AI & SOCIETY
https://doi.org/10.1007/s00146-026-03049-y

REVIEW

Trust in human–AI collaboration in finance: a bibliometric–systematic literature review

Mario Mirabile[1,2] · **Giovanni Emanuele Corazza**[2] · **Jose María Alonso-Moral**[1]

Received: 14 December 2025 / Accepted: 8 April 2026

Abstract
Artificial intelligence (AI) is becoming deeply embedded in financial services, including credit scoring, robo-advisory, trading, compliance, and reporting. In these contexts, failures of trust in human–AI collaboration do not merely affect technology adoption but raise fiduciary, reputational, and systemic concerns. Yet, despite its centrality, trust remains conceptually fragmented and inconsistently operationalized across the literature. To address this fragmentation, this study presents a Bibliometric–Systematic Literature Review of trust in human–AI collaboration in finance. Accordingly, we first stabilize trust in the financial context as a latent evaluative belief under uncertainty by specifying the trustor, trustee, and object of trust, and by distinguishing trust from other constructs. Then following a PRISMA-guided Scopus retrieval (June 9, 2025), 430 records were screened, yielding a final corpus of 114 finance-specific publications published between 2018 and 2025. Using bibliographic coupling, weighted Leiden clustering, and centrality metrics, the review maps the intellectual structure of the field and identifies six research clusters: (i) AI governance in finance, (ii) eXplainable AI for finance, (iii) anthropomorphism in financial AI agents, (iv) user-interface design for human–AI interaction in finance, (v) robo-advisors for financial decision-making, and (vi) infrastructural trust technologies. Across these clusters, trust is variously framed as cognitive, affective, procedural, or infrastructural, with limited integration between analytical levels. Building on the bibliometric mapping and qualitative synthesis, the study develops a multi-level socio-technical framework that organizes how trust is discussed in the literature at the micro-level (user perceptions and calibration), meso-level (organizational design and corporate AI governance), and macro-level (regulatory and infrastructural). The micro–meso–macro framework is operationalized through eight analytically distinct propositions that synthesize recurrent patterns regarding trust calibration, overreliance, transparency-as-assurance, accountability, and systemic trust vulnerability. Four finance-oriented use cases illustrate how trust is treated as a distributed property of individuals, organizations, and institutions rather than as a feature of AI systems alone. By consolidating a fragmented body of work and clarifying its conceptual structure, this review provides a bounded, governance-oriented foundation for future empirical, experimental, and longitudinal research on trust in human–AI collaboration in finance.

Keywords Trust · Artificial intelligence · Human–AI collaboration · Finance · Bibliometric–systematic literature review · Trust framework

✉ Mario Mirabile
mario.mirabile@rai.usc.gal

Giovanni Emanuele Corazza
giovanni.corazza@unibo.it

Jose María Alonso-Moral
josemaria.alonso.moral@usc.es

[1] Centro Singular de Investigacion en Tecnoloxias Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain

[2] Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi", University of Bologna, Bologna, Italy

1 Introduction

Artificial intelligence (AI) is becoming deeply embedded in financial services (Giudici 2018; Maple et al. 2023; Financial Stability Board 2025; Foucault et al. 2025). While this integration promises efficiency and innovation, it also introduces acute fiduciary, reputational, and systemic risks when trust in human–AI collaboration breaks down. In finance, trust failures do not simply slow adoption; they weaken accountability, obscure decision rationales, amplify model-driven bias and correlated behavior, and facilitate fraud and

Published online: 08 May 2026

disinformation. At scale, these dynamics can erode confidence in financial institutions, distort market functioning, and threaten financial stability itself (Aldasoro et al. 2025; Daníelsson et al. 2022; Uuk et al. 2024; Cecchetti et al. 2025). In this context, a growing body of work has examined how AI methods are adapted and applied in financial settings. Specifically, financial machine learning (ML) has emerged as a specialized field focused on the particular challenges of financial data (Dixon et al. 2020; Scholes 2025). Also, Generative AI (Gen-AI) and Agentic AI in financial contexts extend the capability set supporting end-to-end financial workflows, with Gen-AI strengthening generative and analytical functions and Agentic AI supporting more autonomous, goal-driven coordination across customers and teams (Li et al. 2023b; Abou Ali et al.2026 ; Aldridge et al. 2025).

These advances create substantial opportunities, but also introduce significant risks, especially in human–AI collaboration, where humans and AI jointly pursue shared goals through complementary strengths, task co-management, and adaptive feedback (Dellermann et al. 2019; Wang et al. 2019, 2020; Sowa et al. 2021; Sharma et al. 2023). Effective collaboration, however, requires more than technical coordination. It involves processes of mutual adaptation and trust-building between humans and artificial systems (Makarius et al. 2020), recognizing how AI is a tool with agency but no intelligence (Floridi 2025).

In finance, such collaboration has been studied in various domains, including robo-advisory (Jung et al. 2018; Bhatia et al. 2020; Brenner and Meyll 2020), credit scoring (Bussmann et al. 2021; Heng and Subramanian 2023), risk management (Giudici 2018; Heng and Subramanian 2023; Leitner et al. 2024), customer service (Belanche et al. 2019; Graham et al. 2025), fraud detection (Alkhalil et al. 2021), supply chains and regulatory auditing (Garcia et al. 2024). While these domains differ in scope, they share a common thread: the success of human–AI collaboration depends not only on technical accuracy but on the quality of the collaboration, where trust, more than accuracy, is the sector's most vital asset, with delegating savings and investments representing the ultimate act of confidence (Loaiza and Rigobon 2025).

Yet multiple forces threaten this trust. At the technical level, biased or low-quality data, opaque algorithms, and limited auditability can erode confidence (Leitner et al. 2024). Industry-level risks such as provider concentration, and correlated model use may lead to herding, fragility, and competition challenges (Leitner et al. 2024; Dou et al. 2025). On the customer side, AI-enabled phishing, deepfakes, and data leakage raise concerns for privacy and financial safety (Giudici 2018; Alkhalil et al. 2021).

These risks might grow when interpretable models are replaced by opaque algorithms. Although these may perform better in the short term, their opacity can hide errors, biases, and inequities (Breiman 2001; Cramer et al. 2018; Gebru et al. 2021; Garcia et al. 2024).

In response, researchers have increasingly examined conceptual approaches for defining and strengthening trust. Extensions of the Technology Acceptance Model (TAM), originally proposed by Davis (1989), have incorporated trust as a key determinant of adoption (Gefen et al. 2003). Further work has analyzed emotional dynamics and trust-based reactions to AI advisors (Brenner and Meyll 2020). Also, eXplainable AI (XAI) is frequently positioned in the literature as a strategy to enhance interpretability and fairness (Guidotti et al. 2018; Lipton 2018; Floridi 2019; Gunning 2019; Miller 2019; Rudin 2019; Barredo Arrieta et al. 2020; Ali et al. 2023). Specifically, in finance, XAI research has classified methods by task (Černevičienė and Kabašinskas 2024), evaluated robustness in credit risk (Heng and Subramanian 2023), cataloged techniques (Weber et al. 2024), tested model-agnostic approaches (Khan et al. 2025), and demonstrated interpretable statistical methods that improve both transparency and performance (Bussmann et al. 2021).

Despite these valuable contributions, the literature remains fragmented. What is still missing is a cohesive narrative that explains how trust in human–AI collaboration in finance is conceptualized, measured, and fostered. This gap underscores the need for a systematic review that critically analyzes existing works and provides a clearer foundation for research on human–AI collaboration in finance.

To address this gap, the study examines the following research questions (RQs):

RQ1: What are the most distinctive research themes related to trust in human–AI collaboration in finance?

RQ2: How have research communities conceptualized trust in the most influential studies on trust in human–AI collaboration in finance?

RQ3: What methods and tools have been proposed in the identified most influential studies to investigate specific trust dimensions in human–AI collaboration in financial contexts?

RQ4: What emerging themes and research gaps exist in research on trust in human–AI collaboration in finance across identified clusters?

Additionally, building on these research questions, this review also addresses a higher-level, integrative question that emerges from the literature synthesis:

RQ5: Which governance and design levers are most consistently associated with justified trust, understood as calibrated reliance, rather than blind acceptance or overreliance in human–AI collaboration in finance?

To address our RQs, we conducted a Bibliometric Systematic Literature Review (B-SLR), following guidelines given by Marzi et al. (2025), but also integrating PRISMA-guided retrieval (Page et al. 2021a, b), bibliographic coupling (Zupic and Čater 2015; Donthu et al. 2021), and centrality mapping (Bonacich 2007; Fronzetti Colladon and Naldi 2020). From 430 Scopus records, we screened and retained 114 finance-specific studies. We identified document-based clusters using the weighted Leiden algorithm (Traag et al. 2019), mapped the most distinctive themes (RQ1), and performed a manual full-text review of most influential papers to examine trust conceptualization (RQ2) and methodological approaches (RQ3). Also, we analyzed thematic structures to highlight emerging trends and research gaps (RQ4).

A key contribution of this study is the development of a micro–meso–macro (MMM) socio-technical framework of trust in human–AI collaboration in finance (RQ5). Drawing on bibliometric structures and qualitative analysis, the framework organizes trust mechanisms across micro-level user perceptions and calibration, meso-level organizational design and corporate governance arrangements, and macro-level regulatory and infrastructural conditions. To address persistent conceptual fragmentation in the literature, the framework is explicitly operationalized through analytically distinct and testable propositions that specify when and how trust, overreliance, transparency, and accountability are expected to emerge in financial settings.

Before addressing these questions, it is necessary to clarify how this review conceptualizes "trust" in human–AI collaboration in finance and how it is distinguished from related constructs that are often used interchangeably in the literature (Table 1). Based on this classification, the review codes papers according to which of these constructs they explicitly measure or implicitly invoke.

