International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I7P101

Volume 73 Issue 7, 1-15, July 2025
© 2025 Seventh Sense Research Group®

Review Article

The Rise of Foundation Models in Industry: A Cross-Domain Survey of LLM Applications in Healthcare, Finance, Legal, and Education

Chaithanya Reddy Bogadi

Independent Researcher, Texas, USA.

Corresponding Author : chaithanya.bogadi@ieee.org

Received: 26 May 2025 Revised: 18 June 2025 Accepted: 10 July 2025 Published: 27 July 2025

Abstract - *Large Language Models (LLMs) transform artificial intelligence applications across industries. This paper presents the DOME (Domain-Operation-Model-Evaluation) framework, a novel taxonomy for systematically defining and assessing industrial LLM implementations across healthcare, finance, legal services, and education sectors. Through systematic analysis of published case studies, this research identifies domain adaptability patterns and evaluates performance criteria across sectors. Results reveal significant maturity variations: healthcare and finance demonstrate advanced implementations, while legal and Education sectors remain largely experimental. The study identifies critical deployment constraints, including hallucinations, domain adaptation complexity, and regulatory barriers. Key emerging trends include retrieval-augmented generation systems, domain-specific fine-tuning, and edge computing paradigms. The framework provides researchers and practitioners with systematic guidelines for responsible AI deployment in regulated industries, advancing the field's understanding of effective cross-domain LLM implementation strategies.*

Keywords - *Artificial intelligence, Domain adaptation, Evaluation framework, Foundation models, Large language models.*

1. Introduction

Foundation models, particularly Large Language Models (LLMs), have revolutionized artificial intelligence applications across multiple industries [11]. These models, trained on extensive text datasets, have exhibited exceptional proficiency. These models, trained in extensive text datasets, have demonstrated a remarkable ability to understand and generate human language, enabling advanced applications previously considered unfeasible[12].

This survey examines LLMs' specific applications, challenges, and opportunities in four critical industrial sectors: healthcare, finance, legal services, and Education, despite the prevailing emphasis of academic research on general-purpose capabilities and benchmarks.

The deployment of LLMs in high-stakes and regulated sectors presents unique challenges that transcend generic technological concerns. Industry-specific challenges that must be addressed include domain-specific language, regulatory compliance, liability considerations, and workflow integration [86]. The risks and benefits of LLM implementation differ markedly between applications, such as clinical documentation, algorithmic trading, legal contract analysis, and personalized Education.

The paper presents the following contributions.

1. While the previous works have explored LLM applications in cross-domain and general-purpose tasks, this paper presents the DOME framework, an innovative taxonomy for discovering and evaluating LLM applications across diverse industries, including Domain context, Operational purpose, Model characteristics, and Evaluation criteria. However, it does not introduce any new algorithm by combining domain-focused adaptations, regulatory considerations, and evaluation metrics, as this study helps bridge the gap between academic benchmarks and operational requirements in the industry.
2. Offers a comprehensive analysis of more than 60 industrial applications throughout the healthcare, banking, legal, and Education sectors, evaluating adoption patterns, implementation strategies, and sector-specific adaptations.
3. The study defines and examines the essential evaluation metrics and benchmarks utilized to assess LLMs in industrial settings, highlighting the disparity between academic benchmarks and practical performance demands.
4. This study investigates domain-specific obstacles that include legislative limitations, requirements for

explainability, and the various conceptions of “success” in several businesses.

5. The survey examines growing trends and future

Directions for foundation models in industry, encompassing domain-specific fine-tuning tactics, multimodal capabilities, and the progression towards more specialized, vertical applications.

The paper is organized as follows. Section 2 reviews related work and background. Section 3 introduces the Proposed DOME framework. Sections 4-7 examine the applications of LLM in the healthcare, financial, legal, and Education sectors, respectively. Section 8 addresses evaluation criteria and benchmarks. Section 9 analyzes constraints and hazards, and Section 10 investigates future trends and research trajectories. Section 11 concludes the paper.

2. Literature Survey and Background

2.1. Evolution of Foundation Models

Foundation models signify a transformative change in AI development, marked by extensive pretraining on substantial datasets that succeeded in adapting to certain downstream tasks[11]. The progression from older models such as ELMo[63] and BERT [25] to modern LLMs such as GPT-4 [62], LLaMA [76], and PaLM [19] has been characterized by significant expansions in model size, computational demands, and functionalities.

2.2. Prior Surveys and Taxonomies

Numerous surveys have analyzed foundation models from various perspectives. Bommasani et al. [11] presented a thorough analysis of the potential and dangers. Wei et al. [84] reported the emergence of capabilities in big language models. Liu et al. [55] examined rapid engineering methodologies. Nevertheless, these efforts predominantly emphasize general capabilities rather than industry-specific applications. Several studies have focused on domain-specific applications. Thirunavukarasu et al. [74] conducted a survey on LLMs in the healthcare sector, while Huang et al. [36] investigated financial applications. Cui et al. [23] investigated adaptations within the legal domain. In Education, Kasneci et al. [43] examined the impact of LLMs on teaching and Learning. The research distinguishes itself by offering an integrated cross-domain analysis and presenting a thorough taxonomy relevant to many businesses.

2.3. Patterns of Industry Adoption

The adoption of foundation models within the industry has shown distinct tendencies. Initial implementations concentrated on regulated use cases with human supervision, including content generation support and summarizing activities [46]. As models have advanced, application functions have evolved to encompass increasingly autonomous capabilities, including intricate reasoning tasks and interactive decision support[13].

Deployment methods have evolved, and early adopters have predominantly used API access to proprietary models, whereas current trends indicate a growing interest in open-source models that can be optimized and implemented in private infrastructure to meet data privacy and customization needs [92].

3. The DOME Framework

This paper introduces (Domain–Operation–Model–Evaluation), a four-dimensional taxonomy that identifies and assesses LLM applications within high-stakes businesses. This framework fulfils the need for a systematic approach to comprehending how foundation models are tailored to meet industry-specific demands.

3.1. Domain Context

The Domain dimension encompasses the operational environment within which a large language model functions:

3.1.1. Industry Vertical

Healthcare, finance, legal, Education, and subdomains within each.

3.1.2. Regulatory Environment

Applicable regulations such as HIPAA, GDPR, FERPA, and sector-specific requirements.

3.1.3. Task Criticality

The potential impact of model errors or failures, ranging from low (informational content) to high (diagnosis support, financial decisions).

3.1.4. Data Sensitivity

The characteristics and sensitivity of processed data, including Protected Health Information (PHI), Personally Identifiable Information (PII), financial records, and academic data.

3.2. Operational Objective

The Operation dimension defines the functional role of LLMs in business processes.

