CIA Open Access

Trustworthy Artificial Intelligence in High-Stakes Decision Systems: A Cross-Domain Systematic Review of Explainability, Fairness, Privacy, Robustness, and Governance

Tyler J. Brennan

Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
tbrennan@uic.edu

DOI: 10.63575/CIA.2026.40117

Abstract

Artificial intelligence (AI) systems are increasingly entrusted with consequential decisions in finance, healthcare, cybersecurity, public infrastructure, and online platforms. As these systems move from research prototypes to operational deployment, their value depends not only on predictive accuracy but on whether stakeholders can trust them to behave transparently, fairly, privately, robustly, and accountably. This paper presents a cross-domain systematic review of trustworthy AI in high-stakes decision systems, synthesizing 196 recent studies that span financial risk and regulatory technology, clinical and biomedical analytics, network and software security, social-media information integrity, energy and sustainability, and commerce and mobility. We organize the literature along five trustworthiness pillars-explainability, fairness, privacy preservation, robustness and safety, and accountability and governance-and three cross-cutting methodological enablers: multimodal and multi-source data fusion, graph-based learning, and large language models (LLMs) with agentic orchestration. For each domain we characterize the dominant tasks, data modalities, and evaluation practices; for each pillar we summarize representative techniques and their tensions. Our coverage analysis reveals an uneven landscape: explainability and robustness dominate finance and security, privacy concentrates in healthcare and consumer platforms, and fairness remains comparatively underexplored beyond credit scoring. We further discuss recurring trade-offs-privacy versus utility, fairness versus accuracy, transparency versus performance-and the emerging risks introduced by LLM-based and multi-agent systems, including hallucination, memory poisoning, and over-refusal. We conclude with open challenges and a research agenda emphasizing standardized evaluation, lifecycle governance, and the integration of trustworthiness guarantees into agentic AI.

Keywords: *Trustworthy AI; explainable AI; algorithmic fairness; privacy-preserving machine learning; robustness; large language models; agentic AI; systematic review.*

1. Introduction

Over the past few years, machine learning has graduated from a tool for offline analytics into the decision substrate of systems that allocate capital, triage patients, defend networks, and moderate public discourse. In finance, learned models now identify cross-asset liquidity contagion and recalibrate hedging positions [1], score consumer and small-business credit [27], and surface fraudulent transactions in real time [33]. In healthcare, predictive models forecast hospital resource demand during infectious-disease surges [10], stratify readmission risk [41], and screen candidate molecules and treatment plans [131]. In cybersecurity, classifiers recognize malicious traffic and login behavior [2], [17] and reason over attack paths in industrial control networks [8]. On social platforms, temporal and structural models detect coordinated malicious accounts and misinformation campaigns [14], [196]. These deployments share a defining property: the cost of an error is borne by people-borrowers denied credit, patients mistriaged, citizens misinformed-rather than by an abstract loss function.

This shift reframes what it means for an AI system to be good. Accuracy on a held-out test set is necessary but no longer sufficient. A credit model that is accurate on average may nonetheless encode discriminatory patterns against protected groups [110]; a clinical model may be confidently wrong on inputs that drift away from its training distribution [68]; a federated analytics pipeline may leak sensitive attributes through its gradients unless privacy is explicitly engineered [51]; and a security model may be brittle against adversarial manipulation, including the prompt-level attacks that now threaten large language models [128]. The umbrella term trustworthy AI has emerged to capture this broader set of requirements-explainability, fairness, privacy, robustness, and accountability-that govern whether a system deserves to be relied upon in consequential settings.

Although the trustworthy-AI literature is large and growing, it is also fragmented. Existing reviews tend to be vertical, examining a single domain or a single property: systematic treatments of explainability, fairness, and accountability in financial decision-making [105], surveys of privacy-preserving federated learning for medical AI [51], reviews of deep learning in cardiovascular imaging [71], and syntheses of sentiment-driven

market prediction [187]. Broader methodological surveys-such as comprehensive treatments of agentic AI across domains [46]-map capabilities but stop short of analyzing how trustworthiness requirements transfer, conflict, or compound when the same techniques are reused across sectors. The result is that practitioners in one field rarely benefit from hard-won lessons in another, even when they face structurally identical problems: anomaly detection over imbalanced tabular data, attribution of model outputs to inputs, or collaboration across institutions that cannot share raw records.

This paper addresses that gap with a cross-domain synthesis. We assemble a corpus of 196 recent works, deliberately spanning heterogeneous application areas, and read them through a unified lens. Our contributions are fourfold. First, we propose an organizing framework (Figure 1) that couples five trustworthiness pillars with three cross-cutting methodological enablers and a foundation of data governance and evaluation. Second, we provide a structured tour of how AI is actually used across six high-stakes domains, characterizing the tasks and data that make trustworthiness non-negotiable. Third, for each pillar we distill representative techniques, identify which domains drive their development, and surface the trade-offs that recur across sectors. Fourth, through a coverage analysis (Figures 2-4) we quantify where the literature is dense and where it is thin, and we articulate an agenda for the trustworthiness of the agentic and LLM-centric systems now entering production.

The stakes also vary in kind, not merely in degree. Some failures are immediate and visible, as when a trading model misjudges liquidity [1] or a perception system misreads a scene under poor illumination [98]; others are slow and diffuse, as when a credit model entrenches disadvantage [110] or a recommendation and information ecosystem amplifies coordinated manipulation [114] over months. Some harms fall on identifiable individuals-a misclassified patient [42], a wrongly flagged transaction [25]-while others fall on whole populations or institutions, as in systemic financial contagion [5] or the equity consequences of an energy transition [106]. A trustworthy-AI framework must speak to all of these at once, which is why no single property suffices and why the pillars below must be considered jointly rather than as a checklist to be satisfied one item at a time.

A note on positioning is warranted. We do not claim exhaustive coverage of any single subfield; specialized surveys do that better, whether for agentic AI [46], privacy-preserving federated learning in medicine [51], cardiovascular imaging [71], financial decision-making [105], or sentiment-driven market prediction [187]. Our value is orthogonal: by holding a deliberately heterogeneous corpus to a single trustworthiness lens, we expose transferable structure and systematic gaps that vertical reviews, by construction, cannot see. The breadth is the method, not an accident of sampling, and the synthesis it enables-which techniques are mature, which are domain-locked, and which have yet to cross from the field that invented them to the fields that need them-is the contribution we intend.

The remainder of the paper is organized as follows. Section 2 defines scope, taxonomy, and review methodology. Section 3 surveys the application landscape across six domains plus a cross-domain methods category. Sections 4 through 8 examine the five trustworthiness pillars in turn-explainability, fairness, privacy, robustness, and accountability. Section 9 analyzes the three cross-cutting enablers. Section 10 synthesizes the coverage analysis and the trade-offs it exposes. Section 11 lays out open challenges and future directions, and Section 12 concludes.

2. Scope, Taxonomy, and Review Methodology

We scope this review to AI and machine-learning systems that inform or automate consequential decisions, where an incorrect, opaque, unfair, privacy-violating, or manipulable output carries material risk to individuals, organizations, or the public. This deliberately excludes purely entertainment-oriented or low-stakes applications and includes both classical statistical learning and contemporary deep, graph, and foundation-model approaches.

Our taxonomy comprises five trustworthiness pillars. Explainability and interpretability concern whether a model output can be understood and attributed to inputs or mechanisms. Fairness and bias mitigation concern whether outcomes are equitable across individuals and groups. Privacy preservation concerns whether sensitive data can be used for learning without undue disclosure. Robustness, security, and safety concern whether a system maintains correct behavior under distribution shift, adversarial pressure, and operational stress. Accountability and governance concern whether decisions are auditable, compliant, and traceable to responsible actors. Cutting across all five are three methodological enablers that repeatedly determine whether trustworthiness can be achieved in practice: multimodal and multi-source data fusion, graph-based learning, and large language models with agentic orchestration. Beneath these sits a foundation of data-quality governance, evaluation and benchmarking, and the regulatory and ethical context in which systems operate.

The corpus was assembled to maximize cross-domain breadth rather than to exhaust any single venue. It comprises 196 works drawn predominantly from applied AI conferences and journals, with publication years concentrated in 2024-2026. We classify each work by a primary application domain-Finance and RegTech; Healthcare and Life Sciences; Cybersecurity and Digital Trust; Social Platforms and Information Integrity; Infrastructure, Energy, and Sustainability; and Commerce, Mobility, and Creative Applications-plus a Cross-Domain Methods and Foundations category for works whose contribution is primarily methodological [16], [46]. Each work is additionally tagged with the trustworthiness pillars it substantively engages and the enablers

it employs. Because a single study often advances several themes, pillar tags overlap; the domain assignment, by contrast, is exclusive and underlies the distribution reported in Section 10. We emphasize that this is a qualitative, interpretive synthesis: our aim is to map structure and surface tensions across a heterogeneous field, not to perform a quantitative meta-analysis of effect sizes.

Two limitations of this synthesis should be stated plainly. First, our corpus, though broad, is a sample rather than a census, and its composition shapes the coverage patterns we report; the relative sparsity of a cell in our coverage analysis reflects this corpus and should be read as a hypothesis about the field rather than a definitive measurement of it. Second, because we synthesize at the level of problem structure and trustworthiness property, we deliberately abstract away from the quantitative results of individual studies, trading numerical precision for cross-domain perspective. We regard both as acceptable costs for a review whose purpose is to reveal transferable structure and systematic gaps that narrower, deeper treatments cannot, and we are explicit about them so that readers can calibrate the claims that follow.

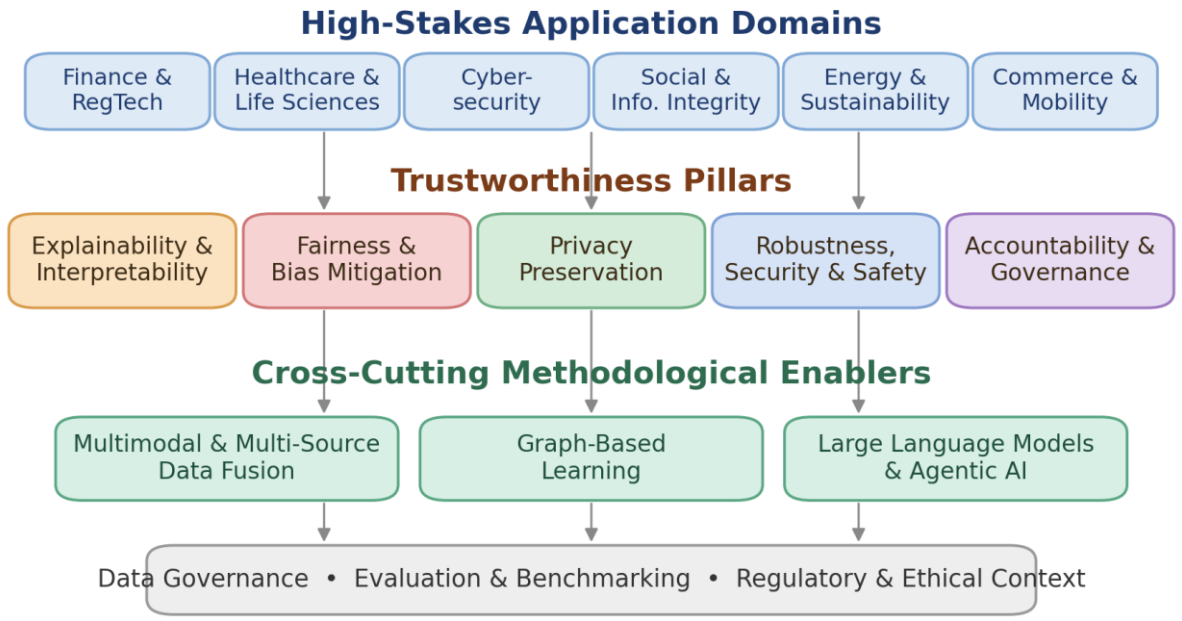

Figure 1. Organizing framework for the review. Six high-stakes application domains (top) impose trustworthiness requirements that we group into five pillars (middle); three cross-cutting methodological enablers (bottom) determine whether those requirements can be met, atop a foundation of data governance, evaluation, and regulatory context.

3. The Application Landscape of AI in High-Stakes Decision Systems

We begin with the demand side: where AI is deployed and why trustworthiness is non-negotiable there. This section surveys six domains and a methods category, characterizing dominant tasks, data modalities, and the failure modes that motivate the pillars examined later.

3.1 Financial and Regulatory Decision Systems

Finance is the most heavily represented domain in our corpus, reflecting both the maturity of quantitative methods and the severity of the consequences of error. A first cluster concerns systemic and market risk. Studies model cross-asset liquidity contagion and optimize dynamic hedging [1], identify cross-market risk-contagion amplifiers with graph attention networks [4], and map risk-contagion pathways between credit and equity markets during stress periods [5]. Related work uses graph neural networks to trace contagion paths in multi-layer financial networks [182] and applies feature-attribution analysis to market-risk stress scenarios [45]. Stress testing itself is being made lighter and more accessible: a variational-autoencoder approach with extreme-value theory generates macroeconomic scenarios for small and medium institutions [22], and ensemble anomaly detection with explainable AI provides real-time multi-risk early warning for community banks [19].

A second cluster concerns credit and default risk. Multi-source data-fusion approaches support credit-default early warning [7] and small-and-medium-enterprise credit assessment [27], while supply-chain finance is examined from a core-enterprise perspective on credit-risk transmission [6]. Green and climate-linked instruments receive dedicated treatment, with machine-learning models integrating climate factors into green-bond default prediction [103]. Consumer-credit default rates are monitored through time-series anomaly detection [65], and customer segmentation is approached via deep embedding clustering [58]. A third cluster targets fraud and financial crime: graph-based representation learning detects fraudulent and anomalous

transactions [12]; behavioral-sequence analysis yields explainable real-time fraud risk scoring [32]; time-decay-aware incremental feature extraction supports streaming fraud detection [33]; and deep-learning and ensemble methods are compared for online-payment fraud [66]. Synthetic-identity fraud and money laundering are addressed through enhanced feature engineering [25], cross-border anomalous fund-flow identification [20], and comparisons of automated versus traditional anti-money-laundering auditing [26].

