

Structural Barriers to the Deployment of AI Foundation Models

Hristo Hristov¹, Elena Somova² and Emil Fuchedzhiev³

¹Department of Software Technologies, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria

²Department of Computer Science, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria

³Department of Political Sciences and National Security, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria

Article history

Received: 14-07-2026

Revised: 06-08-2026

Accepted: 27-08-2026

Corresponding Author:

Elena Somova

Department of Computer
Science, University of Plovdiv
"Paisii Hilendarski", Plovdiv,
Bulgaria

Email: eledel@uni-plovdiv.bg

Abstract: Artificial Intelligence (AI) foundation models excel at benchmarks but are not always cost-effective or successfully deployed in enterprise systems. This paper analyzes the failures in their deployment through interconnected structural components characterized by: Economic constraints (computational costs, memory, and latency that scale with increasing use of AI model services), cognitive constraints (related to training objectives that optimize plausibility rather than causal reasoning and calibrated uncertainty), and socio-technical constraints (requirements for organizational change, accountability, and governance). The study examines why the deployment of AI models in real-world enterprise environments remains difficult. It focuses on four themes – scalability, reliability, interpretability, and practical deployment – aiming to demonstrate why these problems do not disappear even as models improve, and how these issues reinforce each other, thereby leading to higher costs and greater risk. In conclusion, the study outlines alternative development pathways, including the use of alternative architectural paradigms such as Retrieval-Augmented Generation (RAG), hybrid Neuro-symbolic AI, compound and multi-agent systems, and human-in-the-loop designs, which may offer more practical and sustainable ways to align foundation models with the organizational, operational, and governance requirements of the enterprise environment. Additionally, a four-dimensional framework is proposed for assessing a model's deployment readiness based on economic viability, reliability and safety, auditability, and organizational readiness.

Keywords: AI Models, AI Deployment, Retrieval Augmented Generation (RAG), Economic Sustainability, AI Deployment Constraints

Introduction

Since the Transformer network architecture (Vaswani et al., 2017) inaugurated the era of large language models (Brown et al., 2020), foundation models (large-scale AI systems pre-trained on massive datasets and adaptable to a wide spectrum of tasks) have evolved rapidly, expanding into multimodal understanding (Radford et al., 2021), mathematical problem-solving, and chain-of-thought reasoning (Wei et al., 2022; OpenAI, 2024). Nevertheless, most enterprise initiatives encounter serious difficulties when moving beyond pilot projects. Industry analysts project that at least 30% of generative AI projects will be abandoned after the proof-of-concept stage by the end of 2025, owing to poor data quality, inadequate risk

controls, escalating costs, or unclear business value (Gartner, 2024). Other studies also report low success rates, ranging from 15 to 35%, indicating that benchmark capabilities (standardized test datasets used to evaluate and compare different models) do not easily translate into reliable systems (Paley et al., 2023). Despite this, the widespread industry belief persists that deployment problems will be automatically resolved with the advent of larger and more powerful models.

However, practice and analysis of the specialized literature have shown that there are structural barriers rooted in economic processes, cognitive limitations of the models, and organizational conditions in which foundation models are deployed barriers that may persist or even worsen as their application scales. From this

perspective, the paper examines three main categories of constraints.

Economic constraints, where computational and infrastructure costs grow as model capacity increases.

Cognitive constraints, arising from the fact that the statistical text-generation mechanism does not inherently guarantee truthfulness, causal reasoning, or calibrated uncertainty estimation (uncertainty calibration corrects the imbalance between the confidence level the model assigns and the actual probability of its predictions being correct).

Socio-technical constraints (limitations arising from the interaction of AI models with established organizational structures, social norms, regulatory requirements, and user behavior), where deployment requires organizational transformation that no model upgrade can automate.

The paper aims to identify and analyze the structural challenges associated with deploying AI foundation models in domains where decisions are high-stakes and require high reliability and accuracy.

The study is based on four research questions:

- RQ1: What structural barriers prevent foundation models from achieving reliable enterprise deployment, despite advances in model capabilities?
- RQ2: Why does scaling model capacity not constitute a sufficient solution to overcome these barriers?
- RQ3: How do these barriers interact with each other and complicate the deployment of foundation models?
- RQ4: What alternative approaches can address these structural constraints?

Accordingly, the analysis is organized around scalability, reliability, interpretability, and practical deployment, followed by alternative architectural pathways and the proposed deployment readiness framework.

Research Problem

To avoid a one-sided analysis, it is worth first noting where AI foundation models already provide real value. They are particularly useful in code generation and completion, as well as in translating natural-language descriptions into programmatic solutions, when human review remains part of the workflow (Xu et al., 2022). They also power semantic search and information retrieval, dramatically improving the relevance of results without requiring users to accept them as absolute truth. As thinking partners for brainstorming, hypothesis generation, and rapid prototyping, they lower the barrier to experimentation through zero-shot or few-shot capabilities.

A common pattern unites these successes: Tolerance for imperfection, low-stakes consequences, emphasis on

synthesis over factual precision, and appropriate scoping. These patterns reveal why high-stakes, accuracy-critical applications the focus of this paper face structurally different challenges. AI foundation models have limitations both transient and structural.

Transient limitations, such as context window length, inference latency, and specific benchmark gaps, consistently yield to engineering efforts.

Structural limitations arise from the way the paradigm itself operates and do not automatically diminish as models grow.

We call a limitation structural when it satisfies at least two of the following four criteria:

- (1) It derives from proven mathematical properties of the training paradigm
- (2) It persists or worsens across three or more model generations
- (3) It recurs across fundamentally different architectures
- (4) It depends on factors external to the model (markets, institutions, regulations)

A barrier that we term structural would not be considered in this paper if it were resolved through continued scaling within 3–5 generations.

From an economic perspective, scaling laws show that performance improvements go hand in hand with increases in parameters, data, and computational resources (Hoffmann et al., 2022). Despite advances in efficiency, training and especially inference remain key practical bottlenecks in the deployment of large models (Xiao et al., 2025).

At the cognitive level, models learn statistical regularities rather than causal models (Pearl and Mackenzie, 2018). Since next-token prediction is only weakly related to truth, systems remain prone to hallucinations and fragile generalization, even when highly capable.

At the socio-technical level, AI deployment requires changes in workflows, accountability, and governance that cannot be overcome solely through technological means.