Finally, the paper is organized as follows. Section 2 describes the bibliometric–systematic review methodology. Section 3 reports the bibliometric results, thematic clusters, and synthesis of trust conceptualizations. Section 4 integrates these findings into a multi-level socio-technical framework, discusses governance implications, and Sect. 5 concludes with limitations and future research directions.

2 Methodology

This study implemented a B-SLR approach (Marzi et al. 2025), integrating PRISMA-guided protocol (Page et al. 2021a, b), bibliographic coupling (Zupic and Čater 2015; Donthu et al. 2021), centrality mapping (Bonacich 2007), and manual full-text review.

In this research, we use "human–AI collaboration" as an umbrella term for what other works call "partnership" or "teaming", treating these labels interchangeably in our Scopus search to capture settings where humans and AI jointly pursue shared goals (Dellermann et al. 2019; Wang et al. 2019, 2020; Sowa et al. 2021; Sharma et al. 2023). Also, to ensure accessible, high-contrast visuals, we used Tol (2021)'s colorblind-safe muted palette, maintaining consistent encoding across plots.

The methodology followed linear and sequential steps, as depicted in Fig. 1.

Table 1 Core constructs in human–AI collaboration in finance

Construct	Definition and scope
Trust in finance	Here, trust is understood as a latent belief held by a trustor (e.g., clients, professional users such as analysts or advisors, compliance officers, organizational decision-makers, board members, or regulators) toward a trustee. The trustee may include an AI system, an AI model, an AI-enabled interface, a financial institution deploying AI, a third-party AI vendor, or the broader socio-technical arrangement in which these elements are embedded. Trust reflects the belief that the trustee will act competently, with integrity, and in accordance with its designated role, obligations, and constraints under conditions of uncertainty. The object of trust concerns expectations regarding recommendation accuracy, robustness and reliability, fairness and non-discrimination, transparency and auditability, accountability and compliance, and fiduciary alignment. These beliefs may be shaped by technical features (e.g., model performance or explainability), interface design, organizational practices, and regulatory or infrastructural assurances; however, they remain analytically distinct from observable behaviors (e.g., reliance or adoption) and from system capabilities themselves, which constitute antecedent conditions rather than trust per se
Reliance	Observable behavioral dependence on AI outputs (e.g., following recommendations, delegating decisions, or deferring judgment), which may occur with or without justified trust
Calibration	The degree of alignment between subjective trust in an AI system and its actual capabilities and limitations, encompassing both under-reliance and overreliance
Legitimacy	Perceived appropriateness of AI use within normative, organizational, and regulatory frameworks, extending beyond individual trust relationships to broader institutional acceptance