A fourth cluster-regulatory technology-has grown rapidly as compliance becomes data-intensive. Natural-language methods classify the risk level of contingent-liability clauses [59], detect semantic mismatches in XBRL tag mapping for SEC filings [62], and identify disclosure discrepancies for compliance verification [63]. Cross-border contracts are analyzed for jurisdiction clauses [81] and implicit compliance violations [82], and large-scale contract review is optimized with data analytics in IPO audits [83]. Broader governance work targets compliance transparency through automated data governance and intelligent risk reporting [80] and statistical anomaly detection for field-mapping validation in payroll data migration [91]. Robotic process automation is evaluated for audit efficiency in manufacturing [190]. A fifth cluster concerns investment and trading: text mining surfaces risk signals in asset-backed-securities markets [28]; NLP quantifies ESG news sentiment for portfolio outcomes [29] and models ultra-high-net-worth client behavior for risk-aware optimization [30]; machine learning informs dynamic asset allocation for pension funds [31] and investor-asset matching in commercial real estate [159]; deep reinforcement learning optimizes order-book-imbalance high-frequency strategies [111] and cryptocurrency trend forecasting with risk management [191]. Counterparty-credit-risk work spans deep-learning margin-period-of-risk prediction with market sentiment [23] and variance-reduction frameworks for jump-diffusion credit valuation adjustment [24]. Privacy-aware collaboration appears as federated risk monitoring across institutions [53] and secure multi-party computation for transaction analytics [79], while a federated transparent optimizer reduces third-party dependencies in workflow management [77]. Trust-specific contributions include fairness-aware feature attribution for credit scoring [55], fairness-accuracy trade-off evaluation [56], post-hoc attribution on tabular financial data [181], psychological-contract risk detection [125], legal-inquiry classification [123], M&A ownership-structure extraction [112], and systematic reviews of trustworthy AI in financial decision-making [105] and sentiment-driven market prediction [187]. Agentic planning strategies are benchmarked specifically on financial question answering and numerical reasoning [184].

Beneath these clusters, finance illustrates why trustworthiness is inseparable from deployment. Market-facing models such as order-book-imbalance trading strategies [111] and cryptocurrency trend forecasting [191] operate under adversarial, non-stationary conditions where overconfidence is punished immediately, while customer-facing models such as credit scoring [110] and customer segmentation [58] make decisions that are slow to reveal their errors but costly and hard to reverse when they surface. Regulatory technology sits between the two, automating judgments-clause-risk classification [59], jurisdiction identification [81], disclosure verification [63]-that were until recently the exclusive province of trained professionals, which raises the bar for explanation and auditability rather than merely for accuracy. The breadth of the financial cluster in our corpus thus reflects not only commercial incentive but a domain in which every trustworthiness pillar has a concrete, monetizable, and often regulated failure mode.

3.2 Healthcare and Life Sciences

Healthcare is the second-largest domain in the corpus, and the one where explainability and privacy are most tightly coupled to deployment. In population and operational health, AI supports community-infection early warning from public data [3], hospital resource-demand forecasting during infectious-disease surges [10], and explainable risk stratification for hospital readmission with integrated prediction-intervention-evaluation [41]. Claims and insurance integrity is addressed through temporal feature engineering for healthcare-claims anomaly detection [21], deep-learning recognition of anomalous medical-insurance behavior [42], unsupervised detection of anomalous billing patterns for payment integrity [178], and an integrated approach to healthcare data-quality governance [139]. Risk stratification extends to medication safety, with rule-enhanced machine learning for polypharmacy-related adverse outcomes in the elderly [11].

Clinical prediction and imaging form a large cluster. Multimodal fusion supports cardiovascular-disease risk prediction [86] and early chronic-disease risk with fairness awareness [57], while multimodal attention mechanisms enable interpretable biomarker discovery [70]. Wearable HRV data feed adaptive-threshold cardiovascular-risk prediction [92]. Imaging work spans a review of deep learning in cardiovascular CT [71], anatomy-aware contrastive pre-training for label-efficient diagnosis [61], noise suppression for LED medical imaging [134], and trustworthiness evaluation of AI-assisted imaging through confidence calibration and distribution-shift detection [68]. Oncology and therapeutics are well represented: deep reinforcement learning balances efficacy and toxicity in drug combinations [131] and optimizes photodynamic-therapy dosing [133]; uncertainty-aware dose optimization is applied to intensity-modulated radiotherapy [135]; attention-enhanced LSTM predicts breast-cancer recurrence time [137]; and ovarian-stimulation and gonadotoxicity-risk prediction are studied for fertility care [136], [138]. Molecular and structural work includes Bayesian-optimization nanobody screening [130] and detection of dynamic structural changes in inflammatory protein interfaces [132]. Multi-omics drug-target prediction uses graph-attention feature selection [15], and automated eligibility screening accelerates clinical-trial recruitment via multimodal deep learning [129]. Document and language tasks include medical-document classification with multi-engine OCR and deep learning [140], comparison of pre-trained language models for medical-document routing [142], detection and protection of

personally identifiable information in clinical text [76], and definition-enhanced retrieval-augmented generation to mitigate hallucination in medical question answering [101].

Privacy-preserving collaboration is especially salient in healthcare. Federated approaches span a systematic review of privacy-preserving federated learning in medical AI [51], adaptive privacy-budget allocation for multi-institutional learning [50], privacy-preserving collaborative learning with gradient compression and dynamic budget allocation [49], and privacy-aware discovery of rare-disease patients [54]. A further cluster concerns medical communication and education, including adaptive generation of medical-education animations for health literacy [144], culturally intelligent and cross-lingual medical animation [145], [150], and procedural-animation generation for personalized training [149]. Finally, behavioral health is represented through autism-intervention research: transfer-learning evaluation of cross-context behavioral generalization [18], reinforcement-learning prompt selection and fading [165], action recognition for video therapy [166], adaptive-difficulty social-skills training [167], deep-learning classification of verbal operants [168], and adaptive learning-rate optimization for personalized interventions [169]. Dental and biomedical-materials engineering rounds out the domain, with attention-enhanced defect detection in 3D-printed prostheses [170], spectrophotometric shade classification [171], multi-objective optimization of resin-printing parameters [172], data-mining of biomechanical properties [173], and interpretable machine learning for dental-polymer formulation [174].

What unifies the healthcare cluster is that the cost of error is measured in patient outcomes, which sharpens every trustworthiness requirement. Imaging models must be not only accurate but calibrated, since an overconfident false negative can delay treatment [68], [71]; therapeutic-optimization models that titrate dose or combine drugs must respect safety constraints that have no analogue in commercial machine learning [131], [133], [135]; and any model touching patient records inherits stringent privacy obligations that drive the federated and privacy-aware designs discussed later [49], [54]. Even the supporting tasks-document routing [142], OCR-based classification [140], and patient-facing communication [144], [145]-carry trustworthiness weight, because an error in triage routing or in a translated instruction can propagate into a clinical decision. The density of healthcare work in our corpus is therefore matched by an unusually explicit concern for explanation, calibration, and privacy that other domains would do well to emulate.

3.3 Cybersecurity and Digital Trust

Cybersecurity drives much of the robustness and anomaly-detection literature. Threat detection spans ensemble learning with temporal analysis for network-threat behavior [2], machine-learning detection of anomalous login behavior in enterprise networks [17], and graph-learning behavioral detection of software-supply-chain attacks [9]. Explainability enters through attack-path reasoning over knowledge graphs for industrial control networks [8]. Vulnerability and exposure management is addressed through hybrid-analysis firmware vulnerability detection and prioritization [37], anomaly-based zero-day early warning in cloud infrastructure [38], and telemetry-driven anomaly detection for dual-purpose operational and security optimization at the edge [39]. Data protection is treated via a risk-assessment framework for data-leakage prevention [34] and an AI-enhanced federated-learning implementation for financial-network cybersecurity [36], with broader privacy-preserving analysis through federated learning [35]. Large language models appear both as tools and as attack surfaces: they support threat-intelligence analysis and incident response [40] and few-shot cyber-threat-intelligence entity and relation extraction [116], even as the security of LLMs themselves becomes a research subject through studies of jailbreak attacks and defenses [128]. Advertising and click fraud-an economically significant security problem-receives multi-dimensional behavioral analysis [154], privacy-preserving click-pattern anomaly detection [155], and feature-based bot-traffic and click-fraud detection [156].

Cybersecurity is also where the dual-use nature of AI is starkest: the same models that detect intrusions can be turned to evade them, and large language models that extract threat intelligence [40], [116] can equally be coaxed into generating attacks unless their own safety holds [128]. This reflexivity means robustness in security is never settled; defenses and attacks co-evolve, and a detector that is accurate today may be evaded tomorrow. The surveyed work accordingly emphasizes adaptive and anomaly-based detection [38], [39] over static signatures, and prioritization schemes that triage limited analyst attention toward the vulnerabilities most likely to be exploited [37]. Trustworthiness here is less about a one-time guarantee than about a system capacity to keep pace with an adaptive adversary, which is a property of process and monitoring as much as of any individual model.

3.4 Social Platforms and Information Integrity

A focused but coherent cluster targets the integrity of online information ecosystems. Early detection of malicious accounts uses temporal graph-feature learning [13] and graph-based temporal behavior analysis for coordinated accounts [14]. Misinformation is addressed through cross-modal content-consistency verification [195] and temporal-structural propagation-graph analysis for campaign detection and source attribution [196]. Coordinated inauthentic behavior in multilingual political discourse is detected by comparing URL-sharing, content-similarity, and temporal-synchronicity signals [114]. Underlying many of these tasks is the challenge of cross-cultural language understanding, exemplified by context-aware semantic-ambiguity resolution in cross-cultural dialogue [108].

Information-integrity work is small in volume but outsized in societal consequence, and it stresses a different trustworthiness profile: robustness against coordinated, adaptive manipulation and transparency about why content is flagged, more than individual privacy. Because adversaries deliberately mimic legitimate behavior, detectors must rely on subtle temporal and structural signals [13], [196] and on cross-modal consistency that is hard to fake at scale [195], while remaining explainable enough that platform decisions can be contested by those they affect. The cross-cultural and multilingual dimension [108], [114] further complicates fairness, since a detector tuned to one linguistic or cultural context may systematically misjudge another, turning a moderation tool into a source of inequity if deployed without adaptation.

3.5 Infrastructure, Energy, and Sustainability

Sustainability and infrastructure form a substantial domain where optimization and forecasting dominate. Cloud and server operations are addressed through intelligent prediction and dynamic scheduling under burst load [93], convex-optimization-based energy-efficient cloud scheduling [94], machine-learning power-consumption prediction for enterprise servers [95], and multi-source monitoring for performance-degradation prediction in Kubernetes microservices [67]. Carbon-aware computing is a recurring theme, with grid-carbon-variability-driven geo-distributed scheduling [96] and multi-objective deep-reinforcement-learning workload scheduling across data centers [97]. Energy and carbon analytics include building-energy-consumption prediction and carbon-reduction assessment in metropolitan areas [104], [109], [161], quantitative assessment of regional carbon-neutrality policy synergies [107], AI-driven quality assessment and investment-risk identification for carbon-credit projects [102], [162], and equity-focused assessment of vulnerable-population impacts in the energy transition [106], [115], [163]. Transportation and logistics contribute carbon-constrained last-mile delivery via reinforcement learning [99], analysis of route efficiency and carbon-emission correlation in retail distribution [160], and what-if scenario analysis in supply-chain digital twins balancing cost, resilience, and carbon [193]. Site-selection and policy optimization for renewable-energy enterprises uses multi-objective particle-swarm optimization [164].

The sustainability cluster differs from finance and healthcare in that its decisions are often collective and long-horizon-where to schedule computation [96], [97], how to price and vet a carbon-credit project [102], which populations bear the costs of an energy transition [106]-and its trustworthiness concerns accordingly lean toward transparency and equity rather than individual privacy. Yet the same machinery reappears: forecasting under uncertainty [95], multi-objective optimization that must balance cost against carbon and resilience [193], and equity assessment that is, in effect, a fairness analysis applied to populations rather than individuals [115]. That this domain is thinly covered on explainability and fairness in our coverage analysis, despite making increasingly consequential policy-relevant decisions, marks it as an area where trustworthiness research has yet to catch up with deployment.

3.6 Commerce, Mobility, and Creative Applications

This domain aggregates consumer-facing and perception-centric applications. Demand and marketing analytics include multi-source demand forecasting for seasonal retail products [85] and AI-driven seasonal forecasting and resource allocation for luxury-brand marketing [194], with deep reinforcement learning applied to e-commerce return route optimization [192]. Recommendation and click-through prediction are well represented: comparisons of LLM-generated semantic tags against classical text features for short-video recommendation [117], semantic-signal enhancement for CTR prediction [119], high-order feature-interaction operators for conversion-rate prediction in sparse traffic [183], ensemble learning for visitor engagement and content recommendation in virtual museums [179], spatiotemporal preference modeling for ride-hailing [158], and adaptive optimization of advertising-creative visual elements [157]. Differential-privacy techniques recur for consumer platforms, including privacy-preserving feature-attribution explanations for recommendation [72], revenue-transparency frameworks for creator platforms [73], adaptive privacy for multimedia content processing in the cloud [74], and mobile-advertising click-through-rate prediction under differential privacy [75], alongside evaluation of differential privacy and federated learning for customer-service applications [78].

Perception and computer vision span autonomous driving and freight: latency-adaptive feature-fusion weighting for V2X cooperative 3D object detection [88], reliability assessment and adaptive multi-sensor fusion under adverse weather [89], lightweight detection on compact LiDAR-camera configurations for freight [90], illumination-normalization comparison for challenging lighting [98], and depth-estimation benchmarking in unstructured environments [100]. Image-restoration and generation work includes attention-mechanism strategies for single-image super-resolution [69], a comparative evaluation of low-light enhancement from CNNs to diffusion models [176], and cross-modal artifact mining for generalizable deepfake detection [177]. Animation and games contribute predictive animation-state transitions to reduce perceptual latency in FPS games [146], deep-learning prediction of animated facial-expression communication effects [147], and GAN-based keyframe interpolation for character animation [148]. Finally, art and provenance are addressed through generative-AI artwork authentication via style-consistency analysis [151], CNN-based classification of Chinese artwork styles [152], and blockchain provenance verification against counterfeiting [153].

