refer to as

 doi:10.20944/preprints202512.0428.v1

adoption experimentation. To achieve this overarching goal, we began with a state-of-the-art review of known challenges with AI adoption by organizations. We categorized these challenges into Weak or Non-Existent Strategy, Poor Data Readiness and Privacy Concerns, Inadequate Integration with Existing Technology Stack, Inadequate Human Knowledge Skills and Attitudes/Abilities, Scalable and Secure Infrastructure Challenges, Ethical Governance Concerns, Regulatory Framework Lag, Responsibility and Accountability Concerns as well as Reliability Concerns. As a first step in inductive reasoning, we drew on our state-of-the-art review and defined organizational AI readiness as the organization's capacity and disposition to deploy and use AI technology tools in ethical, responsible and accountable ways that add value to the organization. We also traced the evolution of AI technologies that could apply to organizations vis-à-vis ERP.

We inductively mapped the adoption challenges we identified to the keywords in our organizational AI readiness definition and to our recommended adoption approaches to mitigate each challenge identified (see Table1). We further exemplified our recommendations by creating OAAD as a tool for generating integrated AI appliances in organizations. Our core recommendations may be itemized as follows:

1 Take an integrated platform approach to AI technology adoption akin to experience with ERP. The integrated system should include clearly defined points of data integration with existing technology solutions as exemplified by OAAD, especially with ERP that typically has wide stakeholder reach in organizations.
2 Establish an approach to authentication and authorization that harmonizes with existing infrastructure.
3 Have preference for integrated platform that seamlessly enables you to solve dynamic business problems, whether your digital solutions require AI or not.
4 Your integrated infrastructure should include a flexible way to store, fine-tune and use local language models with a privacy-first philosophy, complemented by external LLMs.
5 Build IT capacity (in-house and/or consultants) to identify strengths and limitations of available models for in-house adoption and monitor progressive evolution of model algorithms as limitations are overcome. In other words, know when and why you should upgrade.
6 Regarding Agentic AI, operate with a no-black-box philosophy.
7 Strategically identify and implement standard guardrails for security, ethical and regulatory compliance.
8 Adopt agentic frameworks and tools that support fine-grained observability, explainability and evaluability.
9 Build IT capacity (in-house and/or consultants) that enables you dynamically create agents in response to dynamic business needs and deploy on your integrated platform.
10 In creating agents, use agents workflow type as default for more control, rather than ReAct type agents. Use the latter when you see value in giving the agents workflow decision autonomy.
11 Strategically identify where you can use AI agents to enhance, replace or complement your existing technology solutions.
12 Develop an impact research culture for continuous improvement.

5.1. Limitations and Future Work

We limited our reference to existing organizational technology types to ERP. AI adoption studies could be carried out as well in relation to other well-established technology types e.g., office productivity tools which are also common to organizations or mission-critical technology solutions which can vary with organizations' type of product or service offerings. We also did not go into detail of potential AI transformation of each of the functional features of ERP. Another limitation is the fact that we did not exhaust the possible list of innovations (algorithms, models, libraries, etc.) examples.

We also recommend that academics play a more proactive and umpire role in addressing issues surrounding AI adoption by designing longitudinal impact studies of adoption on various organiza-

 doi:10.20944/preprints202512.0428.v1

tional goals, with requisite theoretical frameworks design, in partnership with organizations. This proactive approach to research engagement, in our opinion, could lead to better informed adoption decisions by organizations.

References

A. A. Abdullah, A. Zubiaga, S. Mirjalili, A. H. Gandomi, F. Daneshfar, M. Amini, A. S. Mohammed, and H. Veisi. Evolution of Meta's Llama models and parameter-efficient fine-tuning of large language models: a survey. Technical report, October 2025. URL https://arxiv.org/pdf/2510.12178v1. Preprint.

D. B. Acharya, K. Kuppan, and B. Divya. Agentic AI: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 13:18912–18936, 2025. doi: 10.1109/ACCESS.2025.3532853.

Monika Agrawal. *Implementing enterprise observability for success: strategically plan and implement observability using real-life examples*. Packt Publishing Ltd., 1 edition, 2023.