Fig. 1 Sequential methodology integrating multiple steps

2.1 PRISMA-guided protocol

A search was conducted in Scopus on June 9, 2025, and records were exported in CSV format. The query string was:

```
TITLE-ABS-KEY ( trust* OR distrust* OR mistrust* ) AND ( "AI" OR "artificial intelligence" ) AND ( coworker* OR colleague* OR partner* OR collaborat* OR team* OR group* ) AND ( "human-AI" OR "human-agent" OR "human-machine" OR "human-bot" OR "human-autonomy" ) AND ( hybrid* OR collect* OR augment* OR extend* ) AND ( financ* ) AND DOCTYPE ( ar OR cp OR re ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

To ensure transparency in construct operationalization, alternative search terms related to trust (i.e., acceptance, adoption, assurance, automation bias, automation complacency, confidence, reliance, safety) were systematically evaluated but excluded to preserve conceptual precision; a complete overview of considered terms and exclusion rationales is provided in the Supplementary Materials.

Following PRISMA 2020 (Page et al. 2021a, b), 430 records were identified in Scopus, as shown in Fig. 2. After cleaning and automatic deduplication, 426 remained. Eligibility required both trust in human–AI collaboration as a primary construct, and an explicit financial application; records in which trust or finance was only incidental were excluded. We limited the corpus to English-language items indexed in Scopus as articles, conference papers, or reviews. In total, 114 records satisfied the eligibility criteria for inclusion in the bibliometric analysis.

2.2 Bibliographic coupling and centrality mapping for community detection

Bibliographic coupling was performed on the 114 selected records, linking documents based on shared references (Zupic and Čater 2015; Donthu et al. 2021). Community detection was conducted using the weighted Leiden algorithm (Traag et al. 2019) in Python v3.12.9 adopting Networkx (2025). This produced six clusters. Distinctiveness centrality was employed to identify the unique themes that characterize each cluster (Fronzetti Colladon and Naldi 2020). To determine the most structurally influential documents, eigenvector centrality served as the primary metric (Bonacich 2007). In addition, degree centrality, betweenness centrality, and closeness centrality were calculated to capture aspects of the network, namely, connectivity among papers, bridging roles, and overall integration within the structure, that complement the information provided by eigenvector centrality (Zhang and Luo 2017). The definitions and interpretations of these metrics are summarized in Table 2 for clarity.

The bibliometric techniques employed in this study support structural and relational insights into the literature, including patterns of co-citation, thematic clustering, and the relative positional influence of documents within the network. These methods do not provide causal explanations, normative evaluations, or empirical validation of conceptual definitions, measurement strategies, or governance mechanisms. Accordingly, all interpretive claims in the Results section are restricted to how trust is framed, clustered, and positioned in the literature, rather than why specific trust configurations arise or which approaches are superior or more effective.

Following the quantitative analysis, we conducted a manual full-text review of the five most structurally influential papers within each cluster, as ranked by raw and locally normalized eigenvector centrality, with the aim of examining trust conceptualizations, methodological approaches, and recurring relational patterns between trust antecedents, governance mechanisms, and outcomes. These insights directly informed the thematic synthesis, the resulting MMM framework, and the analytical propositions articulated.

2.3 B-SLR and synthesis

Building on the above procedures, we structured the methodology around five RQs:

RQ1: **Distinctive cluster themes** (i.e., "What are the most distinctive research themes related to trust in human–AI collaboration in finance?"): keywords were extracted and standardized using a curated alias dictionary to harmonize synonyms (e.g., *"artificial intelligence", "machine learning"* → ai; *"eXplainable AI", "interpretable ML"* → xai). Then we ranked terms using distinctiveness centrality, which favors ties to peripheral, low-degree neighbors over hubs and thus highlights keywords that uniquely characterize clusters (Fronzetti Colladon and Naldi 2020).

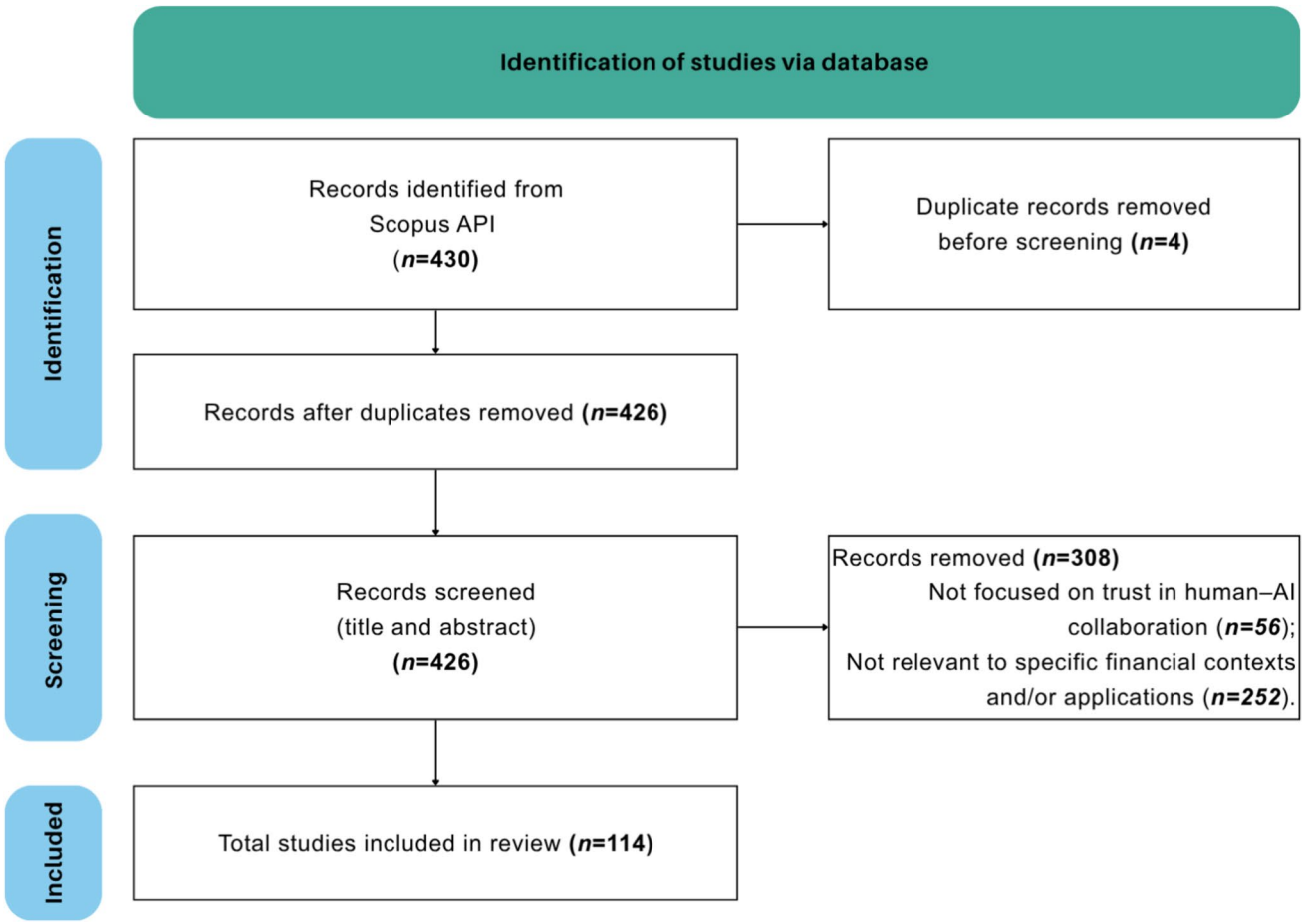

Fig. 2 PRISMA 2020 flow diagram. Adapted from Page et al. (2021a, 2021b)

Table 2 Definitions of the centrality metrics used in this study

Metric	Description
Distinctiveness	Captures the uniqueness of a node by favoring connections to peripheral, low-degree neighbors over hubs, thus highlighting keywords that uniquely characterize clusters
Eigenvector	Measures a node's influence by considering not only its connections but also the importance of the nodes it connects to. Nodes linked to other highly influential nodes receive higher scores
Degree	Counts the number of direct links (edges) a node has, reflecting its immediate connectivity and potential influence
Betweenness	Quantifies how often a node lies on the shortest paths between other nodes, identifying nodes that act as bridges or control points for information flow
Closeness	Indicates how near a node is to all others in the network, based on the shortest path distances. High values mean a node can reach others quickly and efficiently

RQ2: **Trust conceptualization** (i.e., "How have research communities conceptualized trust in the most influential studies on trust in human–AI collaboration in finance?"): the five most central documents in each cluster, sorted by node ID and eigenvector centrality, were examined to identify trust conceptualizations. Data extraction followed an interpretive synthesis procedure, as suggested by Marzi et al. (2025).

RQ3: **Methods, tools, and trust dimensions mapping** (i.e., "What methods and tools have been proposed in the identified most influential studies to investigate specific trust dimensions in human–AI collaboration in financial contexts?"): methodological

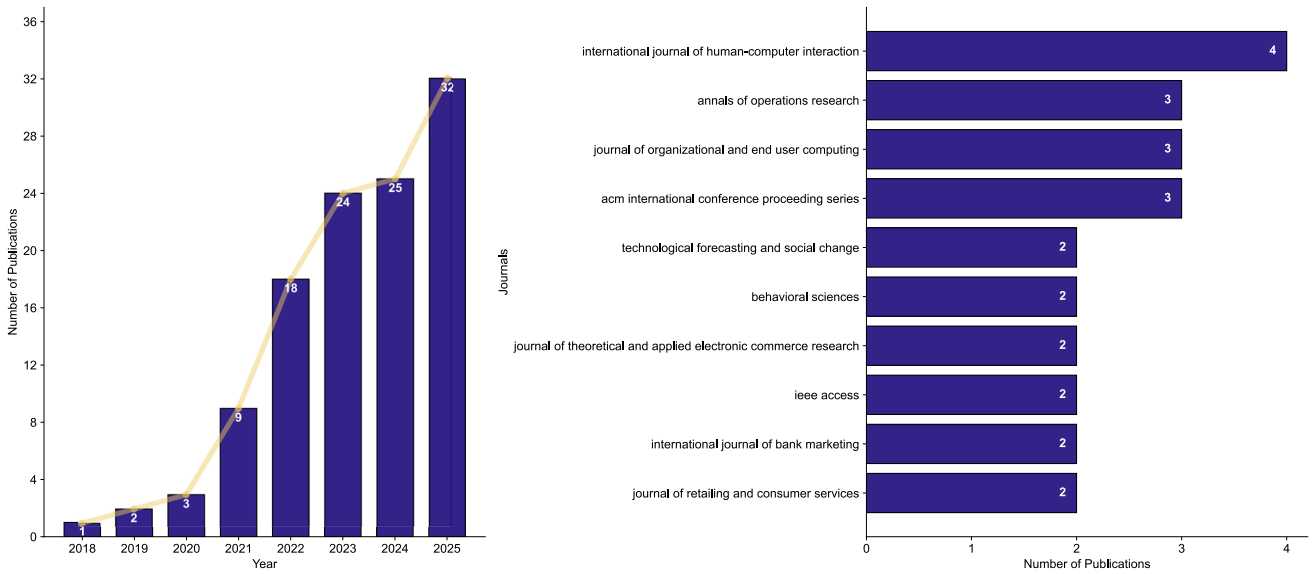

Fig. 3 Annual publication trend from 2018 to 2025 (left panel) and distribution of publications across journals (right panel)

details, including experimental designs, modeling approaches, and tools, were extracted from the reviewed documents. These were mapped to corresponding trust dimensions identified in the literature.

RQ4: **Emerging topics and gap identification** (i.e., "What emerging themes and research gaps exist in research on trust in human–AI collaboration in finance across identified clusters?"): to identify emerging themes and research gaps across clusters, a full-text analysis was conducted on the top five most influential papers within each cluster. The interconnections among the clusters, themes, and gaps were depicted using a Sankey diagram produced with SankeyMATIC (2025).

RQ5: **Integrative framework and propositions** (i.e., "Which governance and design levers are most consistently associated with justified trust, understood as calibrated reliance, rather than blind acceptance or overreliance in human–AI collaboration in finance?"): to address this integrative question, recurrent governance and design levers were synthesized. Based on patterns observed in the bibliometric mapping and the manual review of central papers, these levers were organized into the MMM socio-technical framework and formalized into analytically distinct propositions spanning micro-, meso-, macro-, and cross-level dynamics.

3 Results

3.1 Overview of the dataset and bibliometric structures