Across this domain a tension runs between personalization and protection. The recommendation, click-through, and creative-optimization systems that drive consumer platforms depend on fine-grained behavioral

data [157], [158], [183], yet the same data are precisely what privacy mechanisms must shield, which is why differential privacy recurs so heavily in this cluster [72], [74], [75]. Perception systems face a parallel tension between capability and reliability: richer multi-sensor fusion improves detection but adds failure surfaces that must be hardened against adverse conditions and degraded inputs [88], [89]. Even the generative and authentication work-deepfake detection [177] and blockchain provenance verification [153]-can be read as a trustworthiness response to the very generative capabilities the domain has unleashed. Commerce and mobility thus recapitulate, in a consumer setting, the same pillar tensions seen in heavily regulated domains, often with less external oversight to enforce them.

3.7 Cross-Domain Methods and Foundations

A final category collects works whose contribution is primarily methodological and reusable across domains. Foundational learning topics include filter-based feature selection for high-dimensional data [43], oversampling-ensemble interactions under varying imbalance ratios [44], efficient relational-context perception for knowledge-graph completion [16], and causal effect evaluation via propensity-score matching for welfare-program enrollment [84]. Forecasting methodology is examined through multi-factor customer-service workload prediction under holiday and promotional fluctuations [118]. A large LLM-and-agent methods cluster includes a comprehensive review of agentic AI [46], sequential cooperative multi-agent online learning and coordination [47], web-agent agentic reinforcement learning under cost and failure-risk constraints [48], and the safety of agent memory: memory-poisoning propagation and repair in multi-agent environments [60], [188] and continuous reorganization and performance preservation of agent memory structures [127], [189]. Generation and prompting methodology spans zero-shot and few-shot LLM machine translation for low-resource languages [113], the effect of prompt specificity on edge-case handling in code generation [120], comparative prompting strategies for code generation [121], discrete-diffusion versus autoregressive text generation [122], hallucination-mitigation strategies [126], few-shot example selection for in-context learning [186], retrieval-granularity effects in retrieval-augmented generation [180], over-refusal behavior on pseudo-harmful prompts [185], and prompt-generation strategies for AI agents in programming education [175]. Document-processing methods include adaptive differential privacy for federated document classification [52], adaptive OCR-engine selection for government-document digitization [143], enhanced feature fusion and transfer learning for multi-format government-document classification [141], and machine-learning methods for customer-service dialogue quality assessment [124].

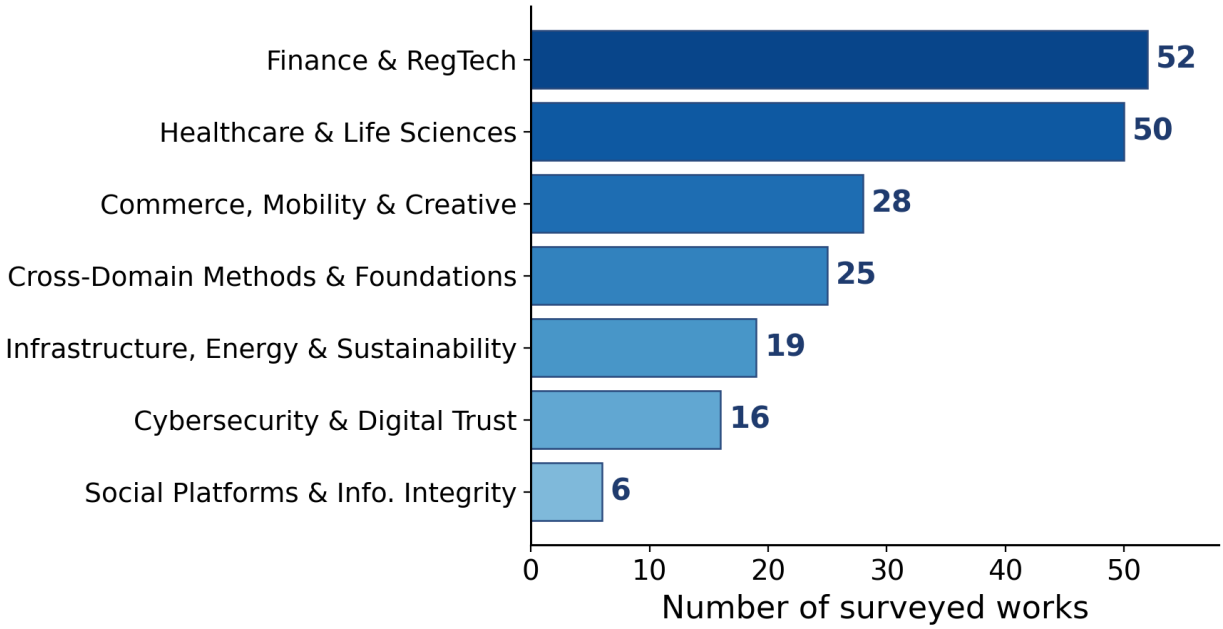

Figure 2. Distribution of the 196 surveyed works across application domains (exclusive assignment by primary domain). Finance and healthcare together account for roughly half of the corpus, while social-platform information integrity, though coherent, is the smallest cluster.

Taken together, the application landscape exhibits striking methodological convergence beneath surface-level diversity. The same handful of problem shapes recur across sectors: detecting rare events in imbalanced streams, attributing a prediction to its drivers, fusing heterogeneous evidence, ranking interventions under uncertainty, and learning from data that cannot be centralized. A fraud detector for payments [33], an anomaly monitor for medical claims [21], and an intrusion detector for enterprise logins [17] are, at the level of the learning problem, close cousins, differing more in their data and consequences than in their underlying machinery. This convergence is precisely what motivates the pillar-centric organization that follows. By abstracting away from domain specifics, we can ask which trustworthiness techniques are mature, which remain domain-locked, and which have begun to diffuse from one field to another, and we can locate the places where deployment has outpaced the study of trustworthiness.

4. Pillar I: Explainability and Interpretability

Explainability is the most frequently invoked trustworthiness property in our corpus, because high-stakes decisions almost always require justification to a regulator, clinician, or affected individual. The literature splits into post-hoc attribution, intrinsically interpretable models, and structured or symbolic reasoning.

Post-hoc attribution dominates tabular high-stakes settings. Feature-attribution methods are used to explain market-risk stress scenarios [45] and are systematically compared on tabular financial data for faithfulness, stability, and computational efficiency [181]. In credit scoring, attribution is combined with causal-path decomposition to produce fairness-aware explanations [55], illustrating how explainability and fairness reinforce one another. Real-time fraud risk scoring is made explainable through behavioral-sequence analysis [32], and community-bank early warning pairs ensemble anomaly detection with explainable AI so that flagged risks can be acted upon [19]. In medication safety, rule-enhanced machine learning yields interpretable risk stratification for polypharmacy outcomes [11], and interpretable machine learning guides dental-polymer formulation by exposing which features drive performance [174].

A second strand emphasizes intrinsic interpretability and trustworthy uncertainty. Multimodal attention mechanisms are designed not only for accuracy but for interpretable biomarker discovery [70], and trustworthiness in medical imaging is evaluated through confidence calibration and distribution-shift detection-recognizing that an explanation is only useful if the underlying confidence is well-founded [68]. Hospital-readmission management embeds explanation within an integrated prediction-intervention-evaluation loop, so that explanations connect to concrete actions [41].

A third strand turns to structured and symbolic reasoning, which offers explanations that follow the topology of a problem. Knowledge graphs enable explainable attack-path reasoning in industrial control security, tracing how an adversary could traverse a network [8]. In privacy-sensitive personalization, differentially private feature-attribution explanations show that interpretability can be delivered without exposing the underlying data [72]. The systematic review of trustworthy AI in financial decision-making confirms that explainability, alongside fairness and accountability, is treated as a first-class requirement rather than an afterthought [105]. Finally, the rise of agentic systems introduces a new form of explanation-the reasoning trace-whose faithfulness is itself under study, for example in benchmarks of planning strategies on financial reasoning tasks [184]. A recurring caution across these works is that an explanation that is plausible but unfaithful can be worse than none, because it manufactures unwarranted trust.

Methodologically, the explainability work in our corpus sorts along two axes: whether an explanation is produced after training (post-hoc) or built into the model (intrinsic), and whether it explains a single prediction (local) or the model as a whole (global). Post-hoc local attribution-assigning credit to input features for a given decision-is the most common pattern, and the comparative study of attribution methods on tabular financial data is notable for treating faithfulness, stability, and computational cost as first-class evaluation criteria rather than assuming that any attribution is adequate [181]. Intrinsic approaches instead constrain the model so that its structure is itself interpretable, as in rule-enhanced risk stratification [11] and interpretable feature-selection pipelines for materials design [174]; these trade some flexibility for transparency that requires no separate explainer and cannot drift out of agreement with the model it purports to describe.

A central but often under-appreciated question is whether an explanation faithfully reflects the model rather than merely appearing reasonable to a human. The medical-imaging trustworthiness work makes this concrete by coupling explanation with confidence calibration and distribution-shift detection, so that an explanation is qualified-or withheld-when the model is operating outside its competence [68]. The same concern motivates structured explanations that follow a problem topology, such as knowledge-graph attack paths whose individual steps can be checked against domain knowledge [8], and feature-attribution analyses of stress scenarios whose drivers can be sanity-checked against economic intuition [45]. As reasoning-capable agents proliferate, the chain-of-thought trace becomes a tempting but treacherous explanation: benchmarks of planning strategies on numerical financial reasoning show that fluent intermediate reasoning does not guarantee a correct final answer [184], underscoring that the field needs faithfulness metrics rather than plausibility heuristics. The practical upshot is that explainability should be evaluated, not merely produced.

5. Pillar II: Fairness and Bias Mitigation

Fairness is the least evenly distributed pillar in our corpus: intensively studied in credit scoring, sparse elsewhere. The canonical setting is algorithmic credit decisions, where bias detection and mitigation are examined directly [110], and where the central tension-between predictive accuracy and fairness constraints-is evaluated by comparing reweighting and resampling strategies under multiple fairness definitions [56]. Fairness-aware feature attribution, using causal-path decomposition, attempts to identify not merely whether a disparity exists but through which features it propagates [55], a step toward actionable remediation rather than mere measurement.

Beyond credit, fairness appears in clinical risk prediction, where fairness-aware multimodal fusion is applied to early chronic-disease risk to avoid systematically disadvantaging subgroups [57], and in organizational analytics, where psychological-contract risk detection in cross-cultural teams incorporates cultural adaptation

so that a model trained in one context does not unfairly penalize another [125]. The relative scarcity of fairness work outside these areas-evident in Figure 4-is itself a finding: many domains that make consequential decisions about people, including healthcare operations, content moderation, and resource allocation, have not yet subjected their models to systematic fairness auditing. We note that fairness is inseparable from the other pillars; a fair outcome that cannot be explained is hard to defend, and fairness guarantees can be undermined when privacy mechanisms perturb subgroup signals unevenly. This coupling, rather than the study of any single definition in isolation, is where the most pressing open questions lie.

Fairness research distinguishes group fairness-parity of some statistic across protected groups-from individual fairness-treating similar individuals similarly-and the credit-scoring literature in our corpus engages mainly the former, comparing reweighting and resampling under several group-fairness definitions [56]. A recurring and unavoidable result, well established in the broader literature and implicit in these comparisons, is that different fairness definitions can be mutually incompatible, so a single model cannot satisfy all of them at once; the practical question therefore shifts from which definition is correct to which is appropriate for a given decision, and who has the authority to decide. Causal approaches attempt to move past purely statistical parity by asking through which pathways a disparity arises, which is the contribution of fairness-aware attribution via causal-path decomposition [55] and connects fairness analysis directly to the explainability methods of the previous section.

Interventions can occur at three points in the pipeline: pre-processing the data, constraining the model during training, or post-processing its outputs. The reweighting and resampling strategies studied for credit operate at the first two stages [56], while the fairness-aware multimodal design for chronic-disease risk builds fairness considerations into the model architecture itself [57], and cultural adaptation in psychological-contract risk detection guards against penalizing a context simply because it differs from the training context [125]. The thinness of fairness coverage outside credit, visible in Figure 4, suggests an immediate transfer opportunity: the metric families, incompatibility results, and intervention taxonomy developed for lending are largely domain-agnostic and could be applied to clinical operations, content moderation, and welfare-program targeting [84], all of which already automate consequential decisions about people without comparable fairness scrutiny. Closing this gap is less a matter of new theory than of disciplined application of existing tools.

6. Pillar III: Privacy Preservation

Privacy preservation is the connective tissue that allows learning across institutions and individuals who cannot or will not share raw data. The corpus reveals three dominant technical families-federated learning, differential privacy, and secure computation-each with characteristic deployment patterns.

Federated learning is most developed in healthcare, where data are siloed by regulation and institution. A systematic review charts techniques, challenges, and the clinical-deployment gap [51], while concrete mechanisms address adaptive privacy-budget allocation for multi-institutional learning [50] and collaborative learning with gradient compression and dynamic budget allocation to control both communication cost and leakage [49]. Privacy-aware methods extend to sensitive populations, such as rare-disease patient discovery and outreach [54]. In finance, federated learning enables collaborative risk monitoring across institutions while balancing regulatory compliance and intelligence sharing [53], and a federated cybersecurity implementation protects financial networks [36]; a federated transparent optimizer further reduces third-party dependencies in workflow management [77]. General implementations demonstrate practical federated data analysis beyond any single sector [35].

Differential privacy concentrates in consumer-facing and content platforms, where formal guarantees can be attached to individual records. Applications include privacy-preserving feature-attribution explanations for large-scale recommendation [72], revenue-transparency frameworks on creator platforms [73], adaptive privacy for multimedia content processing in the cloud [74], mobile-advertising click-through-rate prediction [75], and click-pattern anomaly detection for advertising-fraud protection [155]. Methodologically, adaptive differential privacy with gradient clipping is studied for federated document classification [52], and differential privacy is evaluated jointly with federated learning for customer-service applications [78], reflecting the practical reality that the two families are often combined. The third family, secure multi-party computation, targets settings where even aggregated model updates are too sensitive to expose, as in privacy-preserving financial-transaction analytics [79]. Across all three families, the governing tension is privacy versus utility: stronger guarantees-tighter privacy budgets, more aggressive clipping, heavier cryptographic protocols-degrade accuracy or inflate cost. Much of the engineering effort surveyed here is, in effect, the search for favorable points on that frontier rather than the pursuit of privacy as an absolute.