A. Alexandru, A. Calvi, H. Broomfield, J. Golden, K. Dai, M. Leys, M. Burger, M. Bartolo, R. Engeler, S. Pisupati, T. Drane, and Y. S. Park. Atla Selene Mini: A general-purpose evaluation model. Technical report, January 2025. URL https://arxiv.org/pdf/2501.17195v1. Preprint.

E. Anderson, G. Parker, and B. Tan. The hidden costs of coding with generative AI. *MIT Sloan Management Review*, 67(1), 2025. doi: 10.63383/HADW7619.

Anthropic. Introducing the model context protocol, November 25 2024. URL https://www.anthropic.com/news/model-context-protocol.

Anthropic. Effective context engineering for AI agents, September 29 2025. URL https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents.

Adam Badman. How observability is adjusting to generative AI. https://www.ibm.com/think/insights/observability-gen-ai. Retrieved October 21, 2025.

Chester I. Barnard. *The functions of the executive*. Harvard University Press, 1968.

P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. Muralidharan, Y. C. Lin, and P. Molchanov. Small language models are the future of agentic AI. Technical report, June 2025. URL https://arxiv.org/pdf/2506.02153. Preprint.

BetterUp Labs. Workslop: The hidden cost of AI-generated busywork, September 2025. URL https://www.betterup.com/workslop.

Francesco Bianchini. Retrieval-augmented generation. In F. De Luzi, F. Monti, and M. Mecella, editors, *Engineering Information Systems with Large Language Models*, pages 139–172. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-92285-5_7.

R. Carlson, J. Bauer, and Christopher D. Manning. A new pair of GloVEs. Technical report, July 2025. URL https://arxiv.org/pdf/2507.18103. Preprint.

Paul Carter. *Observability for large language models: understanding and improving your use of LLMs*. O'Reilly Media, Inc., first edition, 2023.

A. Challapally, C. Pease, R. Raskar, and P. Chari. The GenAI divide STATE OF AI IN BUSINESS 2025, July 2025. URL https://example.com/genai-divide-2025.

J. Chen, Y. Lu, X. Wang, H. Zeng, J. Huang, J. Gesi, Y. Xu, B. Yao, and D. Wang. Multi-agent-as-judge: Aligning LLM-agent-based automated evaluation with multi-dimensional human evaluation. Technical report, July 2025. URL https://arxiv.org/pdf/2507.21028. Preprint.

R. Chinni. Authentication and authorization for AI agents and risks involved. TechRxiv, July 19 2025. URL https://www.techrxiv.org/users/939863/articles/1310017-authentication-and-authorization-for-ai-agents-and-risks-involved.

Cristina Criddle and Madhumita Murgia. AI pioneers turn to small language models: Technology less complex products target clients with cost and data privacy concerns. *The Financial Times (London Ed.)*, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of Machine Learning Research*, volume 235, pages 10041–10071, 2024.

Victoria Davies. Rushing out AI likely to backfire, report, September 25 2025. URL https://www.computing.co.uk/news/2025/ai/rushing-out-ai-likely-to-backfire-report.

S. Demigha. Decision support systems (DSS) and management information systems (MIS) in today's organizations. In *European Conference on Research Methodology for Business and Management Studies*, pages 92–100, 2021. doi: 10.34190/ERM.21.062.

 doi:10.20944/preprints202512.0428.v1

J. Dev, N. B. Akhuseyinoglu, G. Kayas, B. Rashidi, and V. Garg. Building guardrails in AI systems with threat modeling. *Digital Government (New York, N.Y. Online)*, 6(1):1–18, 2025. doi: 10.1145/3674845.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://arxiv.org/pdf/1810.04805.

M.-E. Dinculeana. Transforming banking customer service: A detailed exploration of AI adoption with lessons from european countries. *Finance: Challenges of the Future*, 1(26):138–159, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021 - 9th International Conference on Learning Representations*, 2020.

Matthijs Douze, Alexandr Guzhva, Chengda Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. Technical report, January 2024.

Sylvain Duranton. Beyond accuracy: The changing landscape of AI evaluation. Forbes, March 14 2024. URL https://www.forbes.com/sites/sylvainduranton/2024/03/14/beyond-accuracy-the-changing-landscape-of-ai-evaluation/.