3.2.1. Extractive Tasks

Information extraction, Named Entity Recognition (NER), and classification functions.

3.2.2. Generative Tasks

Content creation, summarization, report generation, and other text production functions.

3.2.3. Interactive Tasks

Conversational interfaces, virtual assistants, and interactive decision support.

3.2.4. Reasoning Tasks

Complex analysis, multistep inference, and knowledge synthesis capabilities.

3.3. Characteristics of the Model

The Model dimension pertains to the technical facets of LLM execution:

3.3.1. Model Type

Base models, instruction-tuned models, or domain-specific fine-tuned models.

3.3.2. Access Pattern

Open-source models, proprietary APIs, or enterprise-hosted implementations.

3.3.3. Degree of Customization

Varying from rapid engineering to few-shot Learning, fine-tuning, or retrieval augmented generation (RAG).

3.3.4. Deployment Architecture

Cloud-based or on-premises edge installations, taking into account latency, privacy, and scalability.

3.4. Evaluation and Ethics

The evaluation dimension includes metrics and methodologies for evaluating LLM performance and ensuring responsible utilization:

3.4.1. Performance Metrics

Metrics tailored to certain tasks (e.g. BLEU, ROUGE, F1 score) and evaluation criteria pertinent to particular domains.

3.4.2. Explainability

Criteria and methodologies for model transparency and decision traceability.

3.4.3. Bias and Fairness

Evaluation of potential biases and methodologies to ensure equitable performance in demographic groupings.

3.4.4. Compliance

Strategies for guaranteeing conformity to regulatory mandates and ethical principles

DOME			
Domain	**Operation**	**Model**	**Evaluation**
Industry, Regulation Criticality, Data	Extractive, Generative Interactive, Reasoning	Type, Access, Customization, Deployment	Metrics, Explainability, Fairness, Compliance

Fig. 1 The DOME Framework for classifying and evaluating LLM applications in industry settings

3.5. Case Study Selection Methodology

This study uses a systematic methodology to identify and examine LLM applications in four specified industry sectors. The subsequent selection criteria were implemented:

3.5.1. Source Diversity

The applications were sourced from the peer-reviewed literature (n = 24), industry white papers (n=18), and verified commercial deployments (n =18) between 2021 and 2023 to guarantee comprehensive coverage in research and practice.

3.5.2. Implementation Maturity

The chosen cases exhibit tangible implementation specifics that extend beyond theoretical proposals, accompanied by at least initial evaluation outcomes.

3.5.3. Domain Balance

An almost equal representation was achieved in healthcare (n = 16), finance (n = 15), law (n =14), and Education (n = 15).

3.5.4. Application Diversity

Cases were chosen within each area to exemplify the spectrum of operational objectives specified in the DOME framework.

This selection approach possesses specific restrictions. Prominent deployments (e.g., Med-PaLM, BloombergGPT) They are probably overrepresented due to superior documentation and public disclosure. Moreover, successful implementations may be disproportionately represented relative to unsuccessful attempts, as unfavorable outcomes are disclosed less frequently.

The identified biases are recognized as constraints of the present analysis.

4. LLM Applications in Healthcare

4.1. Current Landscape

Healthcare constitutes a highly promising yet challenging domain for LLM applications, characterized by data-rich environments, specialized terminology, and stringent regulatory constraints [82].

Applications encompass clinical decision support, documentation assistance, patient engagement, research acceleration, and operational efficiency improvements.

4.1.1. Clinical Documentation

Large Language Models show great promise for automating healthcare documentation. Systems such as DAG-NLP [2] and Nuance DAX [41] employ LLMs to transform unstructured clinical dialogues into structured documentation.

Evidence indicates that these applications can alleviate physician burnout by reducing documentation time by 30 to 50 percent [67]. However, apprehensions about hallucination and precision in essential paperwork continue to pose substantial obstacles to wider implementation [74].

4.1.2. Diagnostic Assistance

Foundation models have shown encouraging proficiency in diagnostic support. Med-PaLM 2 attained expert-level proficiency on medical license examination queries [71], while ChatDoctor [50] showed strong performance in simulated patient interactions.

Recent work by Nori et al. [61] has revealed that GPT-4 can achieve competitive performance in diagnostic reasoning tasks relative to medical specialists, especially when enhanced with retrieval mechanisms.

4.1.3. Medical Research and Literature Analysis

Large language models expedite medical research via applications like BioGPT [56] and PubMedGPT [10], which assist in literature review, hypothesis generation, and protocol development. Systems like BEAGLE [49]utilize foundation models to extract structured information from biomedical literature on a large scale, facilitating expedited knowledge synthesis and discovery.

4.2. Domain-Specific Adaptations

Effective healthcare applications require substantial domain adaptation, encompassing:

4.2.1. Medical Vocabulary Enhancement

Models like Clinical-BERT [5] and BioMedLM [9] bridge the specialized terminology gap via domain-specific pretraining or fine-tuning.

4.2.2. Retrieval-Augmented Generation

Systems like Med PaLM [71] integrate large language models with medical knowledge sources and literature retrieval to mitigate hallucinations and improve factual accuracy.

4.2.3. Structured Output Formatting

Applications impose organized outputs consistent with healthcare documentation standards (e.g., SOAP notes, ICD-10 codes) [77].

4.2.4. Explainability Mechanisms

Chain-of-thought prompting, incorporating citation methods, improves transparency and logical reasoning in clinical settings[78].

4.3. Regulatory and Ethical Considerations

Healthcare LLM applications have distinct regulatory challenges:

4.3.1. FDA Regulation

Applications offering clinical decision support may require FDA approval as medical devices, exemplified by Aidoc’s AI solutions receiving approval through the 510(k) process [24].

4.3.2. HIPAA Compliance

Implementations must guarantee the protection and security of Protected Health Information (PHI), frequently requiring on-premise or HIPAA-compliant cloud implementations compared to conventional API access [66].

4.3.3. Liability Frameworks

Questions regarding accountability when LLM outcomes lead to medical errors remain largely unresolved, with some frameworks proposing hierarchical responsibility models[29].

4.3.4. Transparency Requirements

Medical applications often necessitate higher levels of explanation than consumer applications, accompanied by clear justifications and empirical evidence [70].

5. LLM Applications in Finance

5.1. Current Landscape

Financial services have become prominent adopters of Large Language Model (LLM) technology, driven by data-centric processes, measurable outcomes, and significant efficiency opportunities[90]. Applications encompass market analysis, risk assessment, customer service, and regulatory compliance.