This subsection reports the characteristics of the retrieved dataset and the bibliometric structures identified. It presents the PRISMA-guided results, annual publication trend, and journal productivity distributions. Additionally, it visualizes the full bibliographic coupling network, the density contour plot, and the country-level distribution of the connected papers. Specifically, the initial Scopus retrieval produced 426 records. After manual screening of titles and abstracts, 114 finance-specific documents were retained for analysis. Figure 3 illustrates the annual publication trend (left) and the distribution of publications across journals (right). The left subplot shows a consistent rise in research output between 2018 and 2025. The right subplot presents the distribution of publications across the top ten journals.

Bibliographic coupling was applied to this set using the weighted Leiden community detection algorithm, which groups documents based on shared references (Traag et al. 2019). A minimum threshold of one shared reference was applied, yielding a connected network of 105 nodes and 610 coupling links, with 9 documents excluded as isolated. The resulting structure (Fig. 4) reveals 6 distinct clusters ranging from 13 to 21 nodes.

As illustrated in Fig. 5, the network forms six clusters with varying cohesion and connectivity, and an overall

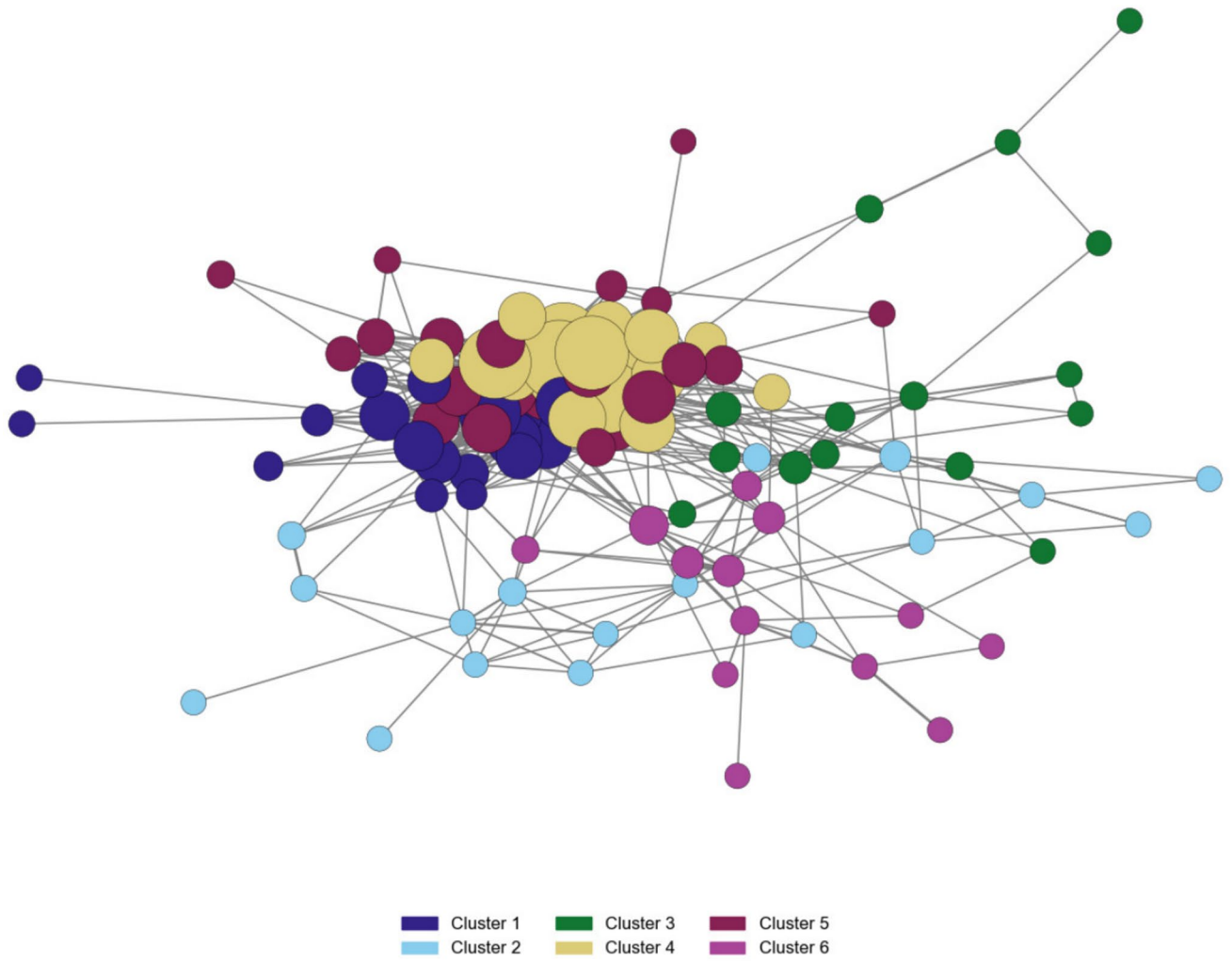

Fig. 4 Bibliographic coupling network illustrating the relationships between documents based on shared references. Node size indicates document citation count, and edge thickness reflects coupling strength

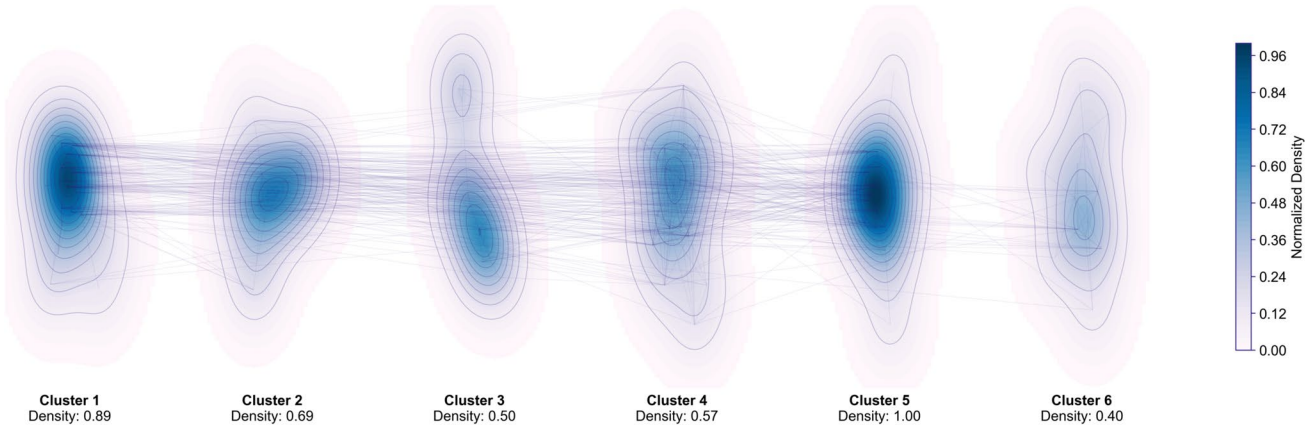

Fig. 5 Density contour plot illustrating cohesion across network clusters. Darker regions indicate stronger internal connectivity

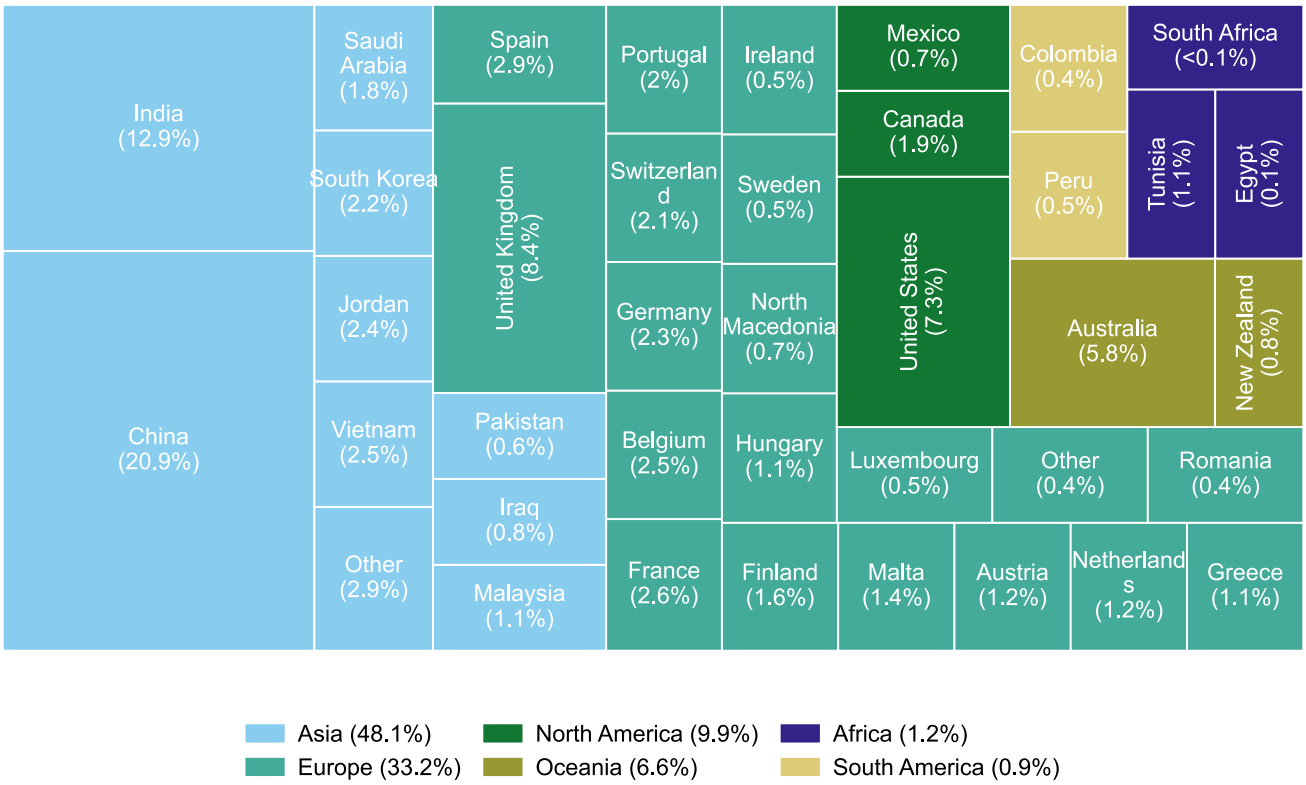

Fig. 6 Geographic distribution of publications based on country affiliation metadata from multi-country papers, using the fractional counting method. Each box is approximately sized proportionally to its global percentage share of fractionalized publication counts and annotated with its percentage. Colors denote continents (colour figure online).

network modularity of 0.38 (Newman 2006). Clusters 5 and 1 are the most cohesive and form part of the network core, while the remaining clusters are smaller and less tightly connected, with weaker links to the rest of the network.

Furthermore, country-level affiliations for the 105 papers were analyzed using full and fractional counting (Perianes-Rodriguez et al. 2016). As shown in Fig. 6, contributions are globally distributed, with China and India providing the largest shares, followed by several countries in Europe, North America, and Oceania. A broader set of countries contributes smaller proportions, reflecting a diverse but uneven international research landscape.

3.2 Distinctive themes within identified clusters (RQ1)

To address RQ1 (*"What are the most distinctive research themes related to trust in human–AI collaboration in finance?"*), we first applied bibliographic coupling to organize the literature into six clusters. We then compute and report Global Weight Distinctiveness (GWD) values for each cluster and keywords in Table 3. Notice that, GWD is computed as defined by Fronzetti Colladon and Naldi (2020). Through this process, we identified the thematic core of each cluster and assigned six overarching thematic labels that capture their primary focus while highlighting the distinctive yet interconnected nature of the research areas represented in the dataset:

Cluster 1: **AI Governance in Finance:** the thematic focus centers on *AI*, *AI governance*, *trust*, and *XAI*, with governance levers such as *responsible AI*, *controllability*, *intelligibility*, *AI transparency*, *AI adoption* and *finance*.

Cluster 2: **XAI for Finance:** key terms include *XAI*, *AI*, and *finance*, alongside assurance terms, *ethics*, *regulatory compliance*, and *fairness*; and design/process cues such as *human-centric design*, *transparency*, and *human experts*. *Trust* is framed in this cluster as being closely associated with transparency and auditability.

Cluster 3: **Anthropomorphism in Financial AI Agents:** the emphasis lies on *finance*, *anthropomorphism*, *AI agents*, and *trust*, with calibration

Table 3 Top ten distinctive keywords per cluster ranked by Global Weight Distinctiveness (GWD)

Top ten distinctive keywords across thematic clusters					
1: AI Governance in Finance		2: XAI for Finance		3: Anthropomorphism in Financial AI Agents	
Keyword	*GWD*	*Keyword*	*GWD*	*Keyword*	*GWD*
AI	74.07	XAI	96.26	Finance	76.21
AI governance	48.10	AI	58.72	Anthropomorphism	71.64
Trust	39.97	Finance	54.88	AI agents	41.44
XAI	31.60	Ethics	35.38	Trust	34.14
Finance	18.27	Regulatory compliance	27.20	Trust repair	17.88
Responsible AI	15.71	Trust	17.49	Human–AI interaction	17.56
Controllability	13.47	Human-centric design	15.65	Attitude towards usage	12.78
AI adoption	11.19	Transparency	15.07	Behavioral finance	10.42
Intelligibility	10.77	Human experts	12.32	CASA paradigm	9.20
AI transparency	10.36	Fairness	10.84	Ethics	7.65
4: User Interface Design for Human–AI in Finance		5: Robo-Advisors for Financial Decision-Making		6: Infrastructural Trust Technologies	
Keyword	*GWD*	*Keyword*	*GWD*	*Keyword*	*GWD*
Chatbot design	97.15	Finance	68.38	Blockchain	120.19
Fintech	74.32	AI	68.13	Supply chain	93.92
Chatbot	63.22	Robo-advisor	66.46	Agrifood traceability	88.18
Trust	56.45	Acceptance	52.46	Transparency	25.48
Acceptance	51.86	Trust	41.14	Trust	20.74
Perception	40.41	Investment decision	29.95	Sustainability	17.51
Privacy	27.69	Overreliance	29.21	Adoption	12.78
AI	26.20	Decision-making	27.65	Privacy and security	12.40
Social presence	11.42	User experience design	22.03	IoT	10.76
Human–AI interaction	8.28	AI anxiety	20.93	AI	4.34

The table highlights the most distinctive terms that characterize each research cluster, with thematic labels, top ten higher GWD values indicating stronger distinctiveness and thematic specificity of the keyword within its cluster

mechanisms like *trust repair* and *human–AI interaction* and attitude constructs such as *attitude towards usage*. Behavioral frames (*behavioral finance*, *CASA paradigm*) and *ethics* appear as supporting themes. Notice that CASA stands for Computer and Social Actors.

Cluster 4: **User Interface Design for Human–AI in Finance:** this cluster emphasizes *chatbot design*, *fintech*, *chatbot*, and *trust*, with adoption-relevant markers, such as *acceptance* and *perception*. Constraints and enablers include *privacy*, *social presence*, and *AI* anchoring *human–AI interaction* context.

Cluster 5: **Robo-Advisors for Financial Decision-Making:** central themes include *finance*, *AI*, and *robo-advisor*, linking *acceptance* and *trust* to decision constructs (*investment decision*, *decision-making*). Risk and user experience factors (*overreliance*, *user experience design*, *AI anxiety*) highlight the need for calibrated reliance on algorithmic advice.

Cluster 6: **Infrastructural Trust Technologies:** this cluster focuses on finance-related socio-technical infrastructures, such as *blockchain*, *supply chain*, and *agrifood traceability*, through which *trust* is partially outsourced to technological and institutional mechanisms. Within these infrastructures, *AI* plays an analytically constitutive role in operationalizing trust at scale, providing a macro-level contrast to user- and organization-centric forms of trust calibration in human–AI collaboration in finance.

Table 4 Top five nodes per cluster ranked by raw eigenvector centrality computed on each cluster's induced subgraph H_c

Top 5 nodes by eigenvector centrality across clusters								
1: AI Governance in Finance			2: XAI for Finance			3: Anthropomorphism in Financial AI Agents		