Pouyan Esmaeilzadeh. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. *Artificial Intelligence in Medicine*, 151:102861, 2024. doi: 10.1016/j.artmed.2024.102861.

Sean Falconer. The future of AI agents is event-driven. Medium, March 12 2025a. URL https://seanfalconer.medium.com/the-future-of-ai-agents-is-event-driven-9e25124060d6.

Sean Falconer. A guide to event-driven design for agents and multi-agent systems, 2025b.

Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative AI. *Business & Information Systems Engineering*, 66(1):111–126, 2024. doi: 10.1007/s12599-023-00834-7.

Tom Fitz. Vector databases - what's our vector, victor? Zenoss Blog, December 21 2023. URL https://www.zenoss.com/blog/ai-explainer-whats-our-vector-victor.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. doi: 10.1007/s11263-021-01453-z.

Maria J. Grant and Andrew Booth. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2):91–108, 2009. doi: 10.1111/j.1471-1842.2009.00848.x.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *37th International Conference on Machine Learning (ICML 2020)*, pages 3845–3854, 2019. Two entries with same PDF – using same eprint for both.

K. Gurjar, A. Jangra, H. Baber, M. Islam, and S. A. Sheikh. An analytical review on the impact of artificial intelligence on the business industry: Applications, trends, and challenges. *IEEE Engineering Management Review*, 52(2):84–102, 2024. doi: 10.1109/EMR.2024.3355973. URL https://doi.org/10.1109/EMR.2024.3355973.

H. Han, H. Shomer, Y. Wang, Y. Lei, K. Guo, Z. Hua, B. Long, H. Liu, and J. Tang. RAG vs. GraphRAG: A systematic evaluation and key insights. Technical report, February 2025. URL https://arxiv.org/pdf/2502.11371. Preprint.

Y. He, E. Wang, Y. Rong, Z. Cheng, and H. Chen. Security of AI agents. Technical report, June 2024. v2.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. Technical report, 2015.

Gregor Hohpe. *Enterprise integration patterns: designing, building, and deploying messaging solutions*. Addison-Wesley, 1 edition, 2003.

K. Hong, A. Troynikov, and J. Huber. Context rot: How increasing input tokens impacts LLM performance. Chroma Research, 2025. URL https://research.trychroma.com/context-rot.

Tim Hornyak. Agentic AI is here—but are we ready? *Research Technology Management*, 68(5):57–58, 2025. doi: 10.1080/08956308.2025.2532320.

X. Hou, Y. Zhao, S. Wang, and H. Wang. Model context protocol (MCP): Landscape, security threats, and future research directions. Technical Report 1, 2025. URL https://arxiv.org/pdf/2503.23278.

 doi:10.20944/preprints202512.0428.v1

D. Hradecky, J. Kennell, W. Cai, and R. Davidson. Organizational readiness to adopt artificial intelligence in the exhibition sector in western europe. *International Journal of Information Management*, 65:102497, 2022. doi: 10.1016/J.IJINFOMGT.2022.102497.

Kevin Huang. *Agentic AI: Theories and Practices*. Springer, 1 edition, 2025. doi: 10.1007/978-3-031-90026-6.

Nathalie Japkowicz and Zois Boukouvalas. *MACHINE LEARNING EVALUATION: Towards Reliable and Responsible AI*. Cambridge University Press, 2024. doi: 10.1017/9781009003872.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2017. doi: 10.1109/TBDATA.2019.2921572.

Lukas Kastner, Miriam Langer, Viktoria Lazar, Annika Schomacker, Timo Speith, and Stefan Sterz. On the relation of trust and explainability: Why to engineer for trustworthiness. In *Proceedings of the IEEE International Conference on Requirements Engineering*, pages 169–175, 2021. doi: 10.1109/REW53955.2021.00031.

H. Khandabattu. The 2025 hype cycle for artificial intelligence goes beyond GenAI. Gartner, July 8 2025. URL https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence.

A. Khanna and A. Bhusri. Navigating AI adoption - an "AI verse" framework for enterprises. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 1488–1491, 2025. doi: 10.1109/CAI64502.2025.00279.