Research Methodology

This paper constitutes a narrative synthesis and position paper, which allows integration of diverse types of evidence – empirical studies, theoretical frameworks, and industry reports. This approach was chosen because the research question spans three distinct disciplinary areas (economics, epistemology, socio-technical systems) and requires synthesis of qualitatively different types of data, for which a unified systematic protocol would be constraining.

The literature review covers the period 2018–2026 and was conducted in three stages: Identification, screening, and selection. During the identification stage, the

following academic databases and repositories were used: ACM Digital Library, IEEE Xplore, Google Scholar, and arXiv. Additionally, proceedings from leading conferences in the field were reviewed (NeurIPS, ICML, AAAI, ACL, ICLR, and CVPR), specialized journals (Nature Machine Intelligence, Journal of Artificial Intelligence Research, Artificial Intelligence, and Communications of the ACM), as well as technical reports from research laboratories (OpenAI, Google DeepMind, Microsoft Research, Anthropic, and Meta AI) and industry reports (Gartner, McKinsey, and Stanford AI Index). The search was conducted using the following keywords and their combinations: “AI limitations”, “LLM challenges”, “enterprise deployment”, “AI robustness”, “model interpretability”, “AI safety”, “production ML”, “scaling laws”, “foundation model deployment”, and related phrases. Priority was given to more recent publications, though fundamental older works were also included.

Since the paper is a narrative synthesis rather than a systematic review, the literature selection follows the logic of purposive sampling rather than an exhaustive protocol. The initial search using the described keywords generated over 200 results. From these, 59 publications were selected that meet the following criteria:

- (1) They address at least one of the three barrier categories economic, cognitive, or socio-technical
- (2) They contain empirical data, formal theoretical analysis, or documented industry observations
- (3) They are published in a peer-reviewed venue or constitute a technical report from an established organization

Publications focused exclusively on architectural improvements without analysis of deployment implications were excluded, as were duplicative results from the same authorial teams. The selection was performed by the authors through an iterative process, with additional sources identified through citation tracking (snowball sampling).

Alongside the academic literature, industry reports were also reviewed, which is justified by the need for the analysis to reflect real deployment challenges rather than solely theoretical considerations. These sources were treated as supplementary evidence and do not form the basis of the argumentation when they cannot be corroborated by academic research.

The identified literature was analyzed through thematic synthesis, and the discovered limitations were grouped into three categories: Economic, cognitive, and socio-technical barriers. The study prioritizes arguments that are robust across different model generations rather than measurements tied to a single vendor.

The choice of literature reflects expert judgment and does

not claim exhaustiveness. Quantitative claims are retained only when they are supported by identifiable sources. Cross-study values are treated as context-specific because models, workloads, hardware configurations, and evaluation protocols differ. Since the paper is a position paper, its conclusions are subject to empirical verification through future longitudinal studies tracking whether the identified barriers continue to exist in subsequent model generations.

Because this study uses purposive rather than exhaustive sampling, the review may be affected by selection bias, publication bias, and the greater visibility of English-language sources. To reduce these risks, we combined academic databases, conference proceedings, citation tracking, and reports from established organizations. Peer-reviewed studies provide the primary evidence, while industry reports and commercial examples are used to illustrate deployment context and current practice. Nevertheless, the final selection and thematic classification reflect expert judgment.

Structural Constraints for Deploying AI Foundation Models

Economic Constraints

Foundation models embody a fundamental economic paradox: The very scaling that drives capability improvement simultaneously increases costs that most enterprise applications cannot justify.

Superlinear Cost Growth

Scaling laws confirm that larger models trained on more data lead to lower losses (Hoffmann et al., 2022), but the economic implications remain contentious. As scale increases, training and inference requirements grow, while the practical value of small quality increments does not always increase at the same rate (Xiao et al., 2025).

Training economics. Frontier-model training requires substantial capital, energy, and specialized computational infrastructure. Available studies show that compute requirements increase with model scale, but precise cost comparisons remain difficult because developers rarely disclose complete information about training runs, failed experiments, hardware utilization, personnel, and energy consumption. This limited transparency contributes to the concentration of frontier-model development among organizations with access to large-scale computational resources (Strubell et al., 2019; Xiao et al., 2025).

Inference economics. Unlike training, inference represents a recurring operational cost that accumulates with every request. Its economic feasibility depends on model size, input and output token volume, request frequency, latency requirements, deployment mode, hardware utilization, and the value generated by the specific use case. Inference costs should therefore be assessed through a use-case-specific total cost of

ownership analysis rather than through a universal monthly estimate.

Energy sustainability. Carbon footprints are non-trivial. For example, BLOOM, with 176 billion parameters, emitted approximately 25 metric tons of CO₂ during training under the conditions reported by Luccioni et al. (2023). More broadly, energy and carbon requirements tend to increase with model size and computational demand (Strubell et al., 2019). However, cross-model comparisons remain uncertain because developers rarely disclose complete information about energy consumption, hardware utilization, and training procedures.

Parametric Efficiency and Its Limits

The self-attention mechanism exhibits $O(n^2)$ complexity with respect to input sequence length, so a context window of 100,000 tokens requires 10,000 times more attention computation than 1,000 tokens. Efficient and sparse attention variants exist that reduce computational cost for long sequences, but practical gains are typically accompanied by architectural trade-offs and do not always preserve quality across all tasks (Tay et al., 2022). Loading a model with 100 billion parameters requires 200 GB of GPU memory at FP16 (16-bit floating point) precision, and the Key-Value (KV) Cache grows linearly with context length and batch size, limiting concurrent request processing (Pope et al., 2023). Meanwhile, the inherent seriality of autoregressive decoding introduces sequential latency that can constrain interactive and real-time applications unless model capacity, output length, or serving configuration is adjusted.

Optimization and the Gap Between Cost and Performance

Techniques such as quantization, pruning, distillation, and efficient architectures can substantially improve computational and memory efficiency. However, the achieved gains are model- and task-dependent and may involve trade-offs in accuracy, robustness, or complex reasoning performance, particularly under aggressive compression (Frantar and Alistarh, 2023). These techniques mitigate, but do not eliminate, the tension between capability, memory requirements, latency, and deployment cost.

Foundation models also lock organizations into specialized hardware accelerators with continuous hardware upgrade cycles unlike traditional software that runs on commodity hardware. And since models require periodic retraining to remain current, computational workloads are ongoing rather than one-time.