Node	EV_r	EV_{nrm}	Node	EV_r	EV_{nrm}	Node	EV_r	EV_{nrm}
30	0.174	0.27	9	0.327	0.61	43	0.353	0.70
36	0.256	0.39	29	0.489	0.93	79	0.326	0.65
38	**0.640**	1.00	41	0.295	0.55	100	0.376	0.75
69	0.614	0.96	**63**	**0.524**	1.00	**107**	**0.501**	1.00
99	0.169	0.26	78	0.402	0.76	111	0.494	0.99
4: User Interface Design for Human–AI in Finance			5: Robo-Advisors for Financial Decision-Making			6: Infrastructural Trust Technologies		
Node	EV_r	EV_{nrm}	Node	EV_r	EV_{nrm}	Node	EV_r	EV_{nrm}
12	**0.393**	1.00	56	0.375	0.91	55	0.149	0.23
23	0.391	0.99	68	0.368	0.89	90	0.366	0.60
27	0.333	0.83	**73**	**0.411**	1.00	92	0.307	0.50
31	0.383	0.97	89	0.328	0.80	108	0.602	1.00
75	0.332	0.82	97	0.373	0.91	**109**	**0.603**	1.00

Raw eigenvector values (EV_r, three decimals) and cluster-normalized values (EV_{nrm}, two decimals). Each subbox is sorted by node number; the top EV_r per cluster is highlighted

Bold text denotes the maximum observed values

Building on these thematic premises, the analysis proceeds to RQ2, focusing on how trust is conceptualized by most influential nodes within the six identified clusters.

3.3 Conceptualizations of trust in human–AI collaboration in finance (RQ2)

To address RQ2 (*"How have research communities conceptualized trust in the most influential studies on trust in human–AI collaboration in finance?"*), this subsection reports how the five most influential studies within each of the six clusters conceptualize trust in human–AI collaboration in finance. The findings are organized by cluster, reflecting the dominant approaches observed in the most structurally influential documents (ranked by node ID and eigenvector centrality). Table 4 summarizes the top five most influential nodes in each cluster based on raw and cluster-normalized eigenvector centrality. Degree, betweenness, and closeness centrality scores (both raw and normalized) are reported in Table 12 in the appendix with the aim of highlighting variations in local connectivity, brokerage potential, and network accessibility, respectively.

Complementing these tabular results, Fig. 7 shows the bibliographic coupling network, illustrating the spatial organization of the six clusters and their internal and cross-cluster link structures.

Building on these network-level insights, the subsequent paragraphs unpack how trust is conceptualized across the six clusters (see Table 5 for further details):

Cluster 1: **AI Governance in Finance:** trust is cast as a governance outcome and is discussed in relation to competent, transparent systems that enable responsible human–AI collaboration (Khushk et al. 2024), relies on transparency and accountability as formal safeguards (Behl et al. 2023), treats algorithmic openness as the basis for managerial reliance (Chowdhury et al. 2023), equates confidence with explainable and auditable decisions aligned to oversight (Sabharwal et al.2025), and embeds assurance within ethical control structures (Lehner et al. 2022).

Cluster 2: **XAI for Finance:** trust is commonly associated with explainability that satisfies regulators and stakeholders. Transparency and fairness are emphasized (Machucho and Ortiz 2025). XAI tools reduce opacity and bolster user confidence (Saarela and Podgorelec 2019). Interpretability supports financial and strategic choices (Behera et al. 2023). Compliance and auditability anchor assurance in high-stakes

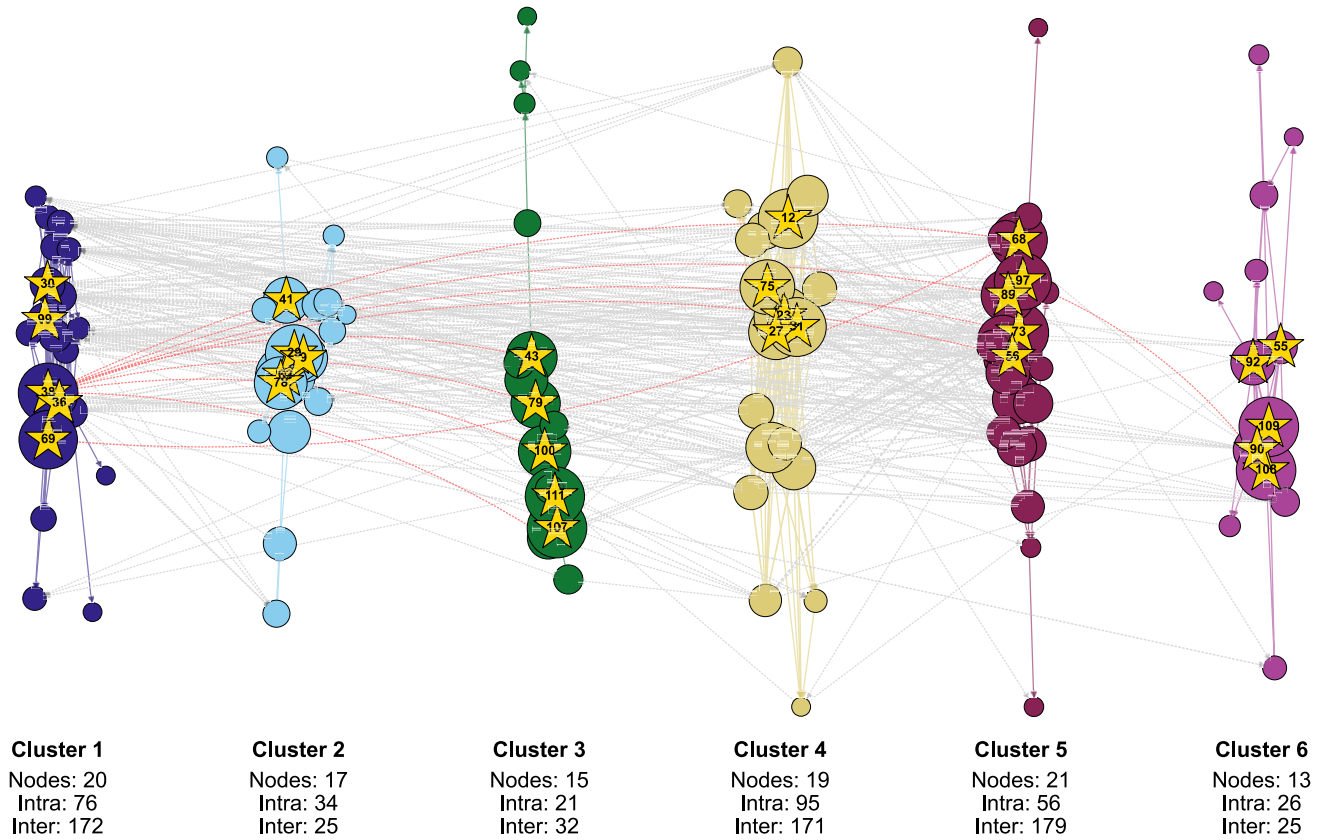

Fig. 7 Cluster-level network visualization derived from bibliographic coupling analysis, with clusters identified using the weighted Leiden algorithm. Each cluster is annotated with its top five nodes (marked with a ⋆) ranked by eigenvector centrality, highlighting the most structurally influential documents within their respective communities (indicated by a red dotted line, ⋯). Intra- and inter-cluster link densities are reported below each cluster to facilitate comparative assessment of internal cohesion and cross-cluster connectivity (colour figure online)

settings (Khan et al. 2025), and transparent, fair, auditable insurance AI sustains both stakeholder and regulatory trust (Owens et al. 2022).

Cluster 3: **Anthropomorphism in Financial AI Agents:** trust is presented as malleable and cue-driven. Apologies and human-like design help restore it after breaches (Kim and Song 2021), perceived competence and human-likeness calibrate it (Schreibelmayr et al. 2023). Some work privileges decentralized reliability over human-like features (Srinivas Aditya et al. 2021). Emotional and relational cues shape competence and integrity perceptions (Bhatia et al. 2021), and reliable algorithms tuned to investor emotions reinforce confidence in financial choices (Bhatia et al. 2020).

Cluster 4: **User Interface Design for Human–AI in Finance:** trust is framed as arising in relation to social design choices, such as warmth–competence impressions formed through interface cues (Yan et al. 2025). Chatbot attributes, including ability, empathy, benevolence, and integrity, are especially important during recovery moments (Agnihotri and Bhattacharya 2024). Usability, coupled with anthropomorphic and emotionally intelligent signals (Wang et al.), as well as aligned human-like cues that enhance social presence and perceived reliability (Chen et al. 2024), further strengthen user trust. Finally, chatbot expertise and responsiveness translate into overall brand trust (Li et al. 2023a).

Cluster 5: **Robo-Advisors for Financial Decision-Making:** trust is discussed as a linking concept between design attributes and reported outcomes. Congruence, clarity, and commitment, amplified by self-efficacy and subtle human-like cues, support user acceptance (Chen et al.

Table 5 Comparative mapping of trust conceptualizations across studies by cluster, node, and author(s) year

Cluster	Node	Author(s) year	Trust conceptualization
1	30	Khushk et al. (2024)	Rooted in competence and transparency, framed as essential to responsible human–AI collaboration
1	36	Behl et al. (2023)	Anchored in transparency and accountability as a governance mechanism ensuring responsible AI
1	**38**	**Chowdhury et al.** (2023)	**Rooted in algorithmic transparency supporting managerial reliance on AI outputs**
1	69	Sabharwal et al. (2025)	Confidence in AI decisions that remain explainable, auditable, and aligned with responsible governance
1	99	Lehner et al. (2022)	Driven by transparency, accountability, and oversight in ethical AI governance
2	9	Machucho and Ortiz (2025)	Regulator-aligned transparency and fairness
2	29	Saarela and Podgorelec (2019)	Built via XAI tools reducing opacity and boosting confidence
2	41	Behera et al. (2023)	Explainability-driven confidence in AI-supported financial and strategic decisions
2	**63**	**Khan et al.** (2025)	**Rooted in compliance and auditing within high-stakes financial AI**
2	78	Owens et al. (2022)	Transparent, fair, and auditable insurance AI ensuring stakeholder and regulatory confidence
3	43	Kim and Song (2021)	Repairable via apologies and anthropomorphic tactics post-trust breach
3	79	Schreibelmayr et al. (2023)	Calibrated by perceived competence and human-likeness shaped through anthropomorphic cues
3	100	Srinivas Aditya et al. (2021)	Rooted in decentralized reliability over anthropomorphism
3	**107**	**Bhatia et al.** (2021)	**Perceived competence and integrity shaped by human-like emotional and relational cues**
3	111	Bhatia et al. (2020)	Algorithmic reliability with human-like sensitivity to investor emotions in financial decisions
4	**12**	**Yan et al.** (2025)	**Arises from warmth and competence perceptions shaped by the design of social cues**
4	23	Agnihotri and Bhattacharya (2024)	Linked to chatbot traits: ability, empathy, benevolence, and integrity in recovery
4	27	Wang et al. ()	Links ease-of-use with relational cues (anthropomorphism and emotional intelligence)