Table 1. Healthcare LLM Applications by Operational Purpose

Purpose	Application	Key Features
Extractive	Clinical NER EHR data extraction [67] for structured analysis Medical concept Mapping unstructured text extraction[56]	EHR data extraction [67] for structured analysis Medical concept Mapping unstructured text extraction [56] to medical ontologies.
Generative	Radiology report drafting [3] Patient education materials [88]	Generating Structured reports from Imaging Findings: Creating accessible health information materials
Interactive	Virtual health assistants [50] Clinical dialogue summaries [41]	Symptom assessment and triage Recommendations Augmenting patient doctor conversation and documentation
Reasoning	Differential diagnosis support [61] Treatment planning[78]	Multistep Diagnostic reasoning assistance Evidence-based therapy Recommendation with citations.

5.1.1. Market Intelligence and Analysis

Large Language Models are revolutionizing financial analysis through applications that synthesize market information at an unprecedented scale. BloombergGPT[90], a domain-specific LLM trained in financial data, exhibits exceptional performance on specific financial tasks compared to general-purpose models.

Systems such as Alpha-GPT [52] facilitate natural language inquiries spanning financial databases that facilitate equitable access to market insights.

5.1.2. Risk Assessment and Fraud Detection

Financial institutions are utilizing LLMs for risk assessment and fraud detection. JPMorgan's IndexGPT [42] evaluates complex financial documents to assess investment risks, while Goldman Sachs has experimented with LLM-based systems to analyze irregular transaction patterns[53].

These applications integrate structured data analysis with unstructured documents, conducting analyzes to provide thorough risk profiles.

5.1.3. Customer Experience and Advisory

Customer-oriented LLM applications include Morgan Stanley's wealth management assistant[60], which provides advisors with immediate access to research and tailored customer recommendations.

Similarly, Bank of America Erica [8] uses NLP capabilities to answer customer queries and provide personalized financial guidance through natural conversation.

5.2. Domain-Specific Adaptations

Financial LLM applications have implemented significant domain-specific modifications:

- Financial Corpora Training: Models like BloombergGPT [90] and FinBERT [6] integrate domain-specific pretraining using financial records, earnings calls, and regulatory filings to improve understanding of financial terminology.
- Numerical Reasoning: Applications utilize specialized techniques for prompting and external instruments to enhance reasoning in financial computations and numerical analysis [91].
- Temporal Awareness: Financial applications require mechanisms to manage model knowledge cutoffs, especially for market data and regulatory changes, which are frequently executed by retrieval augmented generation by utilizing current sources[18].
- Compliance Guardrails: Systems integrate specialized filters and limitations to ensure outputs adhere to financial standards and prevent unauthorized investment advice[28].

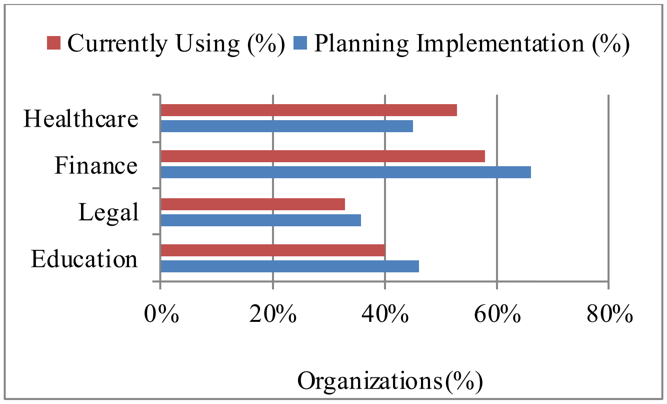

Fig. 2 Organizational adoption of AI and language models by industry sector, based on data from the Stanford HAI AI Index Report 2023 [94] and McKinsey's State of AI 2023 survey [95].

5.3. Regulatory and Ethical Considerations

Financial LLM applications operate in a highly regulated environment with several considerations:

5.3.1. Market Manipulation Risks

LLM outputs discussing financial instruments must carefully avoid language that could be interpreted as market manipulation or unauthorized investment advice [86].

5.3.2. Fairness in Lending

Applications that influence credit decisions must demonstrate compliance with fair lending regulations and avoid discriminatory outcomes [27].

5.3.3. Audit Requirements

Financial applications typically require comprehensive audit trails of model decisions, particularly for regulatory reporting and compliance verification[93].

5.3.4. Data Privacy Frameworks

Applications must navigate complex global data privacy regulations, including GDPR, CCPA, and sector-specific requirements [51].

6. LLM Applications in Finance

6.1. Current Landscape

Legal services have seen transformative applications of LLM, addressing the industry's document-intensive workflows and complex reasoning requirements [34]. Applications encompass document analysis, contract management, legal research, and drafting support.

6.1.1. Document Analysis and Due Diligence

LLM-enabled document analysis has demonstrated considerable efficiency improvements in due diligence procedures. Systems such as Harvey AI [33] may scrutinize several contracts to extract essential terms, identify hazards, and highlight abnormalities. A JPMorgan Chase research indicated that their COIN (Contract Intelligence) system, augmented with NLP capabilities, decreased legal document review time from 360,000 hours to just seconds [20].

6.1.2. Document Analysis and Due Diligence

Legal research applications like Lexis+ AI [57] and Westlaw Edge [75] leverage LLMs to improve search relevance, summarize case law, and analyze legal arguments. CaseText's CoCounsel [15] combines LLMs with legal databases to provide cited answers to complex legal questions, demonstrating performance comparable to experienced attorneys on certain research tasks [44].

6.1.3. Legal Research and Reasoning

LLM applications have improved contract lifecycle management by optimizing drafting, negotiating, and analyzing processes. Platforms such as ContractPodAi [22] and Spellbook[72] produce contract clauses from natural language descriptions, propose alternative provisions, and identify potential compliance concerns.

6.2. Domain-Specific Adaptations

Legal LLM applications have implemented a notable domain adaptations:

- Legal Corpus Training: Models like LexiLaw [23] and LegalBERT [16] undergo specialized training on legal corpora, including case law, statutes, and scholarly articles.
- Jurisdiction-Aware Prompting: Applications integrate jurisdiction-specific limitations and context to deliver legally pertinent responses based on applicable legislation[34].
- Citation Generation: Legal applications implement mechanisms to generate proper legal citations and reference authoritative sources, enhancing credibility and traceability [87].
- Precedent Retrieval: Systems integrate LLMs with vector databases of legal precedents to substantiate responses with established jurisprudence and regulatory directives[23].