V. Koc, J. Verre, D. Blank, and A. Morgan. Mind the metrics: Patterns for telemetry-aware in-ide AI application development using the model context protocol (MCP). Technical report, June 2025. URL https://arxiv.org/pdf/2506.11019v1. Preprint.

D. Kundaliya. UK workers say AI agents are "unreliable", September 25 2025. URL https://www.computing.co.uk/news/2025/ai/uk-workers-say-ai-agents-are-unreliable.

J. Lang, Z. Guo, and S. Huang. A comprehensive study on quantization techniques for large language models. Technical report, November 2024. URL https://arxiv.org/pdf/2411.02530v1. Preprint.

James Larson and Sarah Truitt. GraphRAG: Unlocking LLM discovery on narrative private data. Microsoft Research Blog, February 13 2024. URL https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/.

L. Leoni, G. Gueli, M. Ardolino, M. Panizzon, and S. Gupta. AI-empowered KM processes for decision-making: empirical evidence from worldwide organisations. *Journal of Knowledge Management*, 28(11):320–347, 2024. doi: 10.1108/JKM-03-2024-0262.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *ArXiv.org*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Peter G. Allen. RoBERTa: A robustly optimized BERT pretraining approach. Technical report, 2019.

Mitra Madanchian and Hamed Taherdoost. Barriers and enablers of AI adoption in human resource management: A critical analysis of organizational and technological factors. *Information (Basel)*, 16(1):51, 2025. doi: 10.3390/info16010051.

K. Mandidereza. A2A for enterprise-scale AI agent communication: Architectural needs and limitations. HiveMQ Blog, August 25 2025. URL https://www.hivemq.com/blog/a2a-enterprise-scale-agentic-ai-collaboration-part-1/.

S. Marocco, B. Barbieri, and A. Talamo. Exploring facilitators and barriers to managers' adoption of AI-based systems in decision making: A systematic review. *AI*, 5(4):2538–2567, 2024. doi: 10.3390/ai5040123.

Bernard Marr. A short history of ChatGPT: How we got to where we are today. Forbes, May 19 2023. URL https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/.

Ronald E. McGaughey and Angappa Gunasekaran. Enterprise resource planning (ERP): Past, present and future. *International Journal of Enterprise Information Systems*, 3(3):23–35, 2007. doi: 10.4018/jeis.2007070102.

McKinsey. What are AI guardrails?, November 14 2024. URL https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-ai-guardrails.

McKinsey. The state of AI: Global survey, March 12 2025. URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

J. P. Mendes, M. Marques, and C. Guedes Soares. Risk avoidance in strategic technology adoption. *Journal of Modelling in Management*, 19(5):1485–1509, 2024. doi: 10.1108/JM2-10-2023-0221.

Patrick Mikalef, Kieran Conboy, Jon Erik Lundström, and Aleš Popovič. Thinking responsibly about responsible AI and "the dark side" of AI. *European Journal of Information Systems*, 31(3):257–268, 2022. doi: 10.1080/0960085X.2022.2026621.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

Prateek Mishra. *Practical explainable AI using Python: artificial intelligence model explanations using Python-based libraries, extensions, and frameworks*. Apress, 2021. doi: 10.1007/978-1-4842-7158-2.

E. Nazemi, M. J. Tarokh, and G. R. Djavanshir. ERP: a literature survey. *International Journal of Advanced Manufacturing Technology*, 61(9–12):999–1018, 2012. doi: 10.1007/s00170-011-3756-x.

Oliver Neumann, Katharina Guirguis, and Reto Steiner. Exploring artificial intelligence adoption in public organizations: a comparative case study. *Public Management Review*, 26(1):114–141, 2024. doi: 10.1080/14719037.2022.2048685.

Andrew Ng. Autonomous coding agents, instability at stability AI, Mamba mania, and more. The Batch (DeepLearning.AI), April 10 2024a. URL https://www.deeplearning.ai/the-batch/issue-244/.

Andrew Ng. Four AI agent strategies that improve GPT-4 and GPT-3.5 performance. The Batch (DeepLearning.AI), March 20 2024b. URL https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance.