4	31	Chen et al. (2024)	Aligned anthropomorphic cues enhancing perceived social presence and reliability
4	75	Li et al. (2023a)	chatbot expertise, anthropomorphism, and responsiveness reinforcing brand trust
5	56	Chen et al. (2025)	Stemmed from congruence, clarity, commitment; reinforced by self-efficacy, anthropomorphic cues
5	68	He and Liu (2024)	Rooted in human-like design and user traits, trust underpins engagement, well-being, and loyalty
5	**73**	**Riedel et al.** (2022)	**Affective and ideological mediator linking AI financial advice to user attitudes and behaviors**
5	89	Nourallah (2023)	Initial and culturally shaped confidence linking robo-advisor features to user adoption
5	97	Bawack et al. (2022)	Computational–relational mechanism enhancing confidence in AI-driven recommendations
6	55	Duong et al. (2025)	Confidence in blockchain transparency reinforcing consumer assurance
6	90	Yavaprabhas et al. (2023)	Defined as "trustless trust" through blockchain
6	92	Majeed Parry et al. (2024)	Built through fraud prevention and accountability in sensitive supply chains
6	108	Jellason et al. (2024)	Confidence in blockchain transparency fostering trustworthy and resilient supply chains
6	**109**	**Sri Vigna Hema and Manickavasagan** (2024)	**Distributed trust from blockchain transparency enhancing accountability**

Bold text denotes the maximum observed values

2025). Human-like design features and user dispositions drive engagement, well-being, and loyalty (He and Liu 2024). Affective and ideological responses mediate how AI advice shapes attitudes and behaviors (Riedel et al. 2022), while early confidence reflects cultural factors and feature perceptions associated with adoption (Nourallah 2023). Computational–relational mechanisms enhance confidence in AI recommendations (Bawack et al. 2022).

Cluster 6: **Infrastructural Trust Technologies:** trust is framed as infrastructural assurance via transparency, verification, and distributed

accountability. Blockchain visibility supports assurance (Duong et al.2025), "trustless trust" relies on distributed validation (Yavaprabhas et al. 2023), and fraud prevention/accountability mechanisms reinforce confidence in sensitive supply chains (Majeed Parry et al. 2024), with AI enabling monitoring and validation. Transparency and traceability are repeatedly linked to resilient networks and cross-actor accountability (Jellason et al. 2024; Sri Vigna Hema and Manickavasagan 2024).

Building on these trust conceptualizations, we now address RQ3. Specifically, we focus on the methodological approaches adopted by scholars across clusters to investigate specific trust dimensions.

3.4 Methods and tools to investigate trust dimensions in human–AI collaboration in finance (RQ3)

To address RQ3 (*"What methods and tools have been proposed in the identified most influential studies to investigate specific trust dimensions in human–AI collaboration in financial contexts?"*), this subsection reports the methodological approaches employed across the most central studies within the six identified clusters.

Table 6 and the following paragraph report the main findings from the selected studies, specifying their clusters, nodes, authors, methods, tools, and the addressed trust dimensions. Moreover, acronyms are spelled out in the caption of the table.

Cluster 1: **AI Governance in Finance:** this stream mixes evidence syntheses and empirical/technical studies: PRISMA-based SLRs (Khushk et al. 2024), cross-sectional modeling with Warp PLS-SEM (Behl et al. 2023), a supervised-ML implementation with LIME, an H2O deep classifier and precision–recall evaluation (Chowdhury et al. 2023), ADR with SHAP across classic ML techniques (Sabharwal et al.2025), and a hermeneutic narrative review coded in ATLAS.ti using Rest's ethical framework (Lehner et al. 2022). Trust is parsed into fairness, accountability, sustainability, and transparency (Behl et al. 2023); transparency/explainability with reliability and fairness (Chowdhury et al. 2023; Sabharwal et al.); cognitive/affective/integrity facets with capability, compassion and honesty (Khushk et al. 2024); and ethical–regulatory facets such as objectivity, privacy, and accountability (Lehner et al. 2022).

Cluster 2: **XAI for Finance:** most contributions are PRISMA SLRs that foreground explainability as a compliance and assurance lever (Owens et al. 2022; Saarela and Podgorelec 2019; Khan et al. 2025; Machucho and Ortiz 2025), complemented by a survey analyzed via CB-SEM in AMOS with Harman's test for bias (Behera et al. 2023). Tools concentrate on PRISMA and SEM suites, while the focal dimensions converge on transparency, explainability, and fairness, joined by reliability/trustworthiness, integrity, compliance, responsibility, and reliance in decisions (Owens et al. 2022; Behera et al. 2023; Saarela and Podgorelec 2019; Khan et al. 2025; Machucho and Ortiz 2025).

Cluster 3: **Anthropomorphism in Financial AI Agents:** methods span a 2×2 scenario experiment on apology attribution and anthropomorphism analyzed with ANOVA/ANCOVA and related post hoc/correlation tests (Kim and Song 2021); mixed scenario–survey designs with nonparametric inference and qualitative text analysis (Schreibelmayr et al. 2023); a SLR on decentralized reliability (Srinivas Aditya et al. 2021); and two phenomenological programs-expert interviews with deductive content analysis (Bhatia et al. 2021) and focus groups coded in NVivo (Bhatia et al. 2020). Dimensions traverse cognitive competence and understandability, behavioral reliance, affective responses (benevolence, uncanniness), relational human-likeness/agency, transparency/explainability, data integrity and security, plus situational and human-in-the-loop reliance factors (Bhatia et al. 2020, 2021; Kim and Song 2021; Srinivas Aditya et al. 2021; Schreibelmayr et al. 2023).

Cluster 4: **User Interface Design for Human–AI in Finance:** evidence comes from scenario experiments with mediation tests (Yan et al. 2025), dual-study scenarios with SEM in Mplus and bootstrapping (Agnihotri and Bhattacharya 2024), survey-based SEM with mediation (Wang et al.), a within-subject eye-tracking study analyzed via repeated-measures ANOVA and gaze metrics (Chen et al. 2024), and survey models in AMOS with additional regressions and reliability/validity checks (Li

Table 6 Summary of top influential papers by methods, tools, and trust dimensions

Cluster	Node	Author(s) Year	Methods	Tools	Dimensions
1	30	Khushk et al. (2024)	SLR	PRISMA	Cognitive, affective, integrity, transparency
1	36	Behl et al. (2023)	Survey; FVT, SOIPSVM	Warp PLS; PLS-SEM	Fairness, accountability, sustainability, transparency
1	**38**	**Chowdhury et al.** (2023)	**ML framework**	**LIME; H2O DL**	**Transparency, explainability, fairness, accountability, reliability**
1	69	Sabharwal et al. ()	ADR; ML experiments	SHAP	Transparency, explainability, accountability, reliance, fairness
1	99	Lehner et al. (2022)	Semi-systematic review	ATLAS.ti; Rest's ethics framework	Cognitive, privacy, transparency, integrity
2	9	Machucho and Ortiz (2025)	SLR	PRISMA	Transparency, fairness, integrity, reliance, compliance
2	29	Saarela and Podgorelec (2019)	SLR	PRISMA	Transparency, explainability, reliability, integrity
2	41	Behera et al. (2023)	Survey	CB-SEM; Harman test	Transparency, explainability, reliance, compliance
2	**63**	**Khan et al.** (2025)	**SLR**	**PRISMA**	**Transparency, explainability, fairness, accountability, compliance**
2	78	Owens et al. (2022)	SLR	PRISMA	Cognitive, relational, compliance, transparency
3	43	Kim and Song (2021)	Scenario experiment	ANOVA/ANCOVA; Duncan; Pearson	Cognitive, behavioral, reparation
3	79	Schreibelmayr et al. (2023)	Survey	WMW; Spearman	Cognitive, affective, relational, transparency