Each privacy family carries a distinctive cost structure that shapes where it is adopted. Federated learning keeps raw data local and shares only model updates, but those updates can themselves leak information, which is why the surveyed methods pair federation with privacy-budget control and gradient compression [49], [50], and why a systematic review emphasizes the persistent gap between laboratory federation and clinical deployment [51]. Differential privacy adds calibrated noise to provide a formal, composable guarantee, but the privacy budget is a finite resource that degrades utility as it tightens-an explicit theme in gradient-clipping approaches to federated document classification [52] and in differentially private recommendation

explanations, where the goal is to preserve interpretability without exposing individual records [72]. Secure multi-party computation hides even intermediate computation behind cryptography, offering strong protection at substantial communication and compute cost [79].

In practice these families are combined, and the harder problem is governing the result rather than implementing any single mechanism. Differential privacy is layered atop federated learning for customer-service models [78]; transparent federated optimization is proposed precisely to make collaboration auditable and to reduce opaque third-party dependencies [77]; and cross-institutional financial risk monitoring must reconcile privacy with regulatory reporting obligations [53]. The open challenge is therefore less about inventing new primitives than about composing them into pipelines whose end-to-end privacy guarantees can be stated, verified, and explained to a regulator or an affected individual. This requirement-an auditable, communicable guarantee-connects privacy directly to the accountability pillar of Section 8, and it is where much of the practical difficulty of privacy-preserving AI now resides.

7. Pillar IV: Robustness, Security, and Reliability

Robustness asks whether a system continues to behave correctly under conditions it was not explicitly trained for: distribution shift, imbalance, adversarial manipulation, and operational stress. It is the dominant pillar in cybersecurity and a major concern in finance, healthcare operations, and the emerging study of LLM safety.

Anomaly detection is the workhorse of robustness in operational settings. It underpins network-threat recognition [2], anomalous-login detection [17], healthcare-claims and billing integrity [178], data-quality governance [139], consumer-credit-default monitoring [65], and adaptive thresholds for financial-data-quality monitoring [64]. A persistent obstacle is class imbalance: rare events-fraud, intrusion, disease-are precisely the ones that matter most, motivating careful study of oversampling-ensemble interactions under varying imbalance ratios [44] and comparisons of deep-learning versus ensemble methods for fraud [66]. Reliability under shifting conditions appears in field-mapping validation for payroll-data migration [91] and in multi-sensor fusion that must remain reliable under adverse weather for autonomous systems [89]. Wearable-derived risk prediction must likewise tolerate noisy, non-stationary signals through adaptive thresholds [92].

Security-specific robustness covers vulnerability and exposure management-firmware vulnerability detection and prioritization [37], zero-day early warning in cloud infrastructure [38], telemetry-driven anomaly detection at the edge [39]-and data-leakage prevention [34]. Operational reliability of distributed services is addressed through multi-source monitoring for microservice performance degradation [67]. A rapidly expanding frontier is the robustness of foundation models themselves. The security of large language models is studied through jailbreak attacks and defenses [128]; their behavior under adversarial framing is probed through over-refusal on pseudo-harmful prompts, where excessive caution becomes its own failure mode [185]; and their use in defense is examined through few-shot threat-intelligence extraction [116]. Trustworthiness in medical imaging explicitly incorporates distribution-shift detection so that a model can recognize when it is operating outside its competence [68]. Information-integrity robustness-resisting coordinated manipulation-appears in cross-modal misinformation verification [195] and propagation-graph campaign detection [196]. The unifying lesson is that robustness cannot be bolted on after the fact: it must be designed into data pipelines, model architectures, and monitoring loops, and it must be evaluated against the specific adversaries and shifts a system will face.

It helps to separate the threats that robustness must withstand, because each demands a different defense. Natural distribution shift occurs when deployment data drift away from training data, and is addressed by drift-aware monitoring and by explicit distribution-shift detection, as in trustworthy medical imaging [68]. Class imbalance is a structural rather than adversarial difficulty-rare events are the ones that matter-and is studied through oversampling-ensemble interactions whose behavior changes with the imbalance ratio [44]. Adversarial threats, by contrast, involve an actor deliberately manipulating inputs, which for foundation models takes the form of jailbreaks and prompt injection [128]. Conflating these categories leads to systems hardened against one threat while exposed to another, a danger crisply illustrated by the over-refusal failure mode, in which defenses tuned against malicious prompts begin rejecting benign ones and degrade usefulness in the name of safety [185].

Robustness is also a runtime property, not only a training-time one. Anomaly-detection thresholds must adapt as the underlying process evolves, lest yesterday's calibration generate tomorrow's false alarms [64], [65]; microservice monitoring must surface performance degradation with enough lead time to intervene before users are affected [67]; and data-quality and field-mapping validation must catch corruption before it propagates into downstream decisions [91], [139]. Taken together, these works argue for treating robustness as a continuous monitoring loop spanning the full lifecycle, in which detection, alerting, and retraining are first-class system components rather than afterthoughts. Several of the surveyed early-warning systems already adopt this operational stance, coupling detection with an explicit response pathway so that a flagged risk leads to action rather than to an ignored dashboard [19], [38].

8. Pillar V: Accountability, Governance, and Evaluation

Accountability concerns whether a system decisions can be audited, justified to regulators, and traced to responsible processes. It is most visible in regulatory technology and data governance. Compliance-oriented work detects disclosure discrepancies in filings for regulatory verification [63], identifies semantic mismatches in XBRL tag mapping [62], and flags implicit compliance violations in cross-border contracts [82], while large-scale contract review is systematized for IPO audits [83]. Enterprise governance is addressed through automated data governance and intelligent risk reporting that improve compliance transparency [80], statistical anomaly detection for validating field mappings in data migration [91], and integrated anomaly detection and integrity verification for healthcare data quality [139].

Underpinning accountability is rigorous evaluation, which recurs as a theme in its own right. Workload-forecasting accuracy is benchmarked under realistic holiday and promotional fluctuations [118]; OCR-engine selection for government-document digitization is evaluated adaptively [143]; multi-format government-document classification is improved and assessed through feature fusion and transfer learning [141]; and trustworthy AI in financial decision-making is reviewed precisely along explainability, fairness, and accountability axes [105]. A consistent observation across the corpus is that evaluation practice is fragmented: benchmarks, metrics, and reporting conventions vary by domain, which impedes both comparison and accountability. Where a domain has converged on shared evaluation protocols-as finance has, partly, for fraud and credit-claims about progress are easier to verify; where it has not, accountability suffers. Standardized, decision-relevant evaluation is therefore not a narrow technical concern but a precondition for governance.

Accountability rests on traceability: the ability to reconstruct why a decision was made, on what data, by which model version, and under whose authority. The regulatory-technology work in our corpus operationalizes parts of this for finance, automating the detection of disclosure discrepancies and XBRL tag mismatches so that compliance can be checked at scale [62], [63], and systematizing large-scale contract review so that obligations are not silently missed [82], [83]. Data governance provides the substrate on which such checks depend: automated governance with intelligent risk reporting [80] and integrity verification for healthcare data [139] aim to ensure that the inputs to consequential decisions are themselves trustworthy and that their lineage can be reconstructed when a decision is later questioned.

Accountability and evaluation are ultimately two faces of the same requirement. A claim that a system is fair, private, or robust is only credible if it can be measured against an agreed protocol, which is why the comparative-evaluation studies in the corpus matter beyond their immediate applications-whether benchmarking forecasting under realistic holiday and promotional fluctuations [118], evaluating OCR-engine selection for digitization [143], or assessing multi-format document classification [141], they model the disciplined measurement that accountability demands. Accountability also requires a human locus of responsibility: none of the surveyed systems eliminates the need for human oversight, and the most defensible designs make the human-AI handoff explicit, surfacing calibrated uncertainty [68] and actionable explanations [41] precisely at the points where a person must take responsibility for the decision.

9. Cross-Cutting Methodological Enablers

Three methodological families recur across domains and pillars, and largely determine whether trustworthiness can be achieved in practice. We treat them together to expose patterns invisible from within any single application.

9.1 Multimodal and Multi-Source Data Fusion

Many high-stakes decisions require integrating heterogeneous signals, and fusion quality often dominates model choice. In finance, multi-source fusion supports credit-default early warning [7], and multimodal integration of market sentiment improves counterparty-risk prediction [23]. In healthcare, fusion underlies cardiovascular-disease risk prediction [86], early cancer detection [87], interpretable biomarker discovery via multimodal attention [70], fairness-aware early chronic-disease prediction [57], wearable-HRV risk estimation [92], and multimodal eligibility screening for trials [129]. In perception, latency-adaptive feature-fusion weighting enables V2X cooperative 3D detection under bandwidth constraints [88], and reliability-aware adaptive fusion sustains multi-sensor performance in adverse weather [89]. Consumer analytics fuse weather and social-media signals for demand forecasting [85]. The cross-cutting insight is that fusion is where trustworthiness risks concentrate: missing or biased modalities propagate into unfair or unreliable outputs, and the fusion layer is a natural-but underused-place to attach explanations and uncertainty estimates.

The fusion layer also concentrates failure. When one modality is missing, noisy, or systematically biased for a subgroup, naive fusion can amplify rather than dampen the problem, which is why reliability-aware weighting that down-weights unreliable sensors under adverse conditions is valuable [89] and why bandwidth-adaptive fusion must decide what to trust when communication is constrained [88]. For trustworthiness, the implication is that the fusion mechanism deserves the same scrutiny as the predictor: explanations should be able to attribute a decision to specific modalities, and reported uncertainty should reflect modality quality rather than assume all inputs are equally reliable. The interpretable-attention and fairness-aware fusion works begin to embody this stance [57], [70], but it remains the exception rather than the norm.

9.2 Graph-Based Learning

Graphs model the relational structure that many high-stakes problems share. In finance, graph attention networks identify cross-market contagion amplifiers [4], and graph neural networks trace contagion paths in multi-layer networks [182] and detect fraudulent and anomalous transactions [12]; network-based methods map credit-equity contagion pathways [5]. In security, graph learning detects software-supply-chain attacks [9] and supports knowledge-graph attack-path reasoning [8]. In information integrity, temporal graph features detect malicious and coordinated accounts [13], [14], and propagation-graph analysis attributes misinformation campaigns [196]. In healthcare, graph-attention feature selection advances multi-omics drug-target prediction [15]. Underlying these is methodological work on efficient relational-context perception for knowledge-graph completion [16] and relational fusion for cooperative perception [88]. Graphs offer a natural substrate for explanation-paths and subgraphs are human-readable-yet their robustness to adversarial edge manipulation and their privacy implications for relational data remain comparatively underexamined.

Graphs also raise trustworthiness questions that other representations do not. Because predictions depend on relational structure, an adversary who can add or rewire even a few edges may shift outcomes-an under-studied vulnerability for the contagion and fraud detectors that increasingly rely on graph models [12], [182]. Relational data are also harder to anonymize, since the structure itself can re-identify individuals even when node features are protected, complicating the privacy guarantees discussed earlier. Against these risks graphs offer a compensating advantage: their explanations-influential paths and subgraphs-are unusually legible, as the knowledge-graph attack-path reasoning and propagation-analysis work demonstrate [8], [196]. This makes graphs a promising substrate for explanation if, and only if, their adversarial robustness and relational-privacy implications can be secured alongside their predictive use.

9.3 Large Language Models and Agentic AI

Foundation models are the fastest-moving enabler and the one that most sharply reframes trustworthiness. As capabilities, LLMs support threat-intelligence analysis [40], cyber-threat-intelligence extraction [116], low-resource machine translation [113], code generation [121], document and dialogue understanding [124], and retrieval-augmented generation whose accuracy depends on retrieval granularity [180]. Prompting methodology-prompt specificity for edge cases [120], few-shot example selection [186], and prompt generation for educational agents [175]-shapes both performance and reliability, while generation-quality work compares discrete-diffusion against autoregressive approaches [122]. A comprehensive review situates these capabilities within the broader agentic-AI landscape [46], and concrete agent systems are studied for cooperative multi-agent online learning [47] and web-agent reinforcement learning under cost and failure-risk constraints [48].

Crucially, LLMs introduce trustworthiness failure modes with no classical analogue. Hallucination must be mitigated, both generally [126] and in domain-critical settings such as medical question answering via definition-enhanced retrieval [101]. Safety spans jailbreak attacks and defenses [128] and the opposite failure of over-refusal on benign-but-suspicious prompts [185]. Agentic systems add stateful risks: memory poisoning can propagate through multi-agent collaboration and must be detected and repaired [60], [188], and agent memory must be continuously reorganized to preserve performance under change [127], [189]. Sentiment-driven and reasoning-heavy financial applications inherit these risks, as surveyed for market prediction [187] and benchmarked for numerical reasoning under different planning strategies [184]. The central message of this subsection is that the trustworthiness pillars do not merely carry over to LLM-based and agentic systems-they acquire new, often harder, forms, and the field evaluation machinery has not yet caught up.

The deployment pattern for large language models is increasingly retrieval-augmented and tool-using rather than purely parametric, which redistributes where trust must be established. Retrieval-augmented generation shifts part of the burden onto the retrieval step, so that answer accuracy depends on retrieval granularity and source quality [180]; tool-using agents shift it onto the correctness and safety of tool calls, where cost and failure risk must be reasoned about explicitly [48]; and multi-agent systems shift it onto inter-agent communication and shared memory, where poisoning can propagate silently across collaborators [60], [188]. This decomposition is an opportunity as much as a risk: each interface-a retrieval query, a tool invocation, a memory write-is a natural checkpoint at which provenance can be recorded and policies enforced. The path to trustworthy agents may therefore run through instrumenting these interfaces and continually reorganizing agent memory to preserve integrity under change [127], [189], rather than through the opaque parameters of the underlying model alone.

The deployment pattern for LLMs is increasingly retrieval-augmented and tool-using rather than purely parametric, which redistributes where trust must be established. Retrieval-augmented generation shifts part of the burden onto the retrieval step, so that answer accuracy depends on retrieval granularity and source quality [180]; tool-using agents shift it onto the correctness and safety of tool calls, which must be evaluated under cost and failure-risk constraints [48]; and multi-agent systems shift it onto inter-agent communication and shared memory, where poisoning can propagate silently across collaborators [60], [188]. This decomposition is an opportunity as much as a risk. Each interface-a retrieval query, a tool invocation, a memory write-is a natural checkpoint at which provenance can be recorded, policies enforced, and anomalies detected, which suggests that the path to trustworthy agents runs through instrumenting these interfaces rather than through interrogating the opaque parameters of the underlying model. The continual-reorganization work on agent

memory points in the same direction, treating memory integrity as a maintained property rather than a static asset [127], [189].