Hardware scaling, distributed inference, quantization, pruning, and distillation can substantially reduce cost and latency. Their effect, however, depends on workload

volume, latency targets, model size, hardware utilization, and required accuracy. The claim is therefore conditional: In some high-volume or low-margin use cases, efficiency gains may remain insufficient to offset total lifecycle costs.

Implications for Enterprise Deployment

High-margin workflows (premium professional services, specialized R&D) can absorb foundation model costs. Conversely, routine enterprise operations often cannot justify the costs of deploying and using large AI models. Alternative options exist, including hybrid architectures that route simple tasks to small models and reserve large models only for complex queries. Retrieval-Augmented Generation (RAG) pipelines that offload factual knowledge to cheaper databases (Lewis et al., 2020), and “model-as-a-service” pricing each of these helps, but none eliminates the fundamental tension between capability costs and business value.

Cognitive Constraints

The second structural barrier arises from the way foundation models learn and represent knowledge. Trained through next-token prediction on massive text corpora, these systems learn statistical correlations without developing causal understanding, world models, or mechanisms for distinguishing truth from plausible-sounding fabrication. This epistemological limitation creates fundamental barriers to reliability and trust in critical domains such as healthcare, law, and finance.

Correlation vs. Causation

Foundation models optimize predictive accuracy on the training corpus by learning that certain words, phrases, and patterns co-occur. However, statistical correlation does not imply causation (Pearl and Mackenzie, 2018). A model may learn that “firefighters” and “fires” frequently co-occur without understanding that firefighters respond to fires rather than causing them. This distinction is crucial for counterfactual reasoning, reasoning in novel situations, and robustness to distribution shift.

Counterfactual reasoning: Answering “what if” questions requires causal models capable of simulating interventions (hypothetical changes in conditions). Foundation models struggle with counterfactual claims outside their training data, often fabricating plausible-sounding but incorrect answers. For example, an urban planning system asking “what if we reduce the number of firefighters in areas with frequent fires?” might incorrectly recommend reducing firefighters, since correlation does not reveal that firefighters are a response to fires, not their cause.

Novel situations: Statistical pattern matching fails in situations substantially different from the training data. For example, during the early COVID-19 pandemic,

models trained on pre-pandemic data systematically made incorrect predictions about policy impacts, behavioral changes, and medical outcomes because they lacked causal understanding of disease transmission dynamics.

Robustness to distribution shift: AI models trained on correlational patterns are brittle to distribution shifts where spurious correlations break down while causal relationships are preserved. A hiring model that learns correlations between educational institution prestige and workplace performance may fail when applied to non-traditional candidates due to a lack of understanding of the underlying markers of the sought-after abilities.

The Interpretability Capability Trade-Off

As models become more capable, they become less interpretable not as a side effect, but as a consequence of distributed neural coding. Abstract concepts such as “medical risk” are not encoded in individual neurons but emerge from distributed activation patterns spanning thousands of neurons simultaneously making them resistant to decomposition into human-readable components.

Post-hoc explanation techniques (attention visualization, saliency maps) provide at most a superficial understanding; they can rationalize rather than faithfully describe the model’s computation (Rudin, 2019). Mechanistic interpretability research is beginning to uncover specific internal computational circuits in Transformer architectures (Elhage et al., 2021), but these approaches are computationally expensive and currently difficult to apply to frontier models (the largest contemporary systems). Therefore, interpretability remains a contested and context-dependent requirement (Lipton, 2018), creating an inherent tension between model complexity and the reliability requirements of enterprise deployment.

Hallucinations and Fragile Generalization

Foundation models systematically hallucinate generating convincing but factually incorrect outputs. This is not a bug awaiting a fix; it stems directly from training objectives that reward plausibility over truth. Next-token prediction treats a correct scientific explanation and a plausible-sounding but incorrect explanation equally, as long as both are good continuations. Unlike humans who learn from sensorimotor experience, language models learn solely from text, capturing patterns in how humans describe reality rather than reality itself.

Calibration of trust compounds the problem because model confidence does not always correspond to actual correctness (Guo et al., 2017). Calibration methods can reduce overconfidence on selected benchmarks, but their effectiveness varies across models, datasets, domains, and evaluation protocols. Models may therefore remain

poorly calibrated under distribution shift and in domain-specific applications.

Reported hallucination rates vary substantially across models, tasks, prompting strategies, factuality benchmarks, and evaluation methods. Although recent model and retrieval-based approaches report improvements on specific benchmarks, no single percentage reliably characterizes progress across systems. Hallucination therefore remains a material deployment risk, particularly in domain-specific and high-stakes applications (Ji et al., 2023; Asai et al., 2024; Dong et al., 2025).

Fragile generalization (strong dependence on current inputs or specific conditions, yielding correct operation in expected scenarios but frequent errors when data or context changes) is more problematic than hallucination. Even small, semantically equivalent changes in prompts can lead to substantially different outputs, indicating that sensitivity to phrasing remains an important source of instability (Chatterjee et al., 2024). Adversarial inputs, such as prompt injections or jailbreaks (specifically crafted inputs designed to deceive machine learning models into making incorrect predictions or classifications), exploit fundamental properties of high-dimensional optimization and adversarial transferability (the phenomenon of attacks transferring across different architectures and model generations) (Papernot et al., 2017).

Cognitive Barriers in High Reliability Applications

In domains where errors carry serious consequences, these cognitive barriers become critical. Healthcare, for example, requires causal reasoning about therapeutic mechanisms, drug interactions, and contraindications statistical pattern matching alone risks producing life-threatening recommendations. Law requires distinguishing binding precedents from superficially similar cases, with a widely reported instance of fabricated legal citations in a court filing already leading to attorney sanctions (Mata v. Avianca, Inc., 2023). Finance requires robustness to rare events and market regime shifts where historical correlations lose validity. In every domain, the gap between benchmark performance and real-world reliability underscores the need for architectures that go beyond statistical pattern matching – incorporating explicit causal models, symbolic reasoning, and verified knowledge bases.

Socio Technical Constraints

The third structural barrier reflects a widespread categorization error: Enterprise AI failures are largely problems of organizational change disguised as technical integration challenges.

Bias Amplification

Models trained on historical data inevitably encode historical inequalities. For example, a hiring model learns

past discrimination, while a credit algorithm perpetuates redlining (the historical practice of denying credit based on geographic or demographic criteria).