3	100	Srinivas Aditya et al. (2021)	SLR	None	Integrity, transparency, decentralized reliability
3	**107**	**Bhatia et al.** (2021)	**Expert interviews**	**Deductive/latent coding**	**Cognitive, relational, transparency, situational**
3	111	Bhatia et al. (2020)	Focus groups	NVivo	Cognitive, affective, transparency, reliance
4	**12**	**Yan et al.** (2025)	**Scenario experiments**	**Mediation analysis**	**Cognitive, affective, relational**
4	23	Agnihotri and Bhattacharya (2024)	Scenario experiments	SEM; MPLUS; bootstrap	Ability, benevolence, integrity
4	27	Wang et al. ()	Survey	SEM; mediation	Usability, affective, anthropomorphism
4	31	Chen et al. (2024)	WSE	RM-ANOVA (SPSS)	Cognitive, social presence, satisfaction
4	75	Li et al. (2023a)	Survey	SEM (AMOS); regression; reliability	Competence, interactivity, usability, anthropomorphism, brand trust, human support; risk; security
5	56	Chen et al. (2025)	Survey	PLS-SEM; NCA	Competence; communication/commitment/congruence
5	68	He and Liu (2024)	Science mapping	SciMAT; SPC	Competence, dependability, explainability
5	**73**	**Riedel et al.** (2022)	**Scenario experiments**	**PROCESS**	**Affective, ideology, reliability**
5	89	Nourallah (2023)	Survey	PLS-SEM; bootstrap; k-means	Performance expectancy, hedonic, facilitating conditions, social influence, trust propensity, culture
5	97	Bawack et al. (2022)	Bibliometric review	Louvain; RPYS	Cognition, reliance
6	55	Duong et al. (2025)	Survey	PROCESS	transparency, relational trust, integrity

Table 6 (continued)

Cluster	Node	Author(s) Year	Methods	Tools	Dimensions
6	90	Yavaprabhas et al. (2023)	SLR	Inductive/deductive coding	Cognitive, institutional, affective, temporal, trustor–trustee
6	92	Majeed Parry et al. (2024)	Narrative review	None	Transparency, traceability, authenticity, integrity, accountability, consumer trust
6	108	Jellason et al. (2024)	SLR	PEO	Transparency, traceability, accountability, integrity, governance
6	**109**	**Sri Vigna Hema and Manickavasagan** (2024)	**Comprehensive review**	**None**	**Transparency, traceability, integrity, authenticity, accountability**

ADR Action Design Research, *AMOS* Analysis of Moment Structures, *ANCOVA* Analysis of Covariance, *ANOVA* Analysis of Variance, *CB-SEM* Covariance-Based Structural Equation Modeling, *FVT* Face Validity Test, *H2O DL* H2O Deep Learning, *LIME* Local Interpretable Model-agnostic Explanations, *ML* Machine Learning, *NCA* Necessary Condition Analysis, *PEO* Population–Exposure–Outcome, *PLS-SEM* Partial Least Squares Structural Equation Modeling, *PRISMA* Preferred Reporting Items for Systematic Reviews and Meta-Analyses, *RPYS* Reference Publication Year Spectroscopy, *SEM* Structural Equation Modeling, *SHAP* SHapley Additive exPlanations, *SLR* Systematic Literature Review, *SOIPSVM* Selective Organizational Information Privacy and Security Violations Model, *SPC* Search Path Count, *Warp PLS* Warp Partial Least Squares, *WMW* Wilcoxon–Mann–Whitney test, *WSE* Within-Subject Experiment, *RM-ANOVA* Repeated-Measures ANOVA

Bold text denotes the maximum observed values

et al. 2023a). Reported dimensions emphasize warmth–competence and relational trust (Yan et al. 2025), ability–benevolence–integrity in service recovery (Agnihotri and Bhattacharya 2024), usability and anthropomorphic/emotional cues (Wang et al.), social presence and satisfaction alongside cognitive trust (Chen et al. 2024), responsiveness, ease, anthropomorphism, brand trust, reliance on human support, risk, and privacy/security moderation (Li et al. 2023a).

Cluster 5:
Robo-Advisors for Financial Decision-Making: approaches include scenario surveys with PLS-SEM and NCA (Chen et al. 2025), bibliometric main-path/science mapping using SciMAT (He and Liu 2024), serial-mediation experiments with PROCESS (Riedel et al. 2022), a cross-national PLS-SEM study with bootstrapping and k-means segments (Nourallah 2023), and a bibliometric review leveraging Bibliometrix, Louvain clustering, and RPYS (Bawack et al. 2022). Trust is broken down into cognitive competence with relational communication/commitment/congruence (Chen et al. 2025), dependability and explainability (He and Liu 2024), emotion-based and ideological mediators with perceived reliability (Riedel et al. 2022), adoption-related cognitions (performance/effort), hedonic and situational supports, social/dispositional tendencies, and cultural orientations (Nourallah 2023), plus core cognition–reliance linkages (Bawack et al. 2022).

Cluster 6: **Infrastructural Trust Technologies:** methods range from surveys modeled with PROCESS (Duong et al.2025) to SLRs with inductive/deductive coding (Yavaprabhas et al. 2023), narrative/case syntheses (Majeed Parry et al. 2024), and PEO-guided thematic SLRs (Jellason et al. 2024), alongside a comprehensive review (Sri Vigna Hema and Manickavasagan 2024). The mapped dimensions consistently spotlight transparency and traceability, extended by authenticity, integrity/data immutability, accountability/compliance, provenance governance, temporal (swift vs. long-term), and multi-perspective trust facets spanning cognitive, institutional and affective layers (Yavaprabhas et al. 2023; Duong et al.2025 ; Majeed Parry et al. 2024; Jellason et al. 2024; Sri Vigna Hema and Manickavasagan 2024).

Having established the methods, tools, and trust dimensions, we now proceed to RQ4, which explores the key emerging trends and gaps that may guide future research.

3.5 Emerging themes and research gaps in trust in human–AI collaboration in finance (RQ4)

To address RQ4 (*"What emerging themes and research gaps exist in research on trust in human–AI collaboration in finance across identified clusters?"*), this subsection reports the results of the full-text review of the five most influential

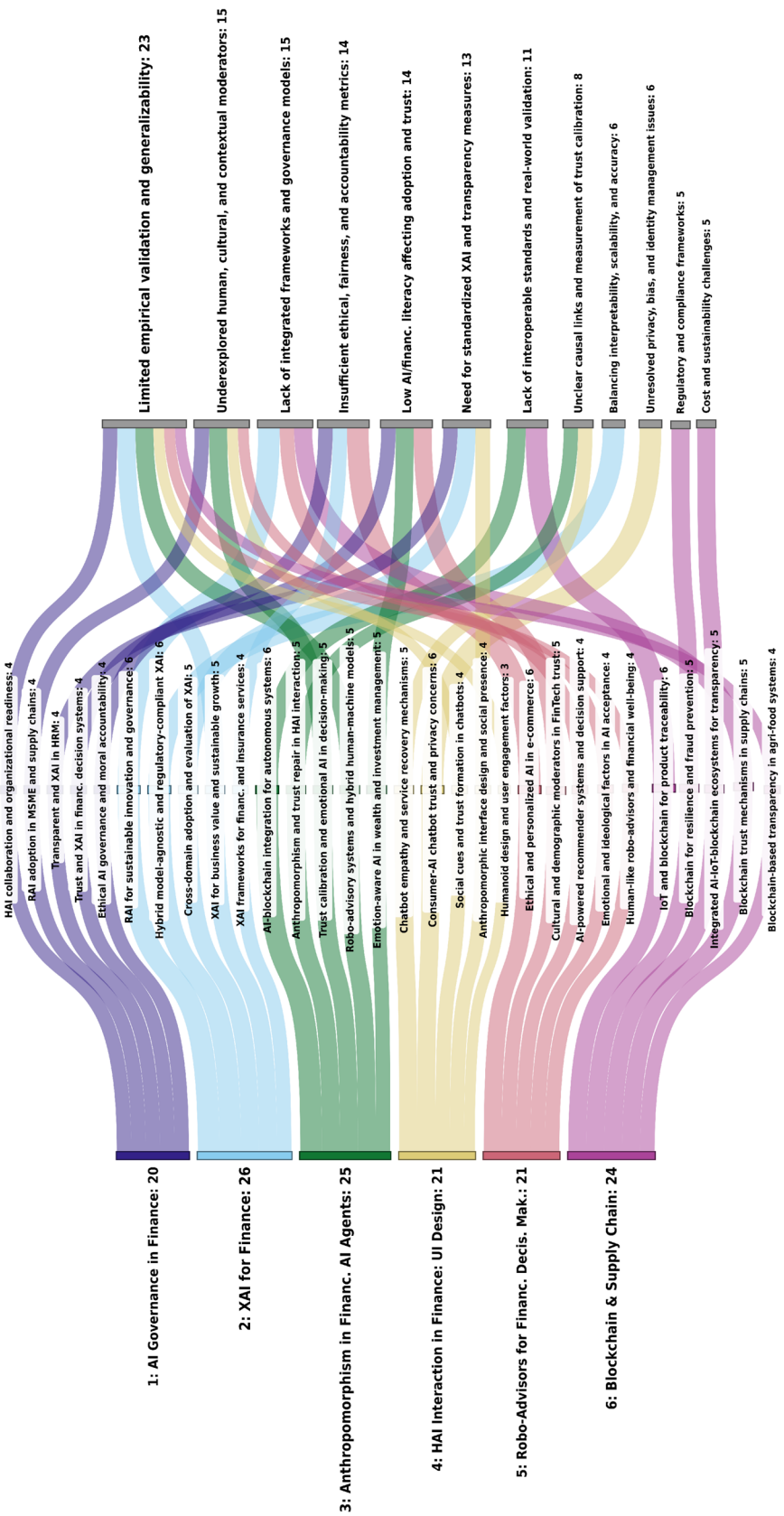

Fig. 8 Sankey diagram showing the distribution and interconnection of research themes and gaps across the six identified clusters. Clusters are shown on the left, emerging themes in the center, and corresponding research gaps on the right. Numbers beside each node indicate their relative frequency, while flow thickness (colored by cluster) represents the strength of thematic linkage

papers within each of the six identified clusters. The aim is to map the emerging themes and the corresponding research gaps identified by the authors related to trust in human–AI collaboration in finance. The resulting visualization (Fig. 8) synthesizes the overall structure of the analysis. Each cluster is positioned on the left, followed by the derived themes in the center and the associated research gaps on the right. The numbers adjacent to each node indicate the relative frequency of occurrence of that element within the corpus, while the width of each colored flow (by cluster) reflects the intensity of connection between clusters, themes, and gaps.

The identified emerging themes and related research gaps are summarized as follows:

Cluster 1: **AI Governance in Finance:** themes span organizational human–AI interaction, transparent financial systems, XAI-based governance, and ethical decision-making in accounting. Though, gaps center on stronger empirical tests, moderators of trust and team dynamics, transition frameworks to transparent AI, AI literacy, and models for human–machine moral collaboration (Lehner et al. 2022; Behl et al. 2023; Chowdhury et al. 2023; Khushk et al. 2024; Sabharwal et al.2025).

Cluster 2: **XAI for Finance:** work highlights XAI for business model innovation, cross-domain XAI, information system value and strategy, and insurance compliance/trust. However, missing are standardized metrics/benchmarks, longitudinal and cross-domain validations, causal evidence for growth advantage, as well as an agreed balance between interpretability, scalability and accuracy (Owens et al. 2022; Behera et al. 2023; Saarela and Podgorelec 2019; Khan et al. 2025; Machucho and Ortiz 2025).