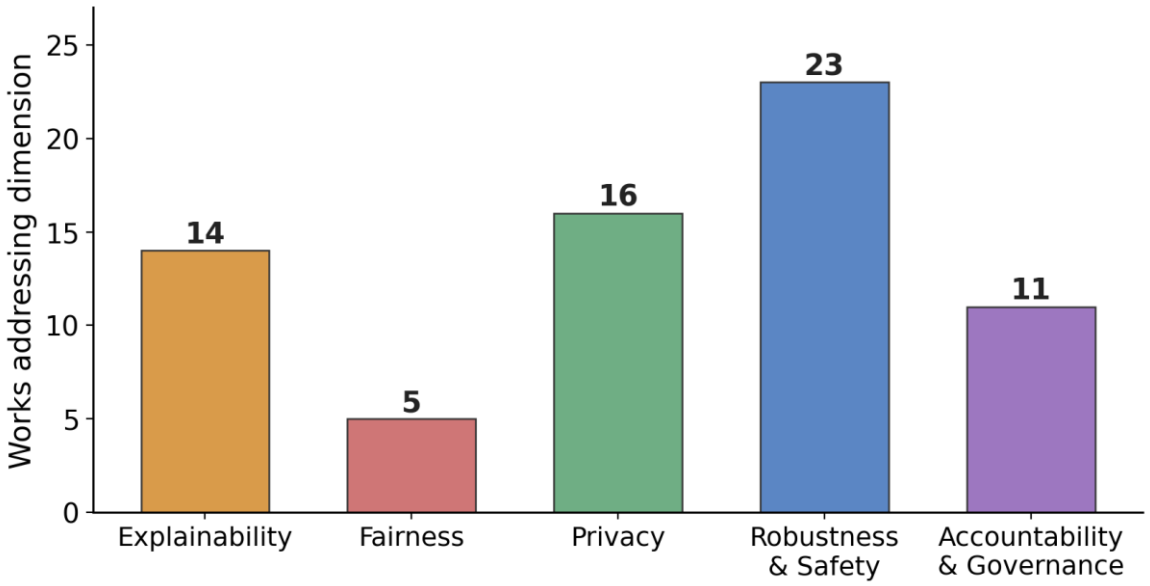

Figure 3. Thematic coverage of the corpus across the five trustworthiness pillars. Because many works engage multiple pillars, counts overlap and do not sum to the corpus size. Robustness and privacy are most heavily addressed; fairness is the least studied.

9.4 Foundational Learning Techniques

Underlying the three headline enablers is a layer of foundational techniques whose quiet importance the corpus makes clear. Class imbalance, ubiquitous in fraud, intrusion, and disease detection, is studied directly through the interaction of oversampling and ensembling across imbalance regimes [44], a reminder that the choice of resampling strategy can matter as much as the choice of model. Feature selection for high-dimensional data improves both performance and interpretability by discarding noise that would otherwise obscure attribution [43], and efficient relational-context perception sharpens the knowledge-graph completion on which many downstream reasoning systems depend [16]. Methodologically careful evaluation also belongs here: causal effect estimation via propensity-score matching shows how to ask what a decision actually caused rather than what it merely correlated with [84], and accuracy benchmarking under realistic operational fluctuations guards against optimistic offline results that quietly fail in production [118].

These foundations are easy to overlook precisely because they are not novel, yet they determine whether the trustworthiness pillars can be realized at all. An attribution method is only as faithful as the features it operates over; a fairness audit is only as meaningful as the imbalance handling that precedes it; and a privacy guarantee is only as useful as the utility that survives it. The most robust contributions in our corpus treat these foundational choices as part of the trustworthiness story rather than as interchangeable preprocessing, and we expect the same discipline to be necessary, and harder to maintain, as the field moves toward agentic systems whose end-to-end behavior is considerably more difficult to evaluate than that of a single classifier.

10. Synthesis: Coverage Analysis and Recurring Tensions

Reading the corpus as a whole exposes structure that individual studies cannot. Figure 2 shows that finance and healthcare together account for roughly half of the 196 works, with finance the single largest domain; commerce, mobility, and creative applications and the cross-domain methods category follow, while social-platform information integrity, though internally coherent, is the smallest cluster. Figure 3 shows that robustness and privacy are the most heavily engaged pillars, explainability is strong, accountability is moderate, and fairness is conspicuously underrepresented. Figure 4 combines these views into a domain-by-pillar coverage matrix and reveals an uneven map: finance is strong across nearly every pillar; healthcare is strong on explainability and privacy but lighter on robustness governance; cybersecurity is intense on robustness yet effectively silent on fairness; and energy, sustainability, and social-integrity domains are thinly covered on most pillars beyond robustness.

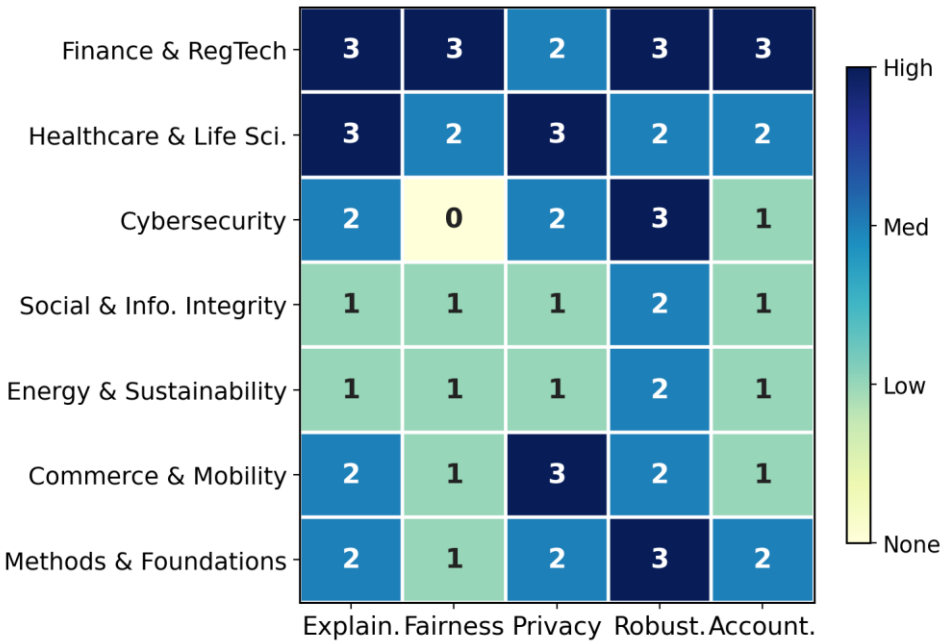

Figure 4. Coverage-intensity matrix (qualitative, 0 = none to 3 = high) of trustworthiness pillars across domains. Darker cells indicate denser treatment in the surveyed literature. The fairness column and several sustainability and social-integrity rows are notably sparse.

Three trade-offs recur across the matrix. The privacy-utility trade-off is most explicit in the federated and differential-privacy literature, where tighter budgets and heavier protocols cost accuracy or latency [50], [52], [74], [79]. The fairness-accuracy trade-off is studied head-on in credit scoring through reweighting and resampling under competing fairness constraints [56] and through fairness-aware attribution that localizes disparities [55], [110]. The transparency-performance trade-off surfaces wherever the most accurate models are the least interpretable, pushing practitioners toward post-hoc attribution [45], [181] or intrinsically interpretable designs [11], [70], [174] and toward calibrated uncertainty as a partial substitute for full transparency [68]. These tensions are not independent: privacy mechanisms can perturb subgroup signals and thereby affect fairness, while explanations that ignore uncertainty can mislead. The most valuable contributions in the corpus are those that treat trustworthiness as a multi-objective design problem rather than optimizing one property in isolation.

It is worth stressing that the pillars interact rather than sum. A privacy mechanism that perturbs subgroup signals can quietly worsen fairness [52], [55]; an explanation that ignores calibrated uncertainty can manufacture misplaced confidence [68]; and a robustness defense tuned too aggressively can degrade utility and fairness at the same time, as the over-refusal phenomenon shows [185]. Treating any pillar in isolation therefore risks improving one trustworthiness metric while silently degrading another. The coverage matrix should accordingly be read not as five independent columns but as a coupled system whose tensions must be managed jointly, and the scarcity of work that does exactly this is, in our view, the single most consequential gap the synthesis reveals.

The coverage gaps are themselves a research agenda. Fairness auditing is largely absent outside credit, even in domains-clinical operations, content moderation, resource allocation-that make consequential decisions about people. Accountability and standardized evaluation are unevenly developed: domains with shared protocols can verify progress, while others cannot [105], [118], [143]. And the trustworthiness of agentic and LLM-based systems-hallucination, memory poisoning, over-refusal, faithful reasoning traces-remains early-stage relative to the speed of deployment [60], [127], [128], [185].

A second-order observation concerns method maturity versus deployment intensity. Some pairings-explainability in finance, robustness in security-show both heavy method development and heavy deployment, and here the literature is reassuringly self-correcting, with comparison studies and reviews keeping claims honest. Other pairings show heavy deployment paired with thin trustworthiness treatment: perception systems for mobility are widely fielded yet their adversarial robustness and fairness are only lightly examined in our corpus [89], [98], [100], and sustainability decision systems increasingly shape policy and investment while receiving comparatively little explainability or fairness scrutiny [102], [107]. These mismatches, rather than the already well-covered cells of Figure 4, are where the risk of unaccountable failure is highest. A risk-proportionate research strategy would therefore steer attention toward the high-deployment, low-trustworthiness corners of the map rather than continuing to deepen the corners that are already well understood.

11. Open Challenges and Future Directions

Standardized, decision-relevant evaluation. The most consistent obstacle to accountability is the absence of shared benchmarks and metrics that reflect real decision costs. Future work should develop evaluation suites that measure not only accuracy but calibrated uncertainty, subgroup fairness, privacy leakage, and robustness to realistic shifts simultaneously, extending the comparative discipline already visible in workload forecasting [118], OCR-engine selection [143], and fraud-method comparison [66] to whole trustworthy-AI pipelines.

Integrated multi-objective trustworthiness. Because privacy, fairness, transparency, and robustness interact, they should be co-designed. Promising directions include fairness-aware attribution that respects privacy constraints [55], [72], multimodal fusion layers that carry uncertainty and explanation jointly [70], [129], and stress testing that probes several properties at once [22]. Treating trustworthiness as a Pareto frontier-rather than a checklist-would let practitioners reason explicitly about acceptable trade-offs for a given decision.

Trustworthy agentic and foundation-model systems. As LLMs and agents enter production, classical pillars must be re-derived for stateful, tool-using systems. Priorities include faithful and verifiable reasoning traces [184], defenses against jailbreak and prompt-level manipulation balanced against over-refusal [128], [185], hallucination control in safety-critical domains [101], [126], and the integrity of agent memory under poisoning and continual change [60], [127], [188], [189]. The cost- and failure-risk-aware framing of web agents [48] and the cooperative-coordination view of multi-agent learning [47] offer useful starting points for embedding trustworthiness constraints into agent objectives.

Privacy-preserving collaboration at scale. Cross-institutional learning is essential where data cannot be centralized, but current methods trade heavily against utility and are hard to govern. Advancing adaptive privacy-budget allocation [50], gradient-compressed collaborative learning [49], and secure computation for sensitive analytics [79], together with transparent federated optimization [77] and clearer accountability for federated pipelines, would close part of the clinical- and financial-deployment gap [51], [53].

Robustness for relational and perceptual systems. Graph-based and fusion-based systems are widely used yet under-hardened. Future work should examine adversarial robustness of graph learners used for contagion and fraud [12], [182], reliability of multi-sensor fusion under degradation [89], and the privacy of relational data-areas where deployment has outpaced the study of failure modes. Equally, the sustainability and information-integrity domains, thinly covered in Figure 4, warrant systematic trustworthiness analysis as their decisions grow more consequential [96], [107], [196].

Bridging domains. Finally, the central motivation of this review-that structurally identical problems recur across sectors-implies an opportunity: techniques matured in one domain can transfer to another with appropriate adaptation. Fraud-style anomaly detection informs claims integrity [21], [178]; credit-fairness methods inform clinical fairness [56], [57]; security-grade robustness informs financial monitoring [64], [128]. Deliberately cross-domain research, supported by shared evaluation and governance practices, is likely to yield faster progress than continued vertical specialization.

Human oversight and calibrated autonomy. As systems become more capable and more autonomous, the question is not whether to keep humans in the loop but where to place them for maximum effect. Calibrated uncertainty and faithful explanations should be concentrated at the decision points of highest consequence, so that scarce human attention is spent where it changes outcomes rather than spread thinly across routine cases [41], [68]. For agentic systems, this principle translates into explicit approval gates around irreversible or high-cost actions, an idea already implicit in cost- and failure-risk-aware agent formulations [48] and in the cooperative-coordination view of multi-agent learning [47]. Designing these gates well-neither so frequent that they negate the benefits of automation nor so rare that they fail to catch consequential errors-is an open human-factors problem as much as a technical one.

Shared, living benchmarks and lifecycle governance. Finally, the field would benefit from shared benchmarks that are maintained over time and that bundle the trustworthiness properties together, so that a reported gain in accuracy cannot quietly come at the expense of fairness, privacy, or robustness. Pairing such benchmarks with lifecycle governance-versioned data and models, recorded provenance, and clear lines of accountability [62], [80], [139]-would turn trustworthy AI from a collection of point techniques into an engineering discipline with reproducible guarantees. The deliberate cross-domain transfer we have emphasized throughout depends on exactly this shared infrastructure: without common evaluation protocols and governance practices, hard-won lessons cannot move from the fields that learn them to the fields that need them, and each sector is condemned to relearn the same mistakes. Building that infrastructure is, in our view, the single highest-leverage investment the community can make.

12. Conclusion

Trustworthy AI has become the decisive factor in whether learned systems can be responsibly deployed in high-stakes settings. Synthesizing 196 recent works across finance, healthcare, cybersecurity, social platforms, infrastructure and sustainability, and commerce and mobility, we organized the field along five pillars-explainability, fairness, privacy, robustness, and accountability-and three cross-cutting enablers: multimodal

fusion, graph learning, and LLM-based and agentic systems. Our coverage analysis shows a mature but uneven landscape: robustness and privacy are well developed, explainability is strong, accountability is emerging, and fairness lags outside credit scoring. The recurring trade-offs among privacy, fairness, transparency, and performance indicate that trustworthiness is fundamentally a multi-objective design problem, not a property to be optimized in isolation. The fastest-moving frontier-foundation models and agents-reproduces every classical trustworthiness challenge in harder, stateful forms while outpacing the field evaluation machinery. We argue that the most valuable next steps are standardized decision-relevant evaluation, integrated multi-objective design, and the deliberate transfer of trustworthiness techniques across domains, so that the lessons learned in one consequential setting can protect people in all of them.