Audits of commercial AI systems have documented substantial demographic disparities a landmark study of commercial gender-classification systems found error rates of up to 34.7% for darker-skinned women, compared with 0.8% for lighter-skinned men, exposing intersecting gender and racial biases (Buolamwini and Gebru, 2018). Software-level debiasing helps but cannot solve the problem because different fairness definitions are mathematically incompatible optimizing demographic parity sacrifices equalized odds and vice versa (Verma and Rubin, 2018). Which criterion matters more is ultimately a political and ethical question, not a technical one. Critically, bias scales with deployment: Systems making millions of automated decisions amplify discriminatory impacts far beyond what individual human decision-makers could produce.

Integration as Organizational Change

AI deployment requires redesigning workflows, not merely automating them. For example, diagnostic AI does not simply assist radiologists it restructures how studies are ordered, triaged, and documented. This requires workforce reskilling across domains (AI literacy for non-IT specialists, machine learning operations skills for IT specialists, compliance auditing skills for validators), but organizations consistently underestimate the required time and costs. AI also fragments accountability when an automated system errs, responsibility disperses among executives, data scientists, engineers, validators, and vendors, leading to paralysis and risk aversion (Selbst et al., 2019). And resistance is not always irrational professionals have legitimate concerns about job displacement (being replaced by AI), deskilling (loss of skills due to over-reliance on AI), and loss of professional autonomy (Attewell, 1987).

Accountability failures are not an unavoidable property of foundation models. They emerge from the interaction between model behavior, workflow design, data ownership, and decision authority, and can be reduced through clear role allocation, human review, escalation paths, audit procedures, and post-deployment monitoring. These controls require continuing organizational coordination and must be embedded in daily practice; formal compliance alone is insufficient.

The Integration Complexity Trap

Most organizations manage IT landscapes spanning decades of technological generations, with 60-80% of AI project effort spent on integration with legacy systems (Paley et al., 2023). Enterprise data is fragmented across data silos, inconsistently formatted, and often privacy-sensitive, leading to “data cascades” that accumulate

errors throughout the pipeline (Sambasivan et al., 2021). Regulatory requirements, including the EU AI Act, add obligations related to documentation, risk management, human oversight, monitoring, and conformity assessment. These requirements can materially increase development costs and deployment time, although their impact depends on the system’s risk classification, application domain, jurisdiction, and the organization’s existing governance capabilities (European Parliament and Council of the European Union, 2024).

Adversarial Dynamics and AI Security Risks

AI security differs fundamentally from traditional cybersecurity. Adversarial attacks (techniques using specially crafted inputs designed to mislead models) evolve faster than models can be retrained, and black-box attack practices show that adversarial examples often transfer across different models and configurations (Papernot et al., 2017).

The use of such data can create backdoors that are difficult to detect, and the backdoor attack literature shows that these vulnerabilities can remain hidden even after model fine-tuning (Zhang et al., 2024), while model extraction attacks undermine intellectual property protection (Tramèr et al., 2016). As AI increasingly mediates social and economic life, these vulnerabilities create systemic risks that traditional security paradigms are not designed to address.

Strategic Implications

Addressing the described socio-technical barriers requires participatory design (a methodology for collaborative design with active involvement of stakeholders from early development stages) that engages stakeholders during development (Raji et al., 2020); gradual adoption starting with low-stakes applications; human-AI collaboration architectures that complement rather than replace expert judgment (Bansal et al., 2021); and governance frameworks including model cards (Mitchell et al., 2019) (standardized documents describing a model’s intended purpose, limitations, and known weaknesses), datasheets (documentation of training data provenance, scope, and quality), and audit trails tracking system decisions.

Scaling is not the Solution

Scaling alone cannot overcome these barriers. Each barrier resists scaling for different reasons. Economically, scaling laws show that improved performance requires additional parameters, data, and computation, while marginal business value often saturates more quickly (Hoffmann et al., 2022). Cognitively, the richer distributed representations formed by larger models make those models less interpretable, not more (Lipton, 2018). Sociologically, organizational change theory shows that

institutional resistance and power dynamics are orthogonal to model capabilities (Orlikowski, 1996).

Empirical evidence. Evidence across model generations indicates that hallucinations, adversarial vulnerabilities, bias, and hidden technical debt remain relevant concerns, although their severity varies across models and evaluation settings (Ji et al., 2023; Papernot et al., 2017; Bender et al., 2021; Sculley et al., 2021). And the hidden technical debt that makes ML systems costly to maintain and slow to bring into production persists despite major advances in capabilities (Sculley et al., 2021).

Counterarguments. Scaling inference-time compute can improve reasoning performance (OpenAI, 2024), but it also increases per-request computation, latency, and cost, thereby reinforcing rather than eliminating the economic trade-off. Reinforcement learning from human feedback and Constitutional AI (AI guided by a predefined set of principles and rules without relying on human feedback) have reduced harmful outputs (Bai et al., 2022), but optimizing for human preferences is not the same as optimizing for truth. There are also clear practical successes: GitHub Copilot is associated with substantial improvements in developer productivity and perceived effectiveness (Ziegler et al., 2024), and generative AI assistance in customer support has been associated with a roughly 15% average increase in issues resolved per hour, with the largest gains about 34% accruing to less-experienced workers (Brynjolfsson et al., 2025).

These successes are consistent with, rather than contradicting, our thesis. Successful cases cluster in contexts that tolerate imperfection, avoid autonomous high-stakes decisions, include rapid feedback loops, and serve high-value workflows that can absorb inference costs. The barriers we identify are most acutely felt where reliability must be near-perfect, ground-truth verification is expensive, data volumes are large, and margins are thin.

Interaction between barriers. The three barriers do not act in isolation on the contrary, they mutually reinforce each other (Fig. 1). Budget constraints limit reliability testing, which exacerbates cognitive weaknesses. Hallucinations undermine stakeholder trust, amplifying socio-technical resistance. Compliance and integration costs further strain budgets, rendering economically marginal use cases infeasible.

For example, in healthcare deployment: High reliability requirements (socio-technical) demand extensive validation, which raises costs (economic). To limit costs, organizations may ship insufficiently tested systems that then hallucinate (cognitive), triggering regulatory scrutiny and clinical pushback (socio-technical) a vicious cycle in which addressing one barrier puts pressure on the others. These feedback dynamics explain why isolated solutions are rarely sufficient and why holistic architectural alternatives, discussed below, become essential.