6.3. Regulatory and Ethical Considerations

Legal LLM applications encounter several considerations:

- Unauthorized Practice of Law: Applications must carefully navigate the boundary between information provision and legal counsel to avoid violating restrictions against unauthorized practice [44].
- Attorney-Client Privilege: Systems that manage confidential legal communications must preserve privilege protections and regulations, often requiring stringent data management practices[87].
- Liability for Errors: Issues of liability assignment and issues pertaining to LLM-generated legal content remain unresolved, with consequences for professional malpractice[34].
- Ethical Obligations: Applications must uphold attorneys' ethical responsibilities. Professional responsibility encompasses responsibilities of proficiency, confidentiality, quality, and oversight of technology [79].

7. LLM Applications in Education

7.1. Current Landscape

Educational applications of LLMs encompass instructional assistance, personalized Education, assessment, and administrative support[43].

These apps have generated much attention about improvement in Learning and issues with academic integrity.

7.1.1. Personalized Tutoring and Learning

LLM-based tutoring systems provide personalized Learning experiences through adaptive instruction and feedback. Applications like Khan Academy's Khanmigo [45] and Duolingo Max [26] deliver customized explanations, practice problems, and conversational learning experiences.

Research by Baidoo Anu et al. [7] suggests these systems can increase student engagement and learning outcomes, particularly for self-directed learners.

7.1.2. Instructional Support and Materials

Educators are using large language models to create teaching resources and instructional programs. Systems such as Cognii[21] and Quizizz [64] help teachers generate diverse learning resources, formative assessments, and differentiated content. These tools allow educators to focus on high-value interactions by automating routine content creation tasks [43].

7.1.3. Assessment and Feedback

LLM applications are revolutionizing assessment methodologies through automated evaluation accompanied by comprehensive feedback.

Gradescope[31] employs language models to assess written responses and provide constructive feedback on scale.

These systems have exhibited correlations with human assessment in subjective evaluations ranging from r = 0.70 to r= 0.85, depending on the type of question [82].

7.2. Domain-Specific Adaptations

Educational LLM applications have implemented several adaptations:

7.2.1. Age-Appropriate Communication

These systems adjust reading limitations and vocabulary modifications grounded in educational attainment and the development phase [43].

7.2.2. Scaffolded Learning

Applications employ techniques such as progressive hint giving and guided exploration rather than the production of immediate responses [59].

7.2.3. Pedagogical Framework Alignment

Systems align responses incorporating recognized learning theories and educational approach methodologies, including Bloom's taxonomy or constructivist theory principles [30].

7.2.4. Learning Style Adaptation

Applications modify explanations of national approaches predicated on learner preferences and observations, Patterns of comprehension [82].

7.3. Regulatory and Ethical Considerations

Educational LLM applications encounter several distinct considerations:

7.3.1. Student Data Privacy

Applications must adhere to regulations such as FERPA in the United States and similar protections around the world when processing student data information [39].

Academic Integrity

Systems must equilibrate educational support to express apprehensions about the facilitation of academic dishonesty, often employing detection techniques and usage policies [59].

Equitable Access

Implementation of educational technology must address issues over the digital divide and guarantee accessibility across socioeconomic strata[68].

Developmental Appropriateness

Younger learners must integrate age-appropriate safety measures, such as sentinels and content moderation systems[43].

8. Evaluation Criteria and Benchmarks

8.1. Task-Level Metrics

Task-level evaluation measures gauge the immediate quality of LLM outputs for designated functions:

8.1.1. Text Quality Metrics

Automated scores evaluate generated text quality; however, these metrics often correlate poorly with human assessments on complex tasks [80].

8.1.2. Domain-Specific Benchmarks

Specialized assessments like MedQA [40] for medical knowledge, FinQA [17] for financial reasoning, and LegalBench [32] for legal tasks provide more relevant assessments than general NLP benchmarks.

8.1.3. Human Evaluation

Expert assessment is the benchmark criterion for high-stakes applications, with organization techniques to assess precision, relevance, and utility [61].

Table 2. LLM Implementation Patterns from Published Case Studies (2021-2023)

Domain	Primary Deployment Models	Common Adaptation Methods	Reference Studies
Healthcare	API access (limited), private cloud deployment	RAG with medical knowledge bases, fine-tuning	Singhal et al. [71], Nori et al. [61
Finance	Private infrastructure, API with strict safeguards	Fine-tuning on financial corpora, tool integration	Wu et al. [90],Lieber man et al. [97]
Legal	Cloud API (research), private deployment (prac- tice)	Retrieval Over case law, citation mechanisms	Guha et al. [32], Henderson et al. [34]
Education	Primarily API- based, edge deployment for privacy	Prompt engineering RAG with educational content	Kasneci et al. [43], Zhang et al. [92]

Note: This table synthesizes patterns from published case studies and does not represent a comprehensive statistical analysis. Each domain shows considerable variation in implementation approaches.

8.2. Model-Level Metrics

Model-level metrics assess overarching system attributes:

8.2.1. Inference Efficiency

Latency measurements, through implementation, and the computational cost affect deployment viability, especially for real-time applications [92].

8.2.2. Hallucination Rate

Validation of claims is essential in regulated sectors, particularly in technology. Methods used include citation validation and factual accuracy assessment [58].

8.2.3. Robustness

The consistency of outputs across prompt fluctuations and adversarial input affects operational reliability [85].

8.3 Cross-Domain Meta-Analysis

The analysis identifies benchmark performance from domains to find trends in LLM capabilities. As demonstrated in Table 3 and Figure 3, several key patterns emerge:

8.3.1. Scaling Patterns

Performance exhibits monotonic improvement with increasing model size across all domains, albeit with diminishing returns beyond 70 billion parameters in legal and educational domain tasks.

8.3.2. Domain Difficulty

Legal reasoning consistently demonstrates decreased performance across all dimensions of the model. The domain poses distinct issues necessitating specialization, cognition, and intricate reasoning.

8.3.3. Domain-Specific Optimization

Domain-specialized models such as Med-PaLM in healthcare and BloombergGPT in finance demonstrate 5 to 12 percent performance gains over general-purpose models of comparable size on domain-specific benchmarks.

8.3.4. Transfer Learning Efficiency

Models that demonstrate robust performance in healthcare activities often demonstrate comparable efficacy in educational tasks, indicating possible common reasoning principles.