References

[1] Han, J., & Jia, R. (2026). AI-Enhanced Cross-Asset Liquidity Contagion Pathway Identification and Dynamic Hedging Strategy Optimization: Evidence from US Equity, Bond, and Derivatives Markets. Journal of Computing Innovations and Applications, 4(1), 89-96.

[2] Ren, W., Wu, X., & Li, J. (2025). AI-Driven Network Threat Behavior Pattern Recognition and Classification: An Ensemble Learning Approach with Temporal Analysis. Journal of Advanced Computing Systems, 5(9), 1-13.

[3] Wang, Y. (2025, December). Practical AI Approaches for Community Infection Early Warning: From Public Data to Actionable Insights. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 1545-1552).

[4] Li, Y., Zhao, F., & Hu, J. (2026). Identifying Cross-Market Risk Contagion Amplifiers via Graph Attention Networks: Empirical Evidence from US Financial Stress Periods. Journal of Computing Innovations and Applications, 4(1), 164-175.

[5] Han, J. (2026). Network-Based Identification of Risk Contagion Pathways Between US Credit and Equity Markets During Stress Periods. Journal of Advanced Computing Systems, 6(2), 50-63.

[6] Wei, C., & Wu, C. (2024). Credit Risk Transmission Mechanism and Prevention Strategies in Supply Chain Finance: A Core Enterprise Perspective. Artificial Intelligence and Machine Learning Review, 5(2), 101-115.

[7] Han, J., & Cao, G. (2024). A Comparative Study of Multi-source Data Fusion Approaches for Credit Default Early Warning. Artificial Intelligence and Machine Learning Review, 5(1), 105-116.

[8] Chen, Y. (2024). Explainable Attack Path Reasoning for Industrial Control Network Security Based on Knowledge Graphs. Journal of Computing Innovations and Applications, 2(1), 128-139.

[9] Hu, J., & Long, X. (2024). Graph Learning-Based Behavioral Detection for Software Supply Chain Attacks. Journal of Advanced Computing Systems, 4(4), 49-60.

[10] Wang, Y. (2026). Accuracy Evaluation of Machine Learning-Based Hospital Resource Demand Forecasting During Infectious Disease Surges: A Comparative Analysis. Journal of Science, Innovation & Social Impact, 2(1), 314-327.

[11] Wang, Y. (2026). Explainable Risk Stratification for Polypharmacy-Related Adverse Outcomes in Community-Dwelling Elderly: A Rule-Enhanced Machine Learning Approach. Journal of Sustainability, Policy, and Practice, 2(2), 18-31.

[12] Wei, C., Ge, L., & Brooks, N. (2024). Graph-based Representation Learning for Financial Fraud and Anomaly Transaction Detection. Journal of Computing Innovations and Applications, 2(1), 153-164.

[13] Deng, M. (2025, September). Early Detection of Malicious Accounts on Social Platforms Based on Temporal Graph Feature Learning. In Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence (pp. 1320-1328).

[14] Deng, M. (2025). Graph-Based Temporal Behavior Analysis for Early Detection of Coordinated Malicious Accounts in Social Media Platforms. Journal of Science, Innovation & Social Impact, 1(2), 96-106.

[15] Cheng, Z. (2025). Graph Attention-Based Feature Selection for Multi-Omics Drug Target Prediction in Cardiovascular Diseases. Journal of Science, Innovation & Social Impact, 1(1), 294-306.

[16] Tu, W., Wan, G., Shang, Z., & Du, B. (2025). Efficient relational context perception for knowledge graph completion. Applied Intelligence, 55(15), 1005.

[17] Zhang, J. (2024). Performance Evaluation and Comparison of Machine Learning Algorithms for Anomalous Login Behavior Detection in Enterprise Networks. Artificial Intelligence and Machine Learning Review, 5(2), 77-90.

[18] Bai, Y., & Liu, M. (2026). A Comparative Evaluation of Transfer Learning Methods for Cross-Context Behavioral Generalization Assessment in Autism Spectrum Disorder Interventions. Journal of Computing Innovations and Applications, 4(1), 176-185.

[19] Li, Y., & Ling, Z. (2026). Real-Time Multi-Risk Early Warning for Community Banks: An Application of Ensemble Anomaly Detection and Explainable Artificial Intelligence. Journal of Advanced Computing Systems, 6(2), 15-27.

[20] Zhong, M. (2026). Multi-Dimensional Feature Analysis and Evaluation Methods for Anomalous Fund Flow Identification in Cross-Border Financial Transactions. Journal of Science, Innovation & Social Impact, 2(2), 1-13.

[21] Han, M., & Lai, J. (2026). Temporal Feature Engineering and Threshold Optimization for Early Warning in Healthcare Claims Anomaly Detection. Journal of Advanced Computing Systems, 6(4), 27-49.

[22] Li, Y., & Long, L. (2026). Lightweight AI-Driven Stress Testing for Small and Medium Financial Institutions: A Variational Autoencoder Approach with Extreme Value Theory for Macroeconomic Scenario Generation. Artificial Intelligence and Machine Learning Review, 7(1), 108-119.

[23] Huang, Y. (2025, August). Deep learning-enhanced dynamic margin period of risk prediction for counterparty credit risk management: A multi-modal approach integrating market sentiment analysis and real-time exposure assessment. In Proceedings of the 2nd International Conference on Intelligent Computing and Data Analysis (pp. 328-335).

[24] Huang, Y. (2024). Adaptive Importance Sampling for Jump-Diffusion CVA A Variance-Reduction Framework. Academia Nexus Journal, 3(3).

[25] Huang, Y. (2025). Enhanced Feature Engineering and Algorithm Optimization for Real-Time Detection of Synthetic Identity Fraud and Money Laundering in Financial Transactions. Journal of Science, Innovation & Social Impact, 1(1), 384-397.

[26] Ge, L. (2025). Efficiency Comparison of Automated Tools versus Traditional Methods in Anti-Money Laundering Compliance Auditing for Banking Institutions. Journal of Science, Innovation & Social Impact, 1(1), 265-277.

[27] Shi, X. (2025, August). Intelligent Credit Risk Assessment for Small and Medium Enterprises Based on Multi-dimensional Data Fusion. In Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business (pp. 186-196).

[28] Han, J. (2025, October). Multi-source Text Mining for Risk Signal Detection in Asset-Backed Securities Market: An NLP-driven Data Analytics Approach. In Proceedings of the 2025 International Symposium on Machine Learning and Social Computing (pp. 497-506).

[29] Cai, Y. (2025). NLP-Quantified ESG News Sentiment and Portfolio Outcomes Evidence from Real-Time Signals. Annals of Applied Sciences, 6(1).

[30] Cai, Y. (2025, June). NLP-Enhanced Predictive Analytics for UHNW Client Investment Behavior: A Risk-Aware Portfolio Optimization Approach in Volatile Markets. In Proceedings of the 2025 2nd International Conference on Digital Economy, Blockchain and Artificial Intelligence (pp. 185-191).

[31] Crowford, A., Cai, Y., & Langford, V. (2024). Machine Learning-Enhanced Dynamic Asset Allocation in Target-Date Investment Strategies for Pension Funds. Journal of Computing Innovations and Applications, 2(2), 122-135.

[32] Deng, M. (2025). Real-Time Fraud Risk Scoring through Behavioral Sequence Analysis: An Explainable Approach for online Transaction Security. Journal of Sustainability, Policy, and Practice, 1(4), 130-142.

[33] Zhong, M. (2024). Time-Decay Aware Incremental Feature Extraction for Real-Time Transaction Fraud Detection. Artificial Intelligence and Machine Learning Review, 5(3), 136-145.

[34] Wu, X., Li, J., & Ren, W. (2024). Risk Assessment Framework for Data Leakage Prevention Using Machine Learning Techniques. Artificial Intelligence and Machine Learning Review, 5(3), 55-66.

[35] Ren, W., Li, J., & Wu, X. (2024). Privacy-Preserving Data Analysis Using Federated Learning: A Practical Implementation Study. Artificial Intelligence and Machine Learning Review, 5(1), 40-50.

[36] Han, J. (2025). AI-Enhanced Cybersecurity for Financial Networks: A Federated Learning Implementation. Journal of Science, Innovation & Social Impact, 1(1), 241-252.

[37] Long, X. (2025). Research on Intelligent Firmware Vulnerability Detection and Priority Assessment Method Based on Hybrid Analysis. Journal of Science, Innovation & Social Impact, 1(1), 350-361.

[38] Long, X. (2026). Performance Evaluation of Anomaly-Based Detection Approaches for Zero-Day Attack Early Warning in Cloud Infrastructure. Journal of Science, Innovation & Social Impact, 2(1), 352-363.

[39] Long, X., Hu, J., & Ling, Z. (2026). A Comparative Analysis of Telemetry-Driven Anomaly Detection Approaches for Dual-Purpose Operational and Security Optimization in Edge Computing Infrastructure. Journal of Computing Innovations and Applications, 4(1), 79-88.

[40] Jia, R., Zhang, J., & Prescot, J. (2024). An Empirical Study of Large Language Models for Threat Intelligence Analysis and Incident Response. Journal of Computing Innovations and Applications, 2(1), 99-110.

[41] Liu, Y. (2025). Explainable Risk Stratification and Resource Coordination for Hospital Readmission Management through Integrated Prediction-Intervention-Evaluation Framework. Journal of Science, Innovation & Social Impact, 1(2), 107-118.

[42] Han, M. (2025). Intelligent Recognition of Anomalous Behaviors in Medical Insurance Through Deep Learning. Journal of Science, Innovation & Social Impact, 1(1), 410-426.

[43] Min, S., & Wei, C. (2023). Comparative Analysis of Filter-based Feature Selection Methods for High-Dimensional Data in Classification Tasks. Journal of Advanced Computing Systems, 3(8), 25-38.

[44] Wei, W., & Shang, Z. (2026). An Empirical Evaluation of Oversampling-Ensemble Interactions Under Varying Imbalance Ratios for Tabular Data Classification. Artificial Intelligence and Machine Learning Review, 7(2), 70-81.

[45] Li, Z., Huang, Y., & Montgomery, I. (2024). Feature Attribution-Based Explainability Analysis for Market Risk Stress Scenarios. Journal of Computing Innovations and Applications, 2(2), 136-150.

[46] Zhang, S., Jia, R., & Li, Z. (2024). Agentic AI Across Domains: A Comprehensive Review of Capabilities, Applications, and Future Directions. Journal of Computing Innovations and Applications, 2(1), 86-98.

[47] Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2025, December). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments. In 2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC) (pp. 692-697). IEEE.

[48] Ma, Q., Yue, L., Xu, S., Shi, Y., & Liu, H. (2026, January). Web Agent Agentic Reinforcement Learning Decision Model Under Multi-Cost and Failure Risk Constraints. In Proceedings of the 2026 5th International Conference on Big Data, Information and Computer Network (pp. 514-520).

[49] Han, M. (2025, December). Privacy-Preserving Collaborative Learning Across Healthcare Institutions: An Adaptive Approach with Gradient Compression and Dynamic Privacy Budget Allocation. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 679-684).

[50] Shi, X. (2024). Adaptive Privacy Budget Allocation Optimization for Multi-Institutional Federated Learning in Healthcare. Journal of Advanced Computing Systems, 4(2), 50-61.

[51] Wei, C., & Guan, H. (2024). Privacy-Preserving Federated Learning in Medical AI: A Systematic Review of Techniques, Challenges, and the Clinical Deployment Gap. Artificial Intelligence and Machine Learning Review, 5(3), 124-135.

[52] Zhang, Q. (2025, December). Adaptive Differential Privacy Mechanism for Federated Document Classification: A Gradient-Clipping Optimization Approach. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 672-678).

[53] Zhong, M. (2026). Privacy-Preserving Federated Learning for Collaborative Risk Monitoring Across Financial Institutions: Balancing Regulatory Compliance and Intelligence Sharing. Journal of Sustainability, Policy, and Practice, 2(2), 44-54.

[54] Pan, Z. (2024). Privacy-Aware AI for Rare-Disease Patient Discovery and Targeted Outreach: An Effectiveness Study. Spectrum of Research, 4(1).

[55] Zhong, M. (2025). Fairness-Aware Feature Attribution for Credit Scoring: A Causal Path Decomposition Approach. Journal of Science, Innovation & Social Impact, 1(1), 442-451.

[56] Wang, Z., & Lai, J. (2026). Fairness-Accuracy Trade-offs in AI Credit Scoring: A Comparative Evaluation of Reweighting and Resampling Strategies Under Multiple Fairness Constraints. Journal of Computing Innovations and Applications, 4(1), 117-126.

[57] Shi, X. (2026). Fairness-Aware Multimodal Fusion for Early Chronic Disease Risk Prediction: A Temporal Deep Learning Approach. Journal of Science, Innovation & Social Impact, 2(1), 217-231.

[58] Weng, H. (2025). Deep Embedding Clustering with Adaptive Feature Selection for Banking Customer Segmentation. Spectrum of Research, 5(2).

[59] Liang, D. (2026). Risk Level Classification of Contingent Liability Clauses in Financial Statement Notes Using NLP Techniques. Artificial Intelligence and Machine Learning Review, 7(1), 53-68.

[60] Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.

[61] Han, M. (2026). Anatomy-Aware Contrastive Pre-training: Leveraging Spatial Consistency for Label-Efficient Medical Image Diagnosis Across Multi-Modal Imaging. Journal of Sustainability, Policy, and Practice, 2(1), 55-70.

[62] Liang, D., Chen, Z., & Wei, C. (2026). Detecting Semantic Mismatches in XBRL Tag Mapping for SEC 10-K Filings: A Text Comparison and Historical Consistency Analysis. Journal of Computing Innovations and Applications, 4(1), 154-163.

[63] Liang, D. (2026). Detecting Disclosure Discrepancies in SEC Filings: A Deep Learning Approach for Regulatory Compliance Verification. Journal of Sustainability, Policy, and Practice, 2(1), 101-114.