Kasey Panetta. Gartner's top 10 technology trends 2017. Gartner, October 18 2016. URL https://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Hao Cao, Xin Cheng, Michael Chung, Xingjian Du, Matteo Grella, G. V. Kranthi Kiran, Xuzheng He, Haowen Hou, Jiaxi Lin, Przemysław Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, 2023. doi: 10.18653/v1/2023.findings-emnlp.936.

Eric G. Poon, Christy Harris Lemak, Javier C. Rojas, John Guptill, and David Classen. Adoption of artificial intelligence in healthcare: Survey of health system priorities, successes, and challenges. *Journal of the American Medical Informatics Association : JAMIA*, 32(7):1093–1100, 2025. doi: 10.1093/jamia/ocaf065.

R. M. Potluri and D. Serikbay. Artificial intelligence (AI) adoption in HR management: Analyzing challenges in kazakhstan corporate projects. *International Journal of Asian Business and Information Management*, 16(1):1–18, 2025. doi: 10.4018/IJABIM.376012.

R. Praveen, A. Shrivastava, G. Sharma, A. M. Shakir, M. Gupta, and S. S. S. R. G. Peri. Overcoming adoption barriers strategies for scalable AI transformation in enterprises. In *2025 International Conference on Engineering, Technology & Management (ICETM)*, pages 1–6, 2025. doi: 10.1109/ICETM63734.2025.11051446.

Rob Procter, Peter Tolmie, and Mark Rouncefield. Holding AI to account: Challenges for the delivery of trustworthy AI in healthcare. *ACM Transactions on Computer-Human Interaction*, 30(2):1–34, 2023. doi: 10.1145/3577009.

X. Qian and D. Klabjan. A probabilistic approach to neural network pruning. Technical report, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of Machine Learning Research*, volume 139, pages 8748–8763, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of Machine Learning Research*, volume 202, pages 28492–28518, 2022.

C. K. Kumar Reddy. *The Power of Agentic AI: In Industry 6.0*. Springer, 1 edition, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, 2019. doi: 10.18653/v1/d19-1410.

Reuters, Thomson. What an ERP system is and how it impacts your business, April 15 2025. URL https://tax.thomsonreuters.com/blog/what-is-an-erp-system-and-why-does-it-matter-for-your-business/.

F. Robert Jacobs and F. C. Ted Weston. Enterprise resource planning (ERP)—a brief history. *Journal of Operations Management*, 25(2):357–363, 2007. doi: 10.1016/j.jom.2006.11.005.

E. Romeo and J. Lacko. Adoption and integration of AI in organizations: a systematic review of challenges and drivers towards future directions of research. *Kybernetes*, 2025. doi: 10.1108/K-07-2024-2002.

 doi:10.20944/preprints202512.0428.v1

Azad M. Salih, Zahra Raisi-Estabragh, Ilaria B. Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz. A perspective on explainable artificial intelligence methods: SHAP and LIME. 2024. doi: 10.1002/aisy.202400304.

Friederike Selten and Bram Klievink. Organizing public sector AI adoption: Navigating between separation and integration. *Government Information Quarterly*, 41(1):101885, 2024. doi: 10.1016/j.giq.2023.101885.

D. Shah, J. Soldatos, M. Hinkle, N. Hoher, N. von I. Seip, V. Just, K. Kechichian, K. Benkrid, M. McDonagh, V. Jesaitis, and W. Abbey. Arm AI readiness index, 2024.

Mahak Sharma, Sunil Luthra, Sudhanshu Joshi, and Anil Kumar. Implementing challenges of artificial intelligence: Evidence from public manufacturing sector of an emerging economy. *Government Information Quarterly*, 39 (4):101624, 2022. doi: 10.1016/j.giq.2021.101624.

X. Song. Flink agents: The agentic AI framework based on Apache Flink. Community Over Code Asia 2025, July 25 2025. URL https://asia.communityovercode.org/sessions/streaming-913378.html.

K. Soule and D. Bergmann. IBM Granite 4.0: Hyper-efficient, high performance hybrid models for enterprise. IBM Announcements, October 2 2025. URL https://www.ibm.com/new/announcements/ibm-granite-4-0-hyper-efficient-high-performance-hybrid-models.

K. M. Stief. Artificial? yes. intelligent? not really, October 1 2025. URL https://www.computing.co.uk/analysis/2025/artificial-yes-intelligent-not-really-copy.