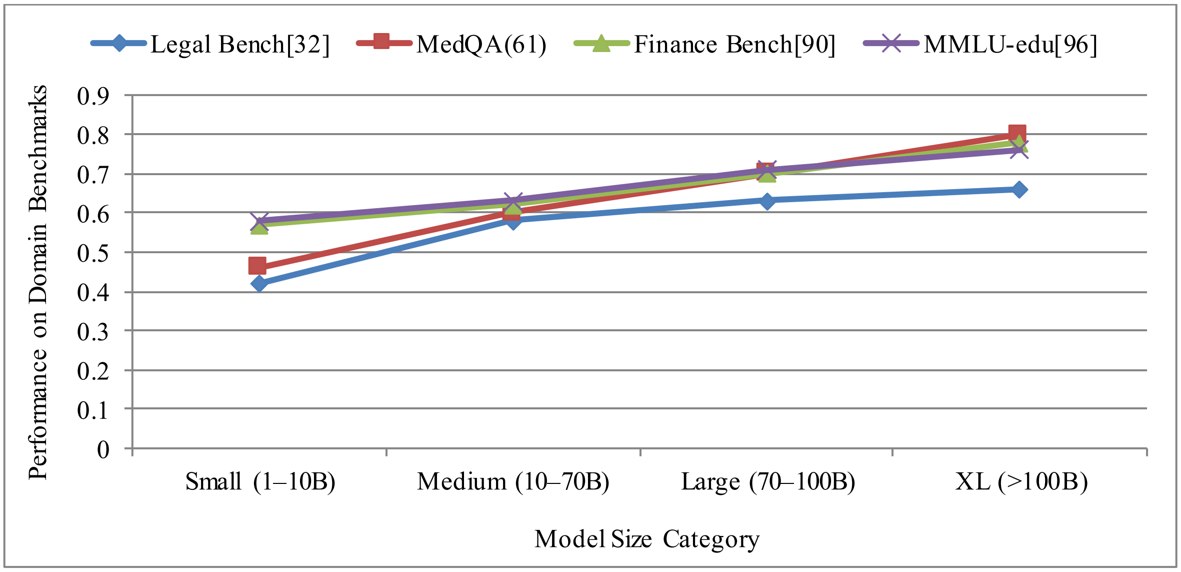

Fig. 3 Performance comparison across domain-specific benchmarks for models of different parameter sizes. Data compiled from published benchmark results in Nori et al. [61], Wu et al. [90], Guha et al. [32], and Hendrycks et al. [96]. Results normalized to a 0-1 scale for comparison

Table 3. Cross-Domain Benchmark Performance Comparison

Model Size	Model	Med QA (%)	Finance Bench (%)	Legal Bench(%)	MMLU-Edu (%)
7-10B	LLaMA- 7B	46	51	40	54
	Flan-T5-XL	44	53	38	56
10-70B	LLaMA13B-	58	62	52	62
	PaLM-32B	62	66	56	67
70-100B	GPT-3.5	72	71	63	71
	LLaMA-70B	70	69	62	70
> 100B	GPT-4	81	77	68	78
	PaLM-2	83	75	67	76

This meta-analysis demonstrates that although general scaling tendencies are applicable across several domains, performance specific to each domain persists. Significantly shaped by knowledge prerequisites and cognitive frameworks.

Data normalized and compiled from Nori et al. [61], Wu et al. [90], Guha et al. [32], and Hendrycks et al. [96]. All values are approximate and normalized to facilitate direct comparison.

8.4. System-Level Metrics

System-level metrics evaluate the integrated application performance:

8.4.1. Retrieval Quality

In retrieval-augmented systems, metrics such as Recall@K, Mean Reciprocal Rank (MRR), and Relevance Score evaluate the efficacy of information retrieval. [73].

8.4.2. User Satisfaction

Metrics such as participation, repeat usage, and direct feedback offer empirical validation of the effectiveness of the system [55].

8.4.3. Business Impact

Measurable indicators such as time efficiency, error mitigation, and return on investment convert technical performance into business value [20].

9. Limitations and Risks

9.1. Technical Limitations

Despite their capabilities, LLMs face significant limitations in industrial applications:

9.1.1. Hallucinations

The creation of credible, yet inaccurate, information continues to be a significant difficulty, especially in high-stakes areas [38]. Research indicates that hallucination rates vary between 5 and 35 percent, contingent upon the specific domain and task involved [58].

9.1.2. Knowledge Cutoffs

Pre-trained models possess static knowledge limited to their training cutoff date, necessitating additional input for dynamic fields such as finance and law [18].

9.1.3. Reasoning Limitations

Current models struggle with complex multistep reasoning, especially in relation to numerical calculations [83].

9.1.4. Contextual Constraints

Restricted context windows limit the amount of information that can be processed in a single inference, posing significant challenges for document-intensive domains such as law[54].

9.2. Implementation Challenges

Practical deployment challenges include:

9.2.1. Integration Complexity

Linking LLMs to enterprise systems and workflows requires considerable engineering effort, especially with regard to legacy systems[92].

9.2.2. Computational Resources

Large-scale model inference remains computationally expensive, limiting deployment options for resource-constrained organizations [35].

9.2.3. Expertise Requirements

Inference of large-scale models is computationally intensive, constraining deployment possibilities for organizations with limited resources [55].

9.2.4. Ongoing Maintenance

LLM applications necessitate perpetual monitoring and updates to ensure alignment with evolving business requirements and developing capabilities [11].

9.3. Emerging Liability Frameworks

Several liability frameworks have been proposed as LLMs are integrated into high-stakes decision-making processes.

9.3.1. Shared Responsibility Models

Specialized professionals are designing cyber and liability insurance frameworks to address claims arising from the use of large language models (LLMs). These frameworks allocate responsibilities among stakeholders, with insurance premiums influenced by model risk assessments, governance mechanisms, and human oversight protocols [29].

9.3.2. Insurance-Based Approaches

To mitigate risks associated with LLM deployment, experts are formulating cyber and liability insurance solutions. These insurance models consider factors such as model evaluation, organizational governance, and human-in-the-loop controls when determining premium structures [27].

9.3.3. Certification Standards

Third-party certification frameworks that assess model safety, reliability, and compliance are being formulated. Domain-specific guidelines are being formulated to set minimum standards for professional utilization [65].

9.3.4. Contractual Risk Allocation

Enterprise installations increasingly employ advanced contractual frameworks that clearly delineate risk between model providers, integrators, and end consumers [87].

These frameworks are still in their infancy, yet indicate the need for tailored approaches to liability allocation rather than generic solutions.