[64] Zhong, M. (2025, September). Adaptive Anomaly Detection Threshold for Financial Data Quality Monitoring Based on Time Series Features. In Proceedings of the 2025 International Symposium on Artificial Intelligence and Computational Social Sciences (pp. 578-587).

[65] Zhong, M. (2026). Optimization of Anomaly Detection Algorithms for Consumer Credit Default Rates Based on Time-Series Feature Extraction. Journal of Sustainability, Policy, and Practice, 2(1), 44-54.

[66] Zhang, J. (2026). A Comparative Evaluation of Deep Learning and Ensemble Algorithms for Online Payment Fraud Detection. Journal of Science, Innovation & Social Impact, 2(1), 164-177.

[67] Cao, H., & Long, L. (2026). Empirical Evaluation of Multi-Source Monitoring Signal Effectiveness and Lead Time for Performance Degradation Prediction in Kubernetes-Based Microservices. Journal of Advanced Computing Systems, 6(4), 15-26.

[68] Chen, Y., & Lai, J. (2026). Multi-Metric Trustworthiness Evaluation of AI-Assisted Medical Imaging Diagnosis: Integrating Confidence Calibration and Distribution Shift Detection. Journal of Global Engineering Review, 4(1), 113-126.

[69] Wang, X., Liu, M., & Long, L. (2026). Effectiveness Evaluation of Attention Mechanism Strategies in Deep Learning-Based Single Image Super-Resolution. Journal of Global Engineering Review, 4(1), 89-98.

[70] Zhang, F., Ye, H., & Wei, C. (2024). Leveraging Multi-Modal Attention Mechanisms for Interpretable Biomarker Discovery and Early Disease Prediction. Journal of Computing Innovations and Applications, 2(2), 111-121.

[71] Zhang, F., Cheng, Z., & Holloway, V. (2024). Deep Learning in Cardiovascular CT Imaging: Evolution, Trends, and Clinical Translation from 2020 to 2025. Journal of Computing Innovations and Applications, 2(2), 88-99.

[72] Li, X. (2025). Privacy-Preserving Feature Attribution Explanations for Large-Scale Recommendation Systems: A Differential Privacy Approach. Journal of Science, Innovation & Social Impact, 1(1), 19-32.

[73] Zhang, J. (2025). Privacy-Preserving Revenue Transparency on Creator Platforms An e-Differential-Privacy Framework. Spectrum of Research, 5(2).

[74] Lei, Y. (2025). Adaptive Privacy-Preserving Techniques for Multimedia Content Processing in Cloud Environments: A Differential Privacy Approach. Journal of Science, Innovation & Social Impact, 1(1), 278-293.

[75] Lu, X. (2025). Research on Mobile Advertising Click-Through Rate Prediction Algorithm Based on Differential Privacy. Journal of Science, Innovation & Social Impact, 1(1), 362-371.

[76] Guan, H. (2025). Intelligent Detection and Protection of Personally Identifiable Information in Clinical Text: An Advanced NLP Approach with Optimized Attention Mechanisms. Journal of Science, Innovation & Social Impact, 1(2), 41-52.

[77] Wang, Z., & Kang, A. (2025). FTAFO: A Federated Transparent Adaptive Financial Optimizer for Reducing Third-Party Dependencies in Workflow Management. Journal of Science, Innovation & Social Impact, 1(1), 329-339.

[78] Zhang, Y. (2026). Evaluation of Differential Privacy and Federated Learning for AI-Driven Customer Service Applications. Journal of Sustainability, Policy, and Practice, 2(2), 55-66.

[79] Wu, Z., Zhang, Z., Zhao, Q., & Yan, L. (2025). Privacy-preserving financial transaction analytics with secure multi-party computation. Working Paper.

[80] Li, Y. (2026). Enhancing Financial Compliance Transparency through Automated Data Governance and Intelligent Risk Reporting. Journal of Science, Innovation & Social Impact, 2(1), 299-313.

[81] Zhang, H. (2026, January). Automated Identification of Jurisdiction Clauses in Cross-Border Financial Contracts: A Comparative Study of Rule-Based, Dictionary-Based, and Transformer-Based Approaches. In Proceedings of the 2026 International Conference on Artificial Intelligence and Fintech (pp. 241-248).

[82] Zhang, H., & Shi, W. (2026). Comparative Evaluation of Automated Detection Approaches for Identifying Implicit Compliance Violations in Cross-border Commercial Contract Clauses. Artificial Intelligence and Machine Learning Review, 7(2), 1-22.

[83] Liang, D., & Cai, C. (2025, December). Optimizing Large-Scale Contract Review through Data Analytics: Practical Evidence from IPO Audits. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 242-249).

[84] Zhou, Y., & Long, L. (2026). Causal Effect Evaluation of Personalized Reminder Strategies on Government Welfare Program Enrollment: A Propensity Score Matching Approach. Journal of Computing Innovations and Applications, 4(1), 106-116.

[85] Wang, J. (2025). Multi-Source Data Fusion for Short-Term Demand Forecasting of Seasonal Retail Products: An Empirical Study Using Weather and Social Media Signals. Journal of Science, Innovation & Social Impact, 1(1), 340-349.

[86] Cheng, Z. (2025). AI Enabled Cardiovascular Disease Risk Prediction through Multimodal Data Fusion: A Predictive Analytics Approach. Journal of Sustainability, Policy, and Practice, 1(2), 98-109.

[87] Zhang, C. (2025). Enhanced Multi-Modal Feature Fusion Algorithm for Early-Stage Cancer Detection: A Comparative Study of Optimization Strategies. Journal of Science, Innovation & Social Impact, 1(1), 318-328.

[88] Guo, Y., & Wei, C. (2026). Latency-Adaptive Feature Fusion Weight Allocation Under Bandwidth Constraints for V2X Cooperative 3D Object Detection. Journal of Advanced Computing Systems, 6(3), 22-31.

[89] Guo, Y. (2025). Reliability Assessment and Adaptive Fusion Algorithm for Multi-Sensor Data in Autonomous Driving under Adverse Weather Conditions. Journal of Sustainability, Policy, and Practice, 1(4), 143-155.

[90] Guo, Y. (2025). Performance Evaluation of Lightweight Detection Algorithms on Compact LiDAR-Camera Configurations for Freight Transportation. Journal of Science, Innovation & Social Impact, 1(1), 398-409.

[91] Cao, H., & Shi, W. (2026). Statistical Anomaly Detection Approach for Field Mapping Validation in Enterprise Payroll Data Migration. Journal of Computing Innovations and Applications, 4(1), 137-153.

[92] Shi, W., & Cheng, Z. (2024). Enhanced Adaptive Threshold Algorithms for Real-Time Cardiovascular Risk Prediction from Wearable HRV Data. Journal of Advanced Computing Systems, 4(1), 46-57.

[93] Lei, Y. (2025, October). Intelligent Prediction and Dynamic Scheduling Optimization Strategy for Cloud Computing Resources under Burst Load Scenarios. In Proceedings of the 2025 International Symposium on Machine Learning and Social Computing (pp. 59-67).

[94] Lei, Y., & Holloway, V. (2024). Adaptive Learning-Enhanced Convex Optimization for Energy-Efficient Cloud Resource Scheduling. Journal of Advanced Computing Systems, 4(11), 73-85.

[95] Long, X. (2025, September). Machine Learning-Based Power Consumption Prediction and Dynamic Adjustment Strategies for Enterprise Servers. In Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence (pp. 1310-1319).

[96] Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling. Artificial Intelligence and Machine Learning Review, 6(4), 18-31.

[97] Chen, Y., & Chen, Z. (2025). Multi-Objective Deep Reinforcement Learning for Carbon-Aware Spatiotemporal Workload Scheduling in Geo-Distributed Data Centers. Journal of Advanced Computing Systems, 5(10), 18-30.

[98] Li, Y. (2025, December). Comparative Analysis of Illumination Normalization Methods for Autonomous Driving Under Challenging Lighting Conditions. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 633-639).

[99] Shi, W., & Wang, J. (2026). Intelligent Path Optimization for Carbon-Constrained Last-Mile Delivery: A Reinforcement Learning and Heuristic Approach. Journal of Advanced Computing Systems, 6(1), 19-31.

[100] Li, Y. (2026). Performance Benchmarking and Optimization Strategies for Depth Estimation Algorithms in Unstructured Environments. Journal of Sustainability, Policy, and Practice, 2(2), 32-43.

[101] Guan, H. (2025). Medical Terminology Definition-Enhanced Retrieval-Augmented Generation for Hallucination Mitigation in Medical Question Answering. Journal of Science, Innovation & Social Impact, 1(1), 222-240.

[102] Zhang, D., & Wang, Y. (2025). AI-driven quality assessment and investment risk identification for carbon credit projects in developing countries. Pinnacle Academic Press Proceedings Series, 3, 76-92.

[103] Zhang, D., & Ma, X. (2025). Machine Learning-Based Credit Risk Assessment for Green Bonds: Climate Factor Integration and Default Prediction Analysis. Journal of Sustainability, Policy, and Practice, 1(2), 121-135.

[104] Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[105] Li, M., & Xu, S. (2026). Trustworthy artificial intelligence in financial decision-making: A systematic review of explainability, fairness, and accountability. Applied Intelligence, 55(15), 1005.

[106] Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[107] Zhang, D., & Feng, E. (2024). Quantitative Assessment of Regional Carbon Neutrality Policy Synergies Based on Deep Learning. Journal of Advanced Computing Systems, 4(10), 38-54.

[108] Guan, H. (2025). Context-Aware Semantic Ambiguity Resolution in Cross-Cultural Dialogue Understanding. Journal of Sustainability, Policy, and Practice, 1(2), 136-147.

[109] Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[110] Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems, 4(2), 36-49.

[111] Dong, B., Zhang, D., & Xin, J. (2024). Deep reinforcement learning for optimizing order book imbalance-based high-frequency trading strategies. Journal of Computing Innovations and Applications, 2(2), 33-43.

[112] Zhang, H. (2026). A Comparative Study of NER Methods for Ownership Structure Extraction from M&A Due Diligence Documents. Journal of Sustainability, Policy, and Practice, 2(1), 71-86.

[113] Li, M., Wang, X., & Yu, M. (2025). Comparative Evaluation of Zero-Shot and Few-Shot Performance of Large Language Models in Low-Resource Language Machine Translation. Journal of Global Engineering Review, 3(2), 59-68.

[114] Wen, S., & Tang, T. (2025). A Comparative Evaluation of URL-Sharing, Content Similarity, and Temporal Synchronicity Signals for Detecting Coordinated Inauthentic Behavior in Multilingual Political Discourse. Journal of Global Engineering Review, 3(2), 69-78.

[115] Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[116] Chen, Y., & Tang, T. (2026). Evaluating Prompt Engineering Strategies for Few-Shot Cyber Threat Intelligence Entity and Relation Extraction from Multi-Source Reports. Journal of Science, Innovation & Social Impact, 2(2), 153-164.

[117] Tang, T., & Yu, M. (2024). A Comparative Evaluation of LLM-Generated Semantic Tags versus Classical Text Features (TF-IDF, LDA, BERT Embeddings) for User-Interest Enrichment in Short-Video Recommendation. Artificial Intelligence and Machine Learning Review, 5(1), 129-140.

[118] Zhang, Y., & Li, M. (2026). Accuracy Evaluation of Multi-Factor Forecasting Methods for Customer Service Workload Prediction Under Holiday and Promotional Fluctuations: Evidence from US Service Industry Data. Journal of Advanced Computing Systems, 6(5), 12-20.

[119] Tang, T., & Yu, M. (2024). A Comparative Empirical Study of Semantic Signal Enhancement Methods for User Interest Features in CTR Prediction: Applicability of TF-IDF Weighting, Sentence-BERT Embeddings, and LDA Topic Fusion. Journal of Computing Innovations and Applications, 2(1), 165-174.

[120] Li, M., Zhao, F., & Tang, T. (2024). How Prompt Specificity Affects Edge Case Handling in LLM-Generated Code: An Empirical Evaluation. Artificial Intelligence and Machine Learning Review, 5(4), 139-149.

[121] Zhao, F., Yu, M., & Luo, C. (2024). A Comparative Evaluation of Prompting Strategies for Code Generation with Large Language Models. Journal of Global Engineering Review, 2(1), 1-11.

[122] Xu, S., Zhao, F., & Wang, X. (2025). An Empirical Comparison of Generation Quality and Diversity Between Discrete Diffusion and Autoregressive Text Generation. Artificial Intelligence and Machine Learning Review, 6(2), 16-26.

[123] Zhang, H. (2025). Classifying Tenant Legal Inquiries: A Comparative Study of Traditional and Deep Learning Approaches. Journal of Science, Innovation & Social Impact, 1(1), 452-462.

[124] Zhang, Y. (2026). A Comparative Study of Machine Learning Methods for Automated Customer Service Dialogue Quality Assessment. Journal of Science, Innovation & Social Impact, 2(1), 328-338.

[125] Long, L., Zou, D., & Shi, W. (2026). NLP-Driven Psychological Contract Risk Detection in Cross-Cultural Teams: An XGBoost Approach with Cultural Adaptation. Artificial Intelligence and Machine Learning Review, 7(2), 43-53.

[126] Xu, S., Li, M., & Zhao, F. (2026). Comparative Empirical Evaluation of Hallucination Mitigation Strategies in LLM-Based Text Generation. Journal of Sustainability, Policy, and Practice, 2(3), 38-49.

[127] Xu, S., Ma, Q., Liu, H., & Yue, L. (2026). Continuous Reorganization and Performance Preservation of Agent Memory Structure Under Distributed Change Environments.

[128] Shang, Z., Wei, W., & Bai, W. (2025). Evolving security in llms: A study of jailbreak attacks and defenses. arXiv preprint arXiv:2504.02080.

[129] Wei, C., & Pan, Z. (2026). Accelerating Clinical Trial Recruitment Through Automated Eligibility Screening with Multi-Modal Deep Learning. Journal of Computing Innovations and Applications, 4(1), 1-11.

[130] Ye, H. (2025). Bayesian Optimization-Based AI Framework for Nanobody Screening: Minimizing Experimental Failures in ELISA Detection Systems. Journal of Sustainability, Policy, and Practice, 1(4), 16-31.

[131] Ye, H. (2025). Deep Reinforcement Learning-Driven Efficacy-Toxicity Balance Optimization Strategy for Personalized Drug Combination in Cancer Patients. Journal of Science, Innovation & Social Impact, 1(1), 307-317.