Mutual reinforcement of structural barriers

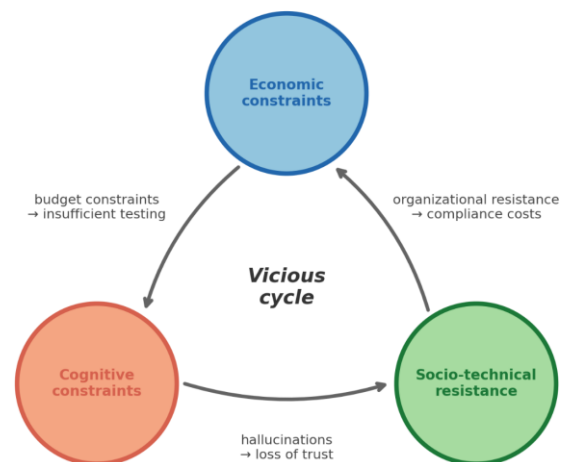


Fig. 1: Vicious cycle of mutual reinforcement among structural barriers

Recommendations for Overcoming Structural Barriers

This section outlines research directions and practical strategies for addressing the structural barriers.

Alternative Architectural Paradigms

Retrieval Augmented Generation (RAG), an architectural approach in AI that enhances LLMs by first retrieving information from external sources (augmenting the context) before generating a response, offloads factual knowledge to external databases that can be verified, updated, and audited (Lewis et al., 2020). By shifting from “knowledge in parameters” to “knowledge in databases,” RAG can reduce hallucinations and, in suitable workloads, computational costs. By 2025, it has become one of the dominant patterns for LLM deployment in production, with retrieval components widely adopted across enterprise systems (Fan et al., 2024).

Architecturally, RAG systems follow a three-stage process:

- (1) Indexing documents are segmented into chunks and encoded into dense vector representations via embedding models
- (2) Retrieval for each query, the most relevant chunks are retrieved via semantic similarity
- (3) Generation the retrieved chunks are provided as context to an LLM to produce grounded responses (Fan et al., 2024)

This design can mitigate aspects of all three structural

barriers: It reduces hallucinations by grounding generated responses in verifiable documentary sources (cognitive), enables the use of smaller generative models (economic), and provides transparent and traceable links to the sources used (socio-technical).

Empirical results demonstrate significant improvements. Self-RAG achieves up to 20+ percentage-point improvement on the FactScore metric (measuring the factual accuracy of generated text) compared to standard generation, and significant gains on factual accuracy benchmarks over PopQA and PubHealth test and training datasets (Asai et al., 2024). Critic-guided agentic RAG approaches also reduce errors through automated retrieval quality assessment and self-correction, improving reliability across multiple RAG scenarios (Dong et al., 2025). In enterprise settings, RAG systems demonstrate significantly higher accuracy on knowledge retrieval tasks at lower computational costs than fine-tuned models (Ram et al., 2023). Their critical advantage is currency: External databases can be updated continuously without expensive model retraining, solving the “knowledge cutoff” problem that plagues static LLMs.

Nevertheless, RAG introduces its own challenges. Retrieval quality becomes a critical point of failure if relevant documents are not retrieved, the generated response cannot be accurate. Document segmentation requires careful balancing chunks that are too small lose context, while chunks that are too large overwhelm the context window and dilute relevance. The retrieval stage introduces additional latency, whose magnitude depends on the retrieval infrastructure, indexing strategy, re-ranking procedure, and number of retrieved documents. Organizationally, RAG requires maintenance of vector stores and indexing pipelines additional infrastructure complexity that some organizations find prohibitive (Asai et al., 2024).

RAG is not a monolithic technique the paradigm has evolved from basic Naive RAG through Advanced to more dynamic Agentic RAG architectures that combine retrieval, planning, and execution control (Leng et al., 2025).

Self-RAG (Asai et al., 2024), which improves response quality by using self-reflection through evaluating and critiquing the relevance of its own generated texts, trains models to decide when retrieval is truly needed and when they already possess sufficient knowledge, avoiding unnecessary retrieval from external sources.

Critic-guided agentic RAG (Dong et al., 2025) employs a self-assessment and planning mechanism where a critic module evaluates errors in retrieval and generation and directs subsequent corrective actions. In this way, the system moves beyond static retrieval and orchestrates automated self-correction within the RAG pipeline (Dong et al., 2025).

Adaptive RAG (Jeong et al., 2024) classifies query complexity and dynamically selects a strategy – no retrieval for simple questions, single-step retrieval for factual queries, and multi-step retrieval for complex questions.

Graph RAG approaches replace flat documents (where vector search is applied to unstructured text) with structured knowledge graphs, improving reasoning over relationships and dependencies and being particularly useful for multi-hop tasks (Hu et al., 2025).

Agentic RAG (Leng et al., 2025) integrates autonomous agents with capabilities for reflection, planning, and tool use into the RAG pipeline, enabling multi-step retrieval and adaptive execution for complex tasks. Each approach involves trade-offs between accuracy, speed, and complexity, with the choice depending on the specific domain and requirements.

Real-world results confirm RAG’s value. Microsoft Copilot and GitHub Copilot Chat, along with many enterprise chatbots, use RAG to ground responses in corporate documents (Fan et al., 2024). Perplexity AI has built an entire product as a “RAG-native” search engine, and Amazon Q provides enterprise RAG for internal data and documentation. In law, RAG systems search case law with expert precision and provide verifiable citations – critical for professional liability. In healthcare, RAG assistants demonstrate improved accuracy and source transparency, which is highly valued by clinicians since the system shows where the information comes from (Xiong et al., 2024).

Hybrid neuro-symbolic systems (Neuro-symbolic AI) attack the cognitive barrier from a different angle than information retrieval. Instead of grounding generated responses in verifiable documentary sources in text documents, they combine neural pattern recognition (deep learning) with formal symbolic reasoning, logical verifiability, and structured ontologies (Garcez and Lamb, 2023). Symbolic reasoning can apply logical rules, check internal consistency, and guarantee that inferences follow defined formal constraints qualities that a purely statistical approach fundamentally cannot provide. This is particularly valuable in domains such as medicine and law, where it is not sufficient for the model to “sound correct” it must be provably correct. The main challenge remains scalability: Integrating symbolic reasoning with models at the scale of contemporary LLMs is still in an early phase, and engineering complexity has grown significantly.