Table 4. Comparison of regulatory frameworks across domains

Legal Aspect	Healthcare Education	Finance
Primary Bar Association Regulation Rules, data privacy	HIPAA (US), FEPRA(US) GDPR(EU), COPPA, GDPR	GDPR, MiFID II, Basel III
Data Privacy Attorney-client Requirements privilege	Protected Health Educational Information (PHI) Records (FERPA)	Personal financial information
Liability Legal Malpractice, Framework unauthorized practise	Medical Malpractice, Professional FDA oversight negligence	Fiduciary duty, regulatory fines
Explainability High: legal requirements reasoning	High: clinical: Moderate: decisive on justification pedagogical transparency	Moderate-High: audit requirements
Proposed bar Association Solutions’ ethical guidelines	FDA Pre-Cert age appropriate medical AI standard design frameworks	Model risk management frameworks

Table 5. Cross-Domain Comparison of Key Implementation Challenges

Challenge Legal	Healthcare Education	Finance
Knowledge Jurisdictional law, Gaps precedents, statutes	Medical vocabulary, Educationional standards, clinical guidelines subject expertise	Financial instruments, market regulations
Critical Citation Accuracy, Failure Modes, and Legal Interpretation	Diagnostic accuracy, Age-appropriate treatment advice content, bias	Calculation precision, risk assessment
Human-AI Attorney Oversight, Integration document review	Clinical workflow Teacher in the loop, integration Grading assistance	Analyst augmentation vs. automation
Data Attorney-client, Constraints, privilege concerns	PHI restrictions, Student privacy, siloed systems FERPA compliance	Financial data access Limitations
Evaluation Legal Accuracy, Challenges Procedural correctness	Clinical expertise Learning outcomes, requirements pedagogical alignment	Quantitative outcomes, regulatory Compliance

9.4. Ethical and Societal Risks

Broader concerns include

9.4.1. Bias Amplification

Models may replicate or exacerbate the biases inherent in training data, resulting in discriminatory results in sensitive areas [86].

9.4.2. Privacy Vulnerabilities

Models may retain and potentially disclose sensitive information from training data, particularly concerning healthcare and finance [14].

9.4.3. Overreliance

Users may demonstrate automation bias, over-trusting model outputs even when incorrect [37].

9.4.4. Labor Market Impacts

Automation of knowledge-work tasks raises concerns about workforce displacement and changing skill requirements [1].

10. Future Trends and Research Directions

10.1. Technical Evolution

Several promising technical directions are emerging:

10.1.1. Domain-Specific Foundation Models

Vertical models like BloombergGPT [90], Med-PaLM [71], and LexiLaw [23] demonstrate the value of domain-specialized pretraining and architecture.

10.1.2. Retrieval-Augmented Generation (RAG)

The integration of large language models with knowledge bases and vector databases facilitates more current and factual responses, essential for regulated industries [48].

10.1.3. Tool-Using Agents

LLMs with external tools (calculators, databases, API calls) exhibit improvement in competencies for intricate tasks requiring a factual foundation [69].

10.1.4. Multimodal Capabilities

The amalgamation of text and visuals, structured data facilitates more thorough applications, especially in healthcare (medical imaging) and financial chart analysis [4].

10.2. Deployment Paradigms

Evolving deployment approaches include the following.

10.2.1. On-Device Inference

Compact, distilled and quantized models provide local inference, preserving privacy without data transmission [47].

10.2.2. Hybrid Architectures

Integrations of on-premise enterprise models with cloud-based functionalities optimize privacy, cost, and capabilities [92].

10.2.3. LLM Orchestration

Intricate workflows integrating Many specialized models for different subtasks demonstrate-Strategize the potential of enterprise applications [89].

10.2.4. Human-AI Collaboration

Systems specifically engineered for augmented intelligence rather than automation maximization [81].

10.3. Governance Frameworks

Emerging governance approaches include

10.3.1. Responsible AI Policies

Organizations are building complete governance frameworks for the implementation of foundation models, focusing on explainability, fairness, and safety. [65].

10.3.2. Regulatory Adaptation

Regulatory bodies are formulating new frameworks explicitly targeting AI and LLM applications, like the EU AI Act and the FDA's proposed software regulations. [27].

10.3.3. Auditing Mechanisms

Emerging third-party audit frameworks and technical tools for bias detection, factual Accuracy verification and compliance assessment are being Developed[65].

10.3.4. Industry Standards

Domain-specific standard bodies are establishing benchmarks and best practices for the responsible deployment of LLMs [65].

11. Conclusion

This survey has analyzed the landscape of foundation model applications in healthcare, finance, legal and Education sectors, introducing the DOME framework as a systematic method to understand and assess industry-specific LLM executions. The analysis identifies prevalent patterns and specialized adaptations defining effective implementations within regulated sectors. Multiple significant conclusions arise from the cross-domain examination. Initially, industry-specific applications are progressively transitioning from generic functionalities to specialized methodologies, including the implementation of domain-specific pretraining, retrieval augmentation, and human-in-the-loop methodologies. Secondly, evaluation frameworks should transcend conventional NLP metrics, including subject expertise, business effect assessments, and compliance factors. Third, regulatory and ethical considerations differ greatly between domains, yet exhibit common characteristics of clarity, equity, and suitable human supervision. Future research should investigate several promising avenues: (1) standardized evaluation frameworks for industry-specific LLM applications that reconcile technical performance with domain-specific requirements; (2) architectural strategies for the efficient integration of domain knowledge with foundational model capabilities;(3) governance frameworks that address the distinct risks and requirements of high-stakes applications; and (4) human-AI collaboration paradigms that optimize complementarity. Leverage complementary capabilities while alleviating the risks associated with automation bias and excessive dependence. As foundation models advance and expand throughout various sectors, systematic methods for understanding their applications, constraints, and governance necessities will become increasingly essential. The DOME framework that offers this analysis serves as a foundational reference, providing researchers and practitioners with a unified terminology to discuss and evaluate LLM implementations in regulated and high-stakes contexts.

References

[1] Daron Acemoglu, and Pascual Restrepo, “Tasks, Automation, and The Rise in US Wage Inequality,” *Econometrica*, vol. 90, no. 5, pp. 1973–2016, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Zabir Ai Nazi, and Wei Peng, “Large Language Models in Healthcare and Medical Domain: A Review,” *Informatics*, vol. 11, no. 3, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Amir Ali Rahsepar, “Large Language Models for Enhancing Radiology Report Impressions: Improve Readability While Decreasing Burnout,” *Radiology*, vol. 310, no. 3, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jean-Baptiste Alayrac et al., “Flamingo: A Visual Language Model for Few-shot Learning,” *Advances in Neural Information Processing Systems*, vol. 35, 2022. [Google Scholar] [Publisher Link]

[5] Emily Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” *arXiv Preprint*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Dogu Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” *arXiv preprint arXiv:1908.10063*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] David Baidoo-Anu, and Leticia Owusu Ansah, “Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Education,” *Journal of AI*, vol. 7, no. 1, pp. 52-62, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Bank of America, Erica, The Guide by Your Side, is Here for You. [Online]. Available: http://info.bankofamerica.com/en/digital-banking/erica

[9] Yu Gu et al., “Domain-specific Language Model Pretraining for Biomedical Natural Language Processing,” *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hao Liu, “PubMed GPT: a Domain-Specific Large Language Model for Biomedical Text,” *Medium*, 2023. [Publisher Link]