Suhas Jayaram Subramanya, Rohan Devvrit, Kadekodi, Ravishankar Krishnaswamy, and Harsha Vardhan Simhadri. DiskANN: Fast accurate billion-point nearest neighbor search on a single node. In *NeurIPS 2019*, 2019. URL https://www.microsoft.com/en-us/research/publication/diskann-fast-accurate-billion-point-nearest-neighbor-search-on-a-single-node/.

S. Sumathi. *Fundamentals of relational database management systems*. Springer, 1 edition, 2007. doi: 10.1007/978-3-540-48399-1.

E. Sánchez, R. Calderón, and F. Herrera. Artificial intelligence adoption in SMEs: Survey based on TOE–DOI framework, primary methodology and challenges. *Applied Sciences*, 15(12):6465, 2025. doi: 10.3390/app15126465.

Araz Taeihagh. Governance of artificial intelligence. *Policy and Society*, 40(2):137–157, 2021. doi: 10.1080/14494035.2021.1928377.

The Linux Foundation. Linux foundation launches the Agent2Agent protocol project to enable secure, intelligent communication between AI agents, June 23 2025. URL https://www.linuxfoundation.org/press/linux-foundation-launches-the-agent2agent-protocol-project-to-enable-secure-intelligent-communication-between-ai-agents.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems 35*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2023.

Gregory Vial, Ann-Frances Cameron, Theodore Giannelia, and Jing Jiang. Managing artificial intelligence projects: Key insights from an AI consulting firm. *Information Systems Journal*, 33(3):669–691, 2023. doi: 10.1111/isj.12420.

L. Wang, S. Chen, L. Jiang, S. Pan, R. Cai, S. Yang, and F. Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):1–64, 2025. doi: 10.1007/S10462-025-11236-4.

H. Wei, Y. Sun, and Y. Li. DeepSeek-OCR: Contexts optical compression. Technical report, October 2025. URL https://arxiv.org/pdf/2510.18234.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, 2022.

S. Williams. Surge in AI adoption amongst sales challenges in Singapore. CFOTech Asia, August 5 2024. URL https://cfotech.asia/story/surge-in-ai-adoption-amongst-sales-challenges-in-singapore.

World Economic Forum. AI governance trends: How regulation, collaboration, and skills demand are shaping the industry, 2024. URL https://www.weforum.org/stories/2024/09/ai-governance-trends-to-watch/.

Jie Yang, Yvette Blount, and Alireza Amrollahi. Artificial intelligence adoption in a professional service industry: A multiple case study. *Technological Forecasting and Social Change*, 201:123251, 2024. doi: 10.1016/J.TECHFORE.2024.123251.

 doi:10.20944/preprints202512.0428.v1

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *11th International Conference on Learning Representations, ICLR 2023*, 2022.

YCombinator Hacker News. Is there a half-life for the success rates of AI agents?, June 2025. URL https://news.ycombinator.com/item?id=44308711.

Ted Young and Austin Parker. *Learning OpenTelemetry*. O'Reilly Media, Incorporated, 1 edition, 2024.

F. Yu. When AIs judge AIs: The rise of agent-as-a-judge evaluation for LLMs. Technical report, August 2025. URL https://arxiv.org/pdf/2508.02994.

Alex L. Zhang. Recursive language models, October 15 2025. URL https://alexzhang13.github.io/blog/2025/rlm/.

Q. Zhang, Z. Liu, S. Pan, and C. Wang. The rise of small language models. *IEEE Intelligent Systems*, 40(1):30–37, 2025. doi: 10.1109/MIS.2024.3517792.

Yutao Zhu, Hao Yuan, Shujin Wang, Jinyi Liu, Wenjie Liu, Chao Deng, Haonan Chen, Zhicheng Liu, Zhiyuan Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. 2025. doi: 10.1145/3748304.

M. Zhuge, C. Zhao, D. R. Ashley, W. Wang, D. Khizbullin, Y. Xiong, Z. Liu, E. Chang, R. Krishnamoorthi, Y. Tian, Y. Shi, V. Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. Technical report, October 2024. URL https://arxiv.org/pdf/2410.10934.