Compound and multi-agent systems represent a broader architectural paradigm that rejects the idea of a single monolithic model responsible for everything. Instead, they compose specialized components LLM generators, retrieval modules, code executors, validators, and critics into orchestrated pipelines (Zaharia et al., 2024). In multi-agent variants, individual agents assume different roles and coordinate task-solving through

structured collaboration or self-collaboration (Wang et al., 2024). This modularity addresses all three structural barriers: verification grounds outputs in reliable sources (cognitive), tool use and APIs facilitate integration with existing systems (socio-technical), and the ability to route simple queries to lightweight components reduces costs (economic). The trade-off is real: orchestrating multiple agents introduces new failure points, and inter-component interactions can generate unpredictable errors that are difficult to debug and audit.

These alternatives are not exempt from structural constraints. RAG adds retrieval, indexing, and data-maintenance costs; neuro-symbolic systems require rule and ontology maintenance, neural-symbolic alignment, and conflict resolution; and compound or multi-agent systems add orchestration overhead, latency, and compound-error risks. Production evaluation should therefore include retrieval recall, source faithfulness, p95 latency, cost per query, component-level error tracking, timeouts, fallback paths, and end-to-end testing. The available evidence supports context-dependent benefits rather than universal superiority.

A Framework for Assessing AI Model Deployment Readiness

High accuracy on standard benchmarks instills false confidence. A model that excels on MMLU or HumanEval benchmarks may fail in a specific enterprise context because benchmarks rarely reflect the complexity of real-world workflows, noisy data, and non-standard queries. Existing AI governance and accountability approaches define regulatory obligations, audit processes, and documentation mechanisms (European Parliament and Council of the European Union, 2024; Raji et al., 2020; Mitchell et al., 2019). The proposed framework complements these approaches by providing a compact deployment-oriented assessment that jointly considers economic viability, reliability and safety, auditability, and organizational readiness. These dimensions are intended as a minimum screening set rather than an exhaustive replacement for established governance frameworks. Therefore, we propose a four-dimensional framework for assessing a model's deployment readiness:

- (1) Economic viability a full total cost of ownership (TCO) analysis, including query volumes, infrastructure needs (GPUs, vector stores, indexing pipelines for RAG systems), and personnel and reskilling costs, compared against realistic estimates of the value generated per query
- (2) Reliability and safety not only average accuracy but also robustness to adversarial inputs, behavior under distribution shift, hallucination rates on domain-specific queries, and calibrated uncertainty/calibrated confidence (the model's ability to produce predictions with probabilities that

match the actual frequency of correctness).

- (3) Auditability documentation of data provenance, bias audits, explainability mechanisms for communicating model decisions to users and auditors, and monitoring dashboards for model drift (the degradation of model effectiveness over time as real-world operational data diverges from training data) and data quality
- (4) Organizational readiness stakeholder analysis, plans for integrating the system into the organization's existing workflows, training programs for personnel who will work with the system, clear accountability structures for errors and undesired system outcomes, assessment of changes to end-user work, and liability analysis for system failures. Systems that pass technical benchmarks but fail to meet organizational readiness criteria create more risks than value and should not be deployed to production

Each dimension is assessed on a five-point ordinal scale. Scores 1, 3, and 5 serve as anchor levels representing critical gaps, partial readiness with unresolved controls, and documented and validated readiness, respectively. Scores 2 and 4 represent intermediate levels and are assigned according to the extent to which the criteria of the adjacent anchor levels are satisfied and supported by documented evidence. As a preliminary screening rule, no dimension should score below 3 before production deployment. Thresholds and weights should be adapted to domain risk, and the rubric requires empirical validation (Table 1).

Fig. 2 shows the framework evaluation of three illustrative profiles: A code assistant (e.g., Copilot strong economic results, low auditability), medical diagnostics (high reliability and auditability, weak economic results), and an internal RAG assistant (balanced results across all dimensions).

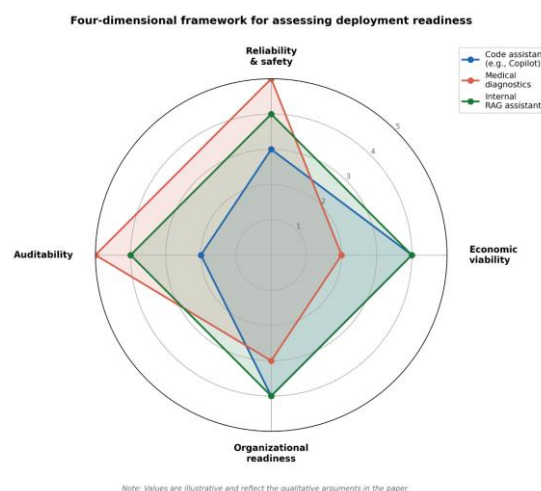


Fig. 2: Four-dimensional framework for assessing AI model deployment readiness

Table 1: Anchor descriptions for the five-point scoring rubric for AI model deployment readiness

Dimension	Score 1	Score 3	Score 5
Economic viability	No TCO or unclear value	Partial TCO; uncertain benefits	Validated TCO and measurable value
Reliability and safety	No domain testing	Basic testing; open risks	Domain validation, monitoring, and fallback
Auditability	No traceability	Partial documentation and logs	Complete provenance, logs, and audit process
Organizational readiness	No owner or workflow plan	Roles defined but incomplete	Clear accountability, training, and integration

Summary of Key Findings and Discussion

Table 2 summarizes the barriers to deploying AI foundation models, the reasons for their persistence, and the alternative approaches analyzed in this work. The reviewed evidence suggests that these barriers are not merely temporary growing pains, but persistent characteristics of the current deployment paradigm. Scaling laws guarantee superlinear cost growth, distributed knowledge representations trade interpretability for broad model capabilities, and institutional inertia is orthogonal to model performance. Empirical evidence accumulated across model generations confirms that benchmark gains do not automatically translate into deployment success.

The analysis shows that several complementary strategies address these barriers from different angles. RAG systems ground the generation process in external, verifiable databases rather than relying entirely on the model's parametric knowledge, thereby reducing hallucinations and lowering computational costs (Asai et al., 2024; Ram et al., 2023). Compound systems distribute tasks among specialized components, reducing dependence on a single monolithic model (Zaharia et al., 2024). None of these strategies is a universal solution each introduces its own trade-offs but together they outline a trajectory beyond simple scaling.

Several recommendations can be made to researchers and practitioners deploying AI foundation models. Researchers should prioritize alternatives to pure scaling.

Intelligent use of external knowledge through RAG, neuro-symbolic integration, uncertainty quantification, and inherently interpretable models (simple machine learning algorithms easily understandable by humans) deserve significantly more attention than they currently receive. Model evaluation frameworks should measure deployment readiness robustness, calibration, source faithfulness, interpretability rather than accuracy on standard benchmarks alone (Fan et al., 2024).