[11] Rishi Bommasani et al., “On the Opportunities and Risks of Foundation Models,” *arXiv preprint arXiv:2108.07258*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tom Brown et al., “Language Models are Few-shot Learners,” *Advances in Neural Information Processing Systems*, 2020. [Google Scholar] [Publisher Link]

[13] Sebastian Bubeck et al., “Sparks of Artificial General Intelligence: Early Experiments with GPT-4,” *arXiv preprint arXiv:2303.12712*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Nicholas Carlini et al., “Extracting Training Data from Large Language Models,” *USENIX Security Symposium*, 2021. [Google Scholar] [Publisher Link]

[15] John M. Polson, and Evan Shenkman, BREAKING NEWS: Casetext’s CoCounsel is Powered by OpenAI’s GPT-4, Fisher Phillips, 2023. [Publisher Link]

[16] Llias Chalkidis et al., “LEGAL-BERT: The Muppets Straight out of Law School,” *arXiv Preprint*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Zhiyu Chen et al., “FinQA: A Dataset of Numerical Reasoning Over Financial Data,” *arXiv Preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Lingjiao Chen, Matei Zaharia, and James Zou, “FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance,” *arXiv preprint arXiv:2305.05176*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Aakanksha Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways,” *JMLR*, 202. [Google Scholar] [Publisher Link]

[20] M. Chui, J. Manyika, and M. Miremadi, “Notes from the AI Frontier: Applications and Value of Deep Learning,” *McKinsey Global Institute*, 2018. [Publisher Link]

[21] Cognii, Cognii Virtual Learning Assistant. [Online]. Available: https://www.cognii.com/technology

[22] LeAh, ContractPod Ai, Automation and AI are your Partners in Strategic Excellence. [Online]. Available: https://contractpodai.com/department/legal/

[23] Jiaxi Cui et al., "ChatLaw: Open-source Legal Large Language Model with Integrated External Knowledge Bases," *CoRR*, 2023. [Google Scholar] [Publisher Link]

[24] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Mesko, "The State of Artificial Intelligence-based FDA-approved Medical Devices and Algorithms: An Online Database," *NPJ Digital Medicine*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Duolingo, "Duolingo Max" Shows the Future of AI Education, 2023. [Online]. Available: https://investors.duolingo.com/news-releases/news-release-details/duolingo-max-shows-future-ai-education

[27] Aniket Deshpande, "Regulatory Compliance and AI: Navigating the Legal and Regulatory Challenges of AI in Finance," *2024 International Conference on Knowledge Engineering and Communication Systems*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Sebastian Gehrmann et al., "Understanding and Mitigating Risks of Generative AI in Financial Services," *arXiv:2504.20086*, 2025. [CrossRef] [Publisher Link]

[29] Sara Gerke, Timo Minssen, and Glenn Cohen, "Ethical and Legal Challenges of Artificial Intelligence-driven Healthcare," *Artificial Intelligence in Healthcare*, pp. 295–336, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ali Zeb, Rafid Ullah, and Rehmat Karim, "Exploring the Role of ChatGPT in Higher Education: Opportunities, Challenges and Ethical Considerations," *International Journal of Information and Learning Technology*, vol. 41, no. 1, pp. 99-111, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] Gradescope. [Online]. Available: https://www.gradescope.com/?locale=es-es#:~:text=Gradescope%20helps%20you%20seamlessly%20administer,how%20your%20students%20are%20doing

[32] Neel Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Model," *arXiv: 2308.11462*, 2023. [CrossRef] [Publisher Link]

[33] Kate Rattray, Clio, Harvey AI for Legal Professionals: Features, Benefits and More. [Online]. Available: https://www.clio.com/blog/harvey-ai-legal/

[34] Peter Henderson et al., "Pile of Law: Learning Responsible Data Filtering from The Law and a 256GB Open-source Legal Dataset," *Advances in Neural Information Processing Systems*, 2022. [Google Scholar] [Publisher Link]

[35] Jordan Hoffmann et al., "Training Compute-optimal Large Language Models," *arXiv preprint arXiv:2203.15556*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[36] Antonio J.G. Busson et al., "Saturn Platform: Foundation Model Operations and Generative AI for Financial Services," *arXiv:2312.07721*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Alon Jacovi et al., "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[38] Ziwei Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[39] Bahar Memarian, and Tenzin Doleck, "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and Higher Education: A Systematic Review," *Computers and Education: Artificial Intelligence*, vol. 5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[40] Razvan V. Marinescu et al., "Disease Knowledge Transfer Across Neurodegenerative Diseases," *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019*, pp. 860-868, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[41] Tyler Haberle et al., "The Impact of Nuance DAX Ambient Listening AI Documentation: A Cohort Study," *Journal of the American Medical Informatics Association*, vol. 31, no. 4, pp. 975-979, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] J.P. Morgan, Quest IndexGPT: Harnessing Generative AI for Investable Indices, 2024. [Online]. Available: https://www.jpmorgan.com/insights/markets/indices/indexgpt

[43] Enkelejda Kasneci et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences*, vol. 103, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[44] Harvard Business Review, A New Survey Reveals How Legal Professionals Expect AI to Impact Their Work, 2023. [Online]. Available: https://hbr.org/sponsored/2023/12/a-new-survey-reveals-how-legal-professionals-expect-ai-to-impact-their-work

[45] Khan Academy, Teachers Share Why Our AI-Powered Teaching Assistant, Khanmigo, Is a Must-Have, 2024. [Online]. Available: https://blog.khanacademy.org/teachers-share-why-our-ai-powered-teaching-assistant-khanmigo-is-a-must-have/

[46] Shanshan Han, "Bridging Today and The Future of Humanity: AI Safety in 2024 and Beyong," *arXiv:2410.18114*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[47] Yaniv Leviathan, Matan Kalman, and Yossi Matias, "Fast Inference from Transformers Via Speculative Decoding," *Proceedings of the 40th International Conference on Machine Learning*, pp. 19274–19286, 2023. [Google Scholar] [Publisher Link]

[48] Patrick Lewis et al., “Retrieval-augmented Generation for Knowledge-intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, vol. 33, 2020. [Google Scholar] [Publisher Link]

[49] Rylan Schaeffer et al., “Correlating and Predicting Human Evaluations of Language Models from Natural Language Processing Benchmarks,” *arXiv:2502.18339*, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[50] Yunxiang Li et al., “ChatDoctor: A Medical Chat Model Fine Tuned on a Large Language Model Meta-AI (LLaMA) using Medical Domain Knowledge,” *Cureus*, vol. 15, no. 6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[51] The Alan Turing Institute, Leveraging Large Language Models in Finance: Pathways to Responsible Adoption. [Publisher Link]