[132] Ye, H. (2025, April). AI-Enhanced Detection of Dynamic Structural Changes in Inflammatory Protein Interfaces: A Case Study of CD11b/Mac-1 Interactions. In 2025 6th International Conference on Computer Engineering and Application (ICCEA) (pp. 2173-2180). IEEE.

[133] Dong, Z., & Jia, R. (2025). Adaptive Dose Optimization Algorithm for LED-based Photodynamic Therapy Based on Deep Reinforcement Learning. Journal of Sustainability, Policy, and Practice, 1(3), 144-155.

[134] Dong, Z., & Zhang, F. (2025). Deep Learning-Based Noise Suppression and Feature Enhancement Algorithm for LED Medical Imaging Applications. Journal of Science, Innovation & Social Impact, 1(1), 9-18.

[135] Zhang, C. (2024). Deep Learning Dose Optimization with Uncertainty Quantification for Intensity-Modulated Radiotherapy: A 3D Radiomics Approach. Artificial Intelligence and Machine Learning Review, 5(2), 116-129.

[136] Zhang, C., & Liu, M. (2026). Integrating Ovarian Reserve Biomarkers with Machine Learning for Gonadotoxicity Risk Prediction in Young Female Cancer Patients: A Scoping Review. Journal of Computing Innovations and Applications, 4(1), 127-136.

[137] Zhang, C., & Xiao, P. (2026). Optimizing Breast Cancer Recurrence Time Prediction with Attention-Enhanced LSTM Networks. Journal of Advanced Computing Systems, 6(1), 80-98.

[138] Zhang, C. (2025, October). Comparative Study of AI Algorithms in Personalized Ovarian Stimulation Protocol Optimization: Predictive Performance Analysis Based on Patient Baseline Characteristics. In Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing (pp. 654-662).

[139] Liu, Y. (2026). AI-Enhanced Healthcare Data Quality Governance: An Integrated Approach for Anomaly Detection and Integrity Verification. Journal of Sustainability, Policy, and Practice, 2(1), 215-229.

[140] Zhang, Q. (2026). Improving Classification Accuracy for Unstructured Medical Documents via Multi-Engine OCR and Deep Learning Collaboration. Journal of Advanced Computing Systems, 6(2), 1-14.

[141] Zhang, Q. (2025). Enhanced Feature Fusion and Transfer Learning for Multi-Format Government Document Classification. Journal of Science, Innovation & Social Impact, 1(1), 427-441.

[142] Zhang, Q. (2025). Comparative Analysis of Pre-Trained Language Models for Medical Document Classification and Priority-Based Workflow Routing. Journal of Sustainability, Policy, and Practice, 1(4), 205-221.

[143] Zhang, Q. (2026). Adaptive OCR Engine Selection and Evaluation for Multi-Format Government Document Digitization. Artificial Intelligence and Machine Learning Review, 7(1), 29-39.

[144] Wang, Z. (2024). Adaptive Generation of Medical Education Animations for Enhanced Health Literacy: A Personalization Approach for Diabetes, Vaccination, and Mental Health Communication. Journal of Advanced Computing Systems, 4(1), 30-45.

[145] Wang, Z. (2025). Cultural-Intelligent Dynamic Medical Animation Generation for Cross-Lingual Telemedicine Communication Enhancement. Journal of Science, Innovation & Social Impact, 1(1), 209-221.

[146] Wang, Z. (2025, April). DeepMotionNet: AI-Driven Predictive Animation State Transitions for Reducing Perceptual Latency in Competitive FPS Games. In 2025 6th International Conference on Computer Engineering and Application (ICCEA) (pp. 01-08). IEEE.

[147] Wang, Z. (2025). Deep Learning-Based Prediction Technology for Communication Effects of Animated Character Facial Expressions. Journal of Sustainability, Policy, and Practice, 1(4), 105-116.

[148] Wang, Z., & Chu, Z. (2025). GAN-Based Intelligent Keyframe Interpolation Method for Character Animation: An Automated In-betweening Approach. Journal of Science, Innovation & Social Impact, 1(2), 29-40.

[149] Li, Z., & Wang, Z. (2024). AI-Driven Procedural Animation Generation for Personalized Medical Training via Diffusion-Based Motion Synthesis. Artificial Intelligence and Machine Learning Review, 5(3), 111-123.

[150] Li, Z., & Wang, Z. (2024). Adaptive Cross-Cultural Medical Animation: Bridging Language and Context in AI-Driven Healthcare Communication. Artificial Intelligence and Machine Learning Review, 5(1), 117-128.

[151] Li, J. (2026). Style Genes: Leveraging Generative AI for Artwork Authentication through Artistic Style Consistency Analysis. Journal of Sustainability, Policy, and Practice, 2(1), 87-100.

[152] Li, J. (2025). Enhanced CNN-based Feature Extraction and Classification for Chinese Artwork Styles. Journal of Science, Innovation & Social Impact, 1(2), 135-148.

[153] Li, J., Zhang, F., & Li, M. (2026). Comparative Effectiveness of Blockchain Provenance Verification on Counterfeit Reduction in Art Transactions: A Multi-Scenario Empirical Assessment. Artificial Intelligence and Machine Learning Review, 7(2), 82-92.

[154] Cao, H. (2024). Detecting Fraudulent Click Patterns in Mobile In-App Browsers: A Multi-dimensional Behavioral Analysis Approach. Artificial Intelligence and Machine Learning Review, 5(2), 130-142.

[155] Cao, H. (2024). Privacy-Preserving Click Pattern Anomaly Detection for Mobile In-App Browser Advertising Fraud. Journal of Computing Innovations and Applications, 2(2), 151-161.

[156] Jia, R., Lu, X., & Whitmore, S. (2024). Feature-Based Detection of Bot Traffic and Click Fraud in Mobile Advertising: A Comparative Analysis. Journal of Computing Innovations and Applications, 2(1), 140-152.

[157] Lu, X. (2025, August). Adaptive Optimization of Advertising Creative Visual Elements Based on Multi-dimensional User Behavior Data. In Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business (pp. 360-368).

[158] Shi, X. (2024). Spatiotemporal Preference Modeling for Ride-Hailing and Context-Aware Recommendations A Machine-Learning Framework. Spectrum of Research, 4(2).

[159] Wang, Z. (2025, October). Machine Learning-Driven Investor-Asset Matching Optimization in Commercial Real Estate Investment Decisions. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 1110-1118).

[160] Wang, Y. (2025). Data-Driven Analysis of Transportation Route Efficiency and Carbon Emission Correlation in Retail Distribution Networks. Journal of Science, Innovation & Social Impact, 1(1), 253-264.

[161] Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[162] Zhang, D., & Wang, Y. (2025). AI-Driven Quality Assessment and Investment Risk Identification for Carbon Credit Projects in Develo Countries. Pinnacle Academic Press Proceedings Series, 3, 76-92.

[163] Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[164] Long, L., & Hu, J. (2026). Multi-Objective Particle Swarm Optimization for Site Selection and Policy Subsidy Maximization of Foreign Renewable Energy Enterprises in the United States. Artificial Intelligence and Machine Learning Review, 7(2), 54-69.

[165] Bai, Y., & Xiao, P. (2026). Adaptive Prompt Selection and Fading Optimization for Autism Skill Acquisition: A Reinforcement Learning Approach. Journal of Advanced Computing Systems, 6(1), 32-44.

[166] Bai, Y. (2025, September). Deep Learning-based Action Recognition for Temporal Analysis and Intervention Effectiveness Assessment in Autism Spectrum Disorder Children's Video Therapy. In Proceedings of the 2025 International Symposium on Artificial Intelligence and Computational Social Sciences (pp. 307-314).

[167] Bai, Y. (2025). Effectiveness Evaluation of Adaptive Difficulty Adjustment Algorithms with Multimodal Feedback for Social Skills Training in Children with Autism Spectrum Disorder. Journal of Sustainability, Policy, and Practice, 1(4), 117-129.

[168] Bai, Y. (2026). Context-Aware Classification of Verbal Operants in Children with ASD Using Deep Learning. Journal of Science, Innovation & Social Impact, 2(1), 232-243.

[169] Shi, W., & Bai, Y. (2024). Adaptive Learning Rate Optimization for Personalized Educational Interventions in Autism Spectrum Disorder: A Multi-Objective Reinforcement Learning Approach. Artificial Intelligence and Machine Learning Review, 5(4), 128-138.

[170] Chung, P. T. (2025). Attention-Enhanced YOLO for Real-Time Defect Detection in 3D-Printed Dental Prostheses. Journal of Science, Innovation & Social Impact, 1(2), 119-134.

[171] Chung, P. T. (2026). Comparative Evaluation of Machine Learning Algorithms for Spectrophotometric Dental Shade Classification. Journal of Sustainability, Policy, and Practice, 2(1), 204-214.

[172] Chung, P. T. (2026). Multi-Objective Optimization of Process Parameters for Dental Resin 3D Printing Using Improved NSGA-II Algorithm. Journal of Science, Innovation & Social Impact, 2(1), 276-287.

[173] Chung, P. T. (2025, December). Data Mining Methods for Biomechanical Property Prediction of Biomedical Materials Based on Optimized Feature Dimensionality Reduction. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 174-180).

[174] Chung, P. T. (2025, December). Enhancing Dental Polymer Formulation through Interpretable Machine Learning: A Comparative Analysis of Feature Selection and Algorithm Performance. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 234-241).

[175] Li, Z., & Chen, Z. (2025). Performance Evaluation of Prompt Generation Strategies for AI Agents in Online Programming Education. Journal of Advanced Computing Systems, 5(9), 14-27.

[176] Zou, D., Chen, Z., & Ling, Z. (2025). A Comparative Evaluation of Deep Learning Paradigms for Low-Light Image Enhancement: From CNNs to Diffusion Models. Journal of Computing Innovations and Applications, 3(2), 85-95.

[177] Weng, H., & Lei, Y. (2024). Cross-Modal Artifact Mining for Generalizable Deepfake Detection in the Wild. Journal of Computing Innovations and Applications, 2(2), 78-87.

[178] Shi, X., & Weng, H. (2024). Comparative Analysis of Unsupervised Learning Approaches for Anomalous Billing Pattern Detection in Healthcare Payment Integrity. Journal of Computing Innovations and Applications, 2(1), 111-127.

[179] Li, J., Liu, M., & Li, M. (2026). Comparative Evaluation of Ensemble Learning Algorithms for Visitor Engagement Prediction and Content Recommendation Optimization in Virtual Museum Environments. Journal of Advanced Computing Systems, 6(2), 64-74.

[180] Wang, X., Fu, X., & Zou, D. (2025). Passage, Sentence, or Proposition? An Empirical Comparison of Retrieval Granularity Effects on LLM Answer Accuracy in Retrieval-Augmented Generation. Journal of Global Engineering Review, 3(1), 81-90.

[181] Xiao, P., & Fu, X. (2026). Comparative Evaluation of Post-Hoc Feature Attribution Methods on Tabular Financial Data: Faithfulness, Stability, and Computational Efficiency. Journal of Science, Innovation & Social Impact, 2(3), 1-11.

[182] Li, Y., & Fu, X. (2026). Comparative Evaluation of Graph Neural Networks for Cross-Market Risk Contagion Path Identification in Multi-Layer Financial Networks. Journal of Sustainability, Policy, and Practice, 2(3), 1-14.

[183] Tang, T., Fu, X., & Luo, C. (2026). An Empirical Comparison of High-Order Feature Interaction Operators for Conversion Rate Prediction in Sparse, High-Cardinality Message-Ads Traffic: Accuracy, Efficiency, and Offline-Online Consistency. Journal of Science, Innovation & Social Impact, 2(3), 12-22.

[184] Fu, X., Tang, T., & Luo, C. (2026). An Empirical Comparison of ReAct, Reflexion, Plan-and-Solve, and Tree-of-Thought Planning Strategies on Financial Question Answering and Numerical Reasoning Tasks. Journal of Science, Innovation & Social Impact, 2(3), 23-34.

[185] Fu, X., & Zou, D. (2025). A Comparative Empirical Study of Over-Refusal Behavior in Closed-Source Large Language Models on Pseudo-Harmful Prompts. Artificial Intelligence and Machine Learning Review, 6(3), 34-45.

[186] Fu, X., & Zhao, F. (2025). An Empirical Comparison of Few-Shot Example Selection Strategies for In-Context Learning on Public Reasoning and QA Benchmarks. Journal of Computing Innovations and Applications, 3(2), 119-128.

[187] Zhao, F., & Tang, T. (2026). AI-Based Sentiment Analysis for Stock Market Prediction: A Systematic Literature Review. Journal of Sustainability, Policy, and Practice, 2(3), 115-124.

[188] Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments. Innovations and Applications, 2(1), 140-152.

[189] Xu, S., Ma, Q., Liu, H., & Yue, L. (2026). Continuous Reorganization and Performance Preservation of Agent Memory Structure Under Distributed Change Environments. Innovations and Applications, 4(1), 127-136.

[190] Ge, L. (2024). Enhancing Financial Audit Efficiency Through RPA Implementation: A Comparative Analysis in Manufacturing Industry. Journal of Computing Innovations and Applications, 2(1), 62-73.

[191] Zhao, F., Zhang, M., Zhou, S., & Lou, Q. (2024). Application of deep reinforcement learning for cryptocurrency market trend forecasting and risk management.

[192] Xiao, P., Wang, Y., & Montgomery, I. (2024). Deep Reinforcement Learning for Route Optimization in E-commerce Return Management. Journal of Computing Innovations and Applications, 2(2), 100-110.

[193] Wang, J., & Jia, R. (2026). AI-Enhanced What-If Scenario Analysis in Supply Chain Digital Twins: A Multi-Objective Trade-Off Perspective on Cost, Resilience, and Carbon Efficiency. Journal of Computing Innovations and Applications, 4(1), 97-105.

[194] Wang, J. (2025, October). Artificial Intelligence-Driven Seasonal Consumption Forecasting and Resource Allocation Optimization in Luxury Brand Marketing. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 1119-1127).

[195] Deng, M., & Zou, D. (2026). Application of Cross-Modal Content Consistency Verification in Social Media Misinformation Detection. Artificial Intelligence and Machine Learning Review, 7(1), 40-52.

[196] Deng, M., & Xu, S. (2026). Temporal-Structural Propagation Graph Analysis for Coordinated Misinformation Campaign Detection and Source Attribution in Social Networks. Journal of Advanced Computing Systems, 6(5), 1-11.