Practitioners should conduct full total cost of ownership analyses before committing to deployment. Starting with human-in-the-loop, low-cost applications – such as information retrieval with human validation or embedded summarization assistants is the most reliable path for organizing AI model deployment. Investments in organizational readiness reskilling, workflow change, clear accountability should at least equal technical investments.

Effective AI system governance should begin with transparency. Model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2021) have already established themselves as standard instruments, but for compound architectures they must be supplemented with documentation for data in every component – provenance, update frequency, coverage, and known data gaps. Systems that rely on external knowledge bases have an additional advantage: The retrieval audit trail allows every response to be traced to specific sources, facilitating independent verification and internal audits (Raji et al., 2020).

Table 2: Summary of structural barriers to AI foundation model deployment

Barrier	Characteristics	Reasons for Failure	Alternative Approaches
Economic	Superlinear cost growth; inference cost growth; energy unsustainability; memory and latency constraints	Costs grow faster than business value; hardware lags behind model size; continuous retraining required	RAG (smaller models using external knowledge sources); tiered pricing; open-source models
Cognitive	Hallucinations; poor calibration; fragile generalization; adversarial vulnerability	Training objectives optimize plausibility, not truth; statistical correlations $\neq$ causation	RAG (grounding in verifiable sources); Self-RAG / RAG-Critic; hybrid neuro-symbolic systems; human-in-the-loop decision-making
Socio-technical	Bias amplification; accountability gaps; integration complexity; organizational resistance	Model improvement does not guarantee organizational deployment readiness; regulatory requirements do not simplify with model advances	RAG (audit trail from data sources); participatory design governance frameworks; gradual organizational adoption

High-stakes domains should require impact assessments with independent third-party audits, drawing on precedents from privacy regulation (Wachter et al., 2017) and environmental impact assessments. Liability rules should be clarified before deployment, not after: Shared responsibility among developers, deployers, and users, combined with mandatory insurance for high-risk applications, would simultaneously incentivize due diligence and ensure compensation for harms.

At the organizational level, the most important shift is cognitive: Acknowledging that structural barriers are not temporary problems that the next model generation will resolve. Deployments should be aligned with foundation models' strengths synthesis, summarization, semantic search, and idea generation and designed around human-AI collaboration rather than full automation (Bansal et al., 2021).

This study is limited by its narrative-review design and its dependence on expert judgment in source selection and thematic classification. The proposed framework has not yet been empirically validated in longitudinal enterprise deployments, and the reviewed evidence is heterogeneous in scope and methodology. Accordingly, the framework should be interpreted as a conceptual decision-support instrument requiring further empirical evaluation.

Conclusion

AI foundation models are a remarkable achievement, but the path from an impressive benchmark to a reliable enterprise system is longer and more complex than many assume. Skepticism toward promises that the next model generation will resolve structural barriers is supported by empirical evidence. The gaps between advertised and actual model capabilities that emerge during deployment are not transient phenomena they reflect fundamental problems in the current paradigm that will not be resolved by updating to the next model.

This paper examined the structural problems facing AI model deployment and proposed a four-dimensional framework for assessing a model's deployment readiness. Future work will validate the framework through comparative enterprise case studies, including code assistants, internal RAG systems, and high-reliability decision-support applications. The four dimensions will be compared with observed outcomes such as total cost, domain-specific errors, audit findings, user adoption, and governance incidents.

Future directions for this research can be oriented along four dimensions.

Architectural. Research should investigate the effective integration of symbolic and neural components at scale, as well as the design of compound systems with clear fault isolation (precisely identifying specific components or causes that led to a system error) when

multiple components (for retrieval, generation, and validation) interact in orchestrated pipelines (Zaharia et al., 2024). Promising directions such as graph knowledge bases (Hu et al., 2025) and multi-agent architectures (Wang et al., 2024) still require systematic evaluation under real enterprise workloads.

Economic. Algorithms with dramatically lower computational requirements for training, sustainable business models, and tools for lifecycle cost management are needed. An important open question is the optimal balance between external and parametric knowledge: Under what conditions is it more efficient to offload knowledge to databases or symbolic systems rather than relying on a larger AI model?

Evaluative. The field needs deployment-oriented benchmarks that measure not only accuracy but also robustness to distribution shift, uncertainty calibration, source faithfulness, quality of attribution (the relative influence of inputs on a model's specific output), and cost-benefit trade-offs (Fan et al., 2024). For compound systems whether using retrieval, symbolic verification, or agent orchestration metrics are needed that assess the reliability of each component and the system as a whole.

Socio-technical. Qualitative empirical studies of real deployment contexts, participatory design methodologies, and AI governance frameworks that translate principles into operational practice remain urgently needed. Of particular importance is the question of how end users build trust in systems whose decision-making mechanisms remain opaque the extent to which they verify the provided justifications and how they react to errors (Bansal et al., 2021).

Acknowledgment

This study is carried out in the frame of the project FP25-FMI-010, "Innovative Interdisciplinary Research in Informatics, Mathematics, and Pedagogy of Education" of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria, 2025-2027.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Hristo Hristov: Contributed to the conceptual framework, research methodology, analysis of AI deployment barriers, supervision of the study, and critical revision of the manuscript.

Elena Somova: Contributed to the methodological design, validation of the analytical framework, academic supervision, manuscript review, and coordination of the final version as corresponding author.

Emil Fuchedzhiev: Contributed to the analysis of socio-technical, regulatory, governance, and security-related aspects of AI deployment, as well as to the review and refinement of the discussion section.

Ethics

This article is based on a narrative synthesis of published academic literature, technical reports, legal documents, and publicly available industry sources. The study does not involve experiments with human participants or animals, and no personal data were collected, processed, or analyzed. Therefore, formal ethical approval was not required.