[52] Hang Yuan, Saizhuo Wang, and Jian Guo, “Alpha-GPT 2.0: Human-in-the-Loop AI for Quantitative Investment,” *arXiv:2402.09746*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[53] Gagan Bhatia et al., “FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models,” *arXiv:2402.10986*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[54] Nelson F. Liu et al., “Lost in the Middle: How Language Models use Long Contexts,” *arXiv preprint arXiv:2307.03172*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[55] Pengfei Liu et al., “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[56] Renqian Luo et al., “BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining,” *Briefings in Bioinformatics*, vol. 23, no. 6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[57] LexisNexis, Lexis+ AI: Transform Your Legal Work. [Online]. Available: https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page

[58] Sewon Min et al., “FactScore: Fine-grained Atomic Evaluation of Factual Precision in Long-form Text Generation,” *arXiv preprint arXiv:2305.14251*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[59] Ethan Mollick, and Lilach Mollick, “Assigning AI: Seven Approaches for Students, with Prompts,” *arXiv Preprint*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[60] OpenAI, Morgan Stanley Uses AI Evals to Shape the Future of Financial Services. [Online]. Available: https://openai.com/index/morgan-stanley/

[61] Harsha Nori et al., “Capabilities of GPT-4 on Medical Challenge Problems,” *arXiv preprint arXiv:2303.13375*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[62] OpenAI, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[63] Matthew E. Peters et al., “Deep Contextualized Word Representations,” *arXiv Preprint*, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[64] Wayground, Quizizz AI. [Online]. Available: https://webflow-dev.quizizz.com/quizizz-ai?source=ai-solution-btn&lng=en

[65] Inioluwa Deborah Raji et al., “Closing the AI Accountability Gap: Defining an End-to-end Framework for Internal Algorithmic Auditing,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[66] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu, “Security and Privacy Challenges of Large Language Models: A Survey,” *arXiv:2402.008888*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[67] Taeyoon Kwon et al., “Large Language Models Are Clinical Reasoners: Reasoning-Aware Diagnosis Framework with Prompt-Generated Rationales,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 18417-18425, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[68] Rasha Halat, and Lina Khair Rahme, Addressing Inequities in Education: AI as a Double-Edged Sword (Part II), Middle East Professional Learning Initiative, 2024. [Publisher Link]

[69] Timo Schick et al., “Toolformer: Language Models can Teach Themselves to Use Tools,” *Advances in Neural Information Processing Systems*, 2023. [Google Scholar] [Publisher Link]

[70] Edward H. Shortliffe, and Martin J. Sepúlveda, “Clinical Decision Support in the Era of Artificial Intelligence,” *JAMA*, vol. 320, no. 21, pp. 2199–2200, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[71] Karan Singhal et al., “Towards Expert-level Medical Question Answering with Large Language Models,” *Nature Medicine*, vol. 31, pp. 943-950, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[72] Spellbook, AI in Legal Departments: 2025 Benchmarking Report. [Online]. Available: https://www.spellbook.legal/blog/counselwell

[73] Nandan Thakur et al., “BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models,” *arXiv preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[74] Arun James Thirunavukarasu et al., “Large Language Models in Medicine,” *Nature Medicine*, vol. 29, pp. 1930–1940, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[75] Thomson Reuters, A Year in Review: How AI Transformed the Legal Profession in 2023. [Online]. Available: https://legal.thomsonreuters.com/blog/how-ai-transformed-the-legal-profession-in-2023/

[76] Hugo Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[77] Scott W. Perkins et al., “Improving Clinical Documentation with Artificial Intelligence: A Systematic Review,” *Advances in Health Information Science and Practice*, vol. 21, no. 2, 2024. [Publisher Link]

[78] Richard M. Schwartzstein, “Clinical Reasoning and Artificial Intelligence: Can AI Really Think?,” *Transaction of the American Clinical and Climatological Association*, vol. 134, pp. 133-145, 2024. [Google Scholar] [Publisher Link]

[79] Brady D. Lund et al., “ChatGPT and a New Academic Reality: AI-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing,” *Journal of the Association for Information Science and Technology*, 2023. [Google Scholar] [Publisher Link]

[80] Yiqiu Shen et al., “ChatGPT and Other Large Language Models are Double-edged Swords,” *Radiology*, vol. 307, no. 2, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[81] Lei Wang et al., “A Survey on Large Language Model based Autonomous Agents,” *Frontiers of Computer Science*, vol. 18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[82] Alaa Abd-alrazaq et al., “Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions,” *Medical Education*, vol. 9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[83] Jason Wei et al., “Chain of thought Prompting Elicits Reasoning in Large Language Models,” *Advances in Neural Information Processing Systems*, vol. 35, 2022. [Google Scholar] [Publisher Link]

[84] Jason Wei et al., “Emergent Abilities of Large Language Models,” *arXiv preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[85] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt, “Jailbroken: How Does LLM Safety Training Fail?,” *Advances in Neural Information Processing Systems*, 2023. [Google Scholar] [Publisher Link]

[86] Laura Weidinger et al., “Ethical and Social Risks of Harm from Language Models,” *arXiv preprint arXiv:2112.04359*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[87] Jonathan H. Choi, Amy B. Monahan, and Daniel Schwarcz, “Lawyering in the Age of Artificial Intelligence,” *Minn. L. Rev*, 2024. [Google Scholar] [Publisher Link]

[88] Malik Sallam, “ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns,” *Healthcare*, vol. 11, no. 6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[89] Qingyun Wu et al., “AutoGen: Enabling Next-gen LLM Applications Via Multi-agent Conversation,” *arXiv preprint arXiv:2308.08155*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[90] Shijie Wu et al., “BloombergGPT: A Large Language Model for Finance,” *arXiv preprint arXiv:2303.17564*, 2023. [Google Scholar] [Publisher Link]

[91] Xiang Yue et al., “Mammoth: Building Math Generalist Models through Hybrid Instruction Tuning,” *arXiv preprint arXiv:2309.05653*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[92] Wei Luo, and Dihong Gong, “Pre-trained Large Language Models for Financial Sentiment Analysis,” *arXiv:2401.05215*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[93] Nestor Maslej et al., “Artificial Intelligence Index Report 2023,” *arXiv preprint*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[94] Michael Chui et al., “The State of AI in 2023: Generative AI’s Breakout Year,” *McKinsey Global Institute*, McKinsey & Company, 2023. [Google Scholar] [Publisher Link]

[95] Dan Hendrycks et al., “Measuring Massive Multitask Language Understanding,” *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[96] Yupeng Cao et al., “RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data,” *arXiv:2404.07452*, 2024. [Google Scholar] [Publisher Link]