References

- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. International Conference on Learning Representations (ICLR 2024), Vienna, Austria.
<https://doi.org/10.48550/arXiv.2310.11511>
- Attewell, P. (1987). The Deskilling Controversy. *Work and Occupations*, 14(3), 323–346.
<https://doi.org/10.1177/0730888487014003001>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., & Mirhoseini, A. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv.
<https://doi.org/10.48550/arXiv.2212.08073>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.
<https://doi.org/10.1145/3411764.3445717>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems* (pp. 1877–1901).
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. In *The Quarterly Journal of Economics* (Vol. 140, Issue 2, pp. 889–942). <https://doi.org/10.1093/qje/qjae044>
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT 2018)*, PMLR, 77–91.
- Chatterjee, A., Renduchintala, H. S. V. N. S. K., Bhatia, S., & Chakraborty, T. (2024). POSIX: A Prompt Sensitivity Index For Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14550–14565).
<https://doi.org/10.18653/v1/2024.findings-emnlp.852>
- Dong, G., Jin, J., Li, X., Zhu, Y., Dou, Z., & Wen, J.-R. (2025). RAG-Critic: Leveraging Automated Critic-Guided Agentic Workflow for Retrieval Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3551–3578).
<https://doi.org/10.18653/v1/2025.acl-long.179>
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., & Conerly, T. (2021). A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread, Anthropic. *Transformer Circuits Thread*.
<https://transformer-circuits.pub/2021/framework/index.html>
- European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations and Directives*.
- Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6491–6501).
<https://doi.org/10.1145/3637528.3671470>
- Frantar, E., & Alistarh, D.-A. (2023). SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot. *Proceedings of the 40th International Conference on Machine Learning*, PMLR, 10323–10337.
<https://doi.org/10.48550/arXiv.2301.00774>
- Garcez, A. d’Avila, & Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. In *Artificial Intelligence Review* (Vol. 56, Issue 11, pp. 12387–12406).
<https://doi.org/10.1007/s10462-023-10448-w>
- Gartner. (2024). Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept by End of 2025. In *Gartner Press Release*. <https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025>

- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. In *Communications of the ACM* (Vol. 64, Issue 12, pp. 86–92). <https://doi.org/10.1145/3458723>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning, PMLR*, 1321–1330.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. D. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. V. D., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems 35* (pp. 30016–30030). <https://doi.org/10.52202/068431-2176>
- Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., & Zhao, L. (2025). GRAG: Graph Retrieval-Augmented Generation. *Findings of the Association for Computational Linguistics: NAACL 2025*, 4145–4157. <https://doi.org/10.18653/v1/2025.findings-naacl.232>
- Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. (2024). Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7036–7050. <https://doi.org/10.18653/v1/2024.naacl-long.389>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Leng, Y., Lei, Y., Liu, X., Zhong, M., Xiong, B., Zhang, Y., Gao, Y., Yiwu, Hu, Y., & Xiong, D. (2025). DecEx-RAG: Boosting Agentic Retrieval-Augmented Generation with Decision and Execution Optimization via Process Supervision. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1412–1425. <https://doi.org/10.18653/v1/2025.emnlp-industry.99>
- Lewis, P., Piktus, E., Petroni, A., Karpukhin, F., Goyal, V., Küttler, N., Lewis, H., Yih, M., Rocktäschel, W., Riedel, T., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 9459–9474.
- Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, 24(253), 1–15.
- Mata v. Avianca, Inc. (2003). Opinion and Order on Sanctions. *United States District Court for the Southern District of New York*.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- OpenAI. (2024). Learning to Reason with LLMs. *OpenAI Technical Blog*. <https://openai.com/index/learning-to-reason-with-llms/>
- Orlikowski, W. J. (1996). Improvising Organizational Transformation Over Time: A Situated Change Perspective. *Information Systems Research*, 7(1), 63–92. <https://doi.org/10.1287/isre.7.1.63>
- Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2023). Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Computing Surveys*, 55(6), 1–29. <https://doi.org/10.1145/3533378>
- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519. <https://doi.org/10.1145/3052973.3053009>
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*.
- Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., Xiao, K., Agrawal, S., & Dean, J. (2023). Efficiently Scaling Transformer Inference. *Proceedings of Machine Learning and Systems*, 606–624.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning, PMLR*, 8748–8763.
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>

- Ram, O., Levine, Y., Dalmedigos, I., Muhlgaay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11, 1316–1331. https://doi.org/10.1162/tacl_a_00605
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3411764.3445518>
- Sculley, D., G. Ernst, N., Kazman, R., & Delange, J. (2021). Technical Debt in Machine Learning Systems. *Technical Debt in Practice: How to Find It and Fix It*, 177–192. <https://doi.org/10.7551/mitpress/12440.003.0011>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. <https://doi.org/10.1145/3287560.3287598>
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. <https://doi.org/10.18653/v1/p19-1355>
- Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2023). Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6), 1–28. <https://doi.org/10.1145/3530811>
- Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction APIs. *Proceedings of the 25th USENIX Security Symposium (USENIX Security 16)*, 601–618.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. <https://doi.org/10.1145/3194770.3194776>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, 31(2), 47. <https://doi.org/10.2139/ssrn.3063289>
- Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., & Ji, H. (2024). Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 257–279. <https://doi.org/10.18653/v1/2024.naacl-long.15>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35, 24824–24837. <https://doi.org/10.52202/068431-1800>
- Xiao, C., Cai, J., Zhao, W., Lin, B., Zeng, G., Zhou, J., Zheng, Z., Han, X., Liu, Z., & Sun, M. (2025). Densing law of LLMs. *Nature Machine Intelligence*, 7(11), 1823–1833. <https://doi.org/10.1038/s42256-025-01137-0>
- Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking Retrieval-Augmented Generation for Medicine. *Findings of the Association for Computational Linguistics ACL 2024*, 6233–6251. <https://doi.org/10.18653/v1/2024.findings-acl.372>
- Xu, F. F., Alon, U., Neubig, G., & Hellendoorn, V. J. (2022). A systematic evaluation of large language models of code. *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 6233–6251. <https://doi.org/10.1145/3520312.3534862>
- Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., Zou, J., Carbin, M., Frankle, J., Rao, N., & Ghodsi, A. (2024). The Shift from Models to Compound AI Systems. *Berkeley AI Research*.
- Zhang, S., Pan, Y., Liu, Q., Yan, Z., Choo, K.-K. R., & Wang, G. (2025). Backdoor Attacks and Defenses Targeting Multi-Domain AI Models: A Comprehensive Review. In *ACM Computing Surveys* (Vol. 57, Issue 4, pp. 1–35). <https://doi.org/10.1145/3704725>
- Ziegler, A., Kalliamvakou, E., Li, X. A., Rice, A., Rifkin, D., Simister, S., Sittampalam, G., & Aftandilian, E. (2024). Measuring GitHub Copilot’s Impact on Productivity. In *Communications of the ACM* (Vol. 67, Issue 3, pp. 54–63). <https://doi.org/10.1145/3633453>