Cluster 3: **Anthropomorphism in Financial AI Agents:** themes address trust repair processes, calibration via transparency/emotion, decentralized reliability (AI–blockchain), and hybrid robo-advisory models with emotion-aware decisions. Nevertheless, gaps involve precise anthropomorphism dimensions and boundary conditions, validated multi-item trust measures, real-world testbeds/standards for blockchain-based robotics, quantitative validation, richer data integration, and larger/diverse samples (Bhatia et al. 2020, 2021; Kim and Song 2021; Srinivas Aditya et al. 2021; Schreibelmayr et al. 2023).

Cluster 4: **User Interface Design for Human–AI in Finance:** emerging themes focus on social-cue design, warmth–competence pathways, humanoid traits (ease, emotional intelligence, personalization), visual–conversational interplay, and multi-factor chatbot trust (privacy, risk, brand, human support). But, needed are longitudinal and cross-cultural tests, broader cue taxonomies, guidance for empathetic design under failure, field methods beyond labs, and mapping trust across exposure–adoption with clarity on human support vs. autonomy (Li et al. 2023a; Agnihotri and Bhattacharya 2024; Chen et al. 2024; Wang et al. ; Yan et al. 2025).

Cluster 5: **Robo-Advisors for Financial Decision-Making:** themes include human-like features and self-efficacy as trust antecedents, optimization/personalization and sentiment, affect–attitude mechanisms, cross-cultural initial trust and Unified Theory of Acceptance and Use of Technology (UTAUT), and recommenders. However, gaps point to longitudinal trust–loyalty dynamics, cultural/demographic moderators, real-world validation of humanoid cues and recommenders, ethical and sustainable XAI adoption, discrete emotions and ideology effects, multi-country panels, and voice-assistant support (Bawack et al. 2022; Riedel et al. 2022; Nourallah 2023; He and Liu 2024; Chen et al. 2025).

Cluster 6: **Infrastructural Trust Technologies:** this cluster emphasizes blockchain-enabled traceability and transparency in food and wine supply chains, often combined with IoT and smart contracts, where AI is mainly positioned as an enabling layer for monitoring, verification, and anomaly detection. Key gaps include limited cross-cultural and longitudinal validation, unresolved tensions between "trustless" and trust-enabling framings, challenges around standards, interoperability, and scalability, governance and compliance cost barriers, energy impacts, unclear stakeholder roles and accountability, and insufficient long-term safety and efficiency assessments (Yavaprabhas et al. 2023; Duong et al. ; Jellason et al. 2024; Majeed Parry et al. 2024; Sri Vigna Hema and Manickavasagan 2024).

4 Discussion

4.1 Discussion of results

Guided by the results reported in Sect. 3, we next discuss RQ1–RQ4, and in doing so answer RQ5 through a novel multi-level socio-technical framework with four financial use cases and grounded propositions.

First, bibliographic coupling and distinctiveness analysis revealed six interrelated thematic clusters structuring the research landscape of trust in human–AI collaboration in finance: i) AI Governance in Finance; ii) XAI for Finance; iii) Anthropomorphism in Financial AI Agents; iv) User Interface Design for Human–AI in Finance; v) Robo-Advisors for Financial Decision-Making; and vi) Infrastructural Trust Technologies. These clusters highlight how trust acts as an integrative bridge between micro-level (user perceptions and calibration), meso-level (organizational design and corporate AI governance), and macro (regulatory and infrastructural). Also, this structure indicates present but limited cross-fertilization among clusters.

Second, across clusters, trust is conceptualized through complementary yet fragmented lenses. In the governance and XAI clusters, trust is predominantly cognitive and procedural, anchored in transparency, accountability, and explainability as governance levers for assurance (Khushk et al. 2024; Sabharwal et al. ; Khan et al. 2025; Machucho and Ortiz 2025). In anthropomorphism and interaction clusters, trust takes a socio-affective form, shaped by warmth–competence perceptions, empathy, and anthropomorphic cues that enhance social presence and perceived reliability (Bhatia et al. 2021; Agnihotri and Bhattacharya 2024; Yan et al. 2025). In robo-advisor research, trust functions as a behavioral mediator linking design attributes and user adoption, balancing confidence and overreliance (Riedel et al. 2022; Chen et al. 2025). Finally, infrastructural trust studies reframe trust as embedded in transparency, traceability, and decentralized verification (Sri Vigna Hema and Manickavasagan 2024).

Third, methodologically, survey-based SEM dominates the literature, especially within the governance, XAI, interaction, and robo-advisor clusters (Riedel et al. 2022; Behl et al. 2023; Wang et al.2025). Scenario experiments, qualitative interviews, and eye-tracking studies provide complementary but less frequent evidence, often in user-interface and anthropomorphism contexts (Bhatia et al. 2021; Yan et al. 2025; Kim and Song 2021). Technical research integrates XAI tools such as SHAP and LIME for model interpretability (Chowdhury et al. 2023) while blockchain research employs thematic reviews and process frameworks focused on transparency and accountability (Yavaprabhas et al. 2023; Jellason et al. 2024). The methodological landscape remains cross-sectional, with limited triangulation across behavioral and technical evidence.

Fourth, across clusters, five broad research frontiers emerge:

1. Governance and ethics research calls for frameworks translating responsible AI principles into auditable practice (Behl et al. 2023; Khushk et al. 2024).
2. XAI requires standardized metrics, longitudinal validation, and scalable model-agnostic tools (Khan et al. 2025; Machucho and Ortiz 2025).
3. Human–AI interaction and anthropomorphism need empirically validated cue taxonomies, longitudinal and cross-cultural testing, and real-world deployments (Kim and Song 2021; Yan et al. 2025).
4. Robo-advisor studies demand examination of cultural and emotional moderators of trust and long-term user loyalty (Riedel et al. 2022; Nourallah 2023; Chen et al. 2025).
5. Infrastructural trust research should address the challenges of interoperability, sustainability, and governance, particularly the trade-offs between trustless and trust-enabling architectures (Duong et al. 2025; Jellason et al. 2024).

Across the six clusters, our bibliometric and qualitative synthesis reveals a consistent multi-level pattern in how trust is shaped within human–AI collaboration in finance. Cluster 1 (AI Governance in Finance) and Cluster 2 (XAI for Finance) predominantly engage meso–macro mechanisms, emphasizing organizational governance practices, regulatory alignment, and system-level assurance structures. Cluster 3 (Anthropomorphism in Financial AI Agents), Cluster 4 (User Interface Design for Human–AI in Finance), and Cluster 5 (Robo-Advisors for Financial Decision-Making) operate chiefly at the micro–meso interface, centering on individual cognition, affect, social presence, and how organizational design choices shape user experience. Cluster 6 (Infrastructural Trust Technologies), by contrast, is rooted in macro-level trust infrastructures, where verification, traceability, and decentralized validation mechanisms reframe trust as a system-level assurance rather than an individual evaluative judgment. Within this cluster, trust is discussed as being embedded in technological and institutional arrangements, with AI serving a role in supporting monitoring and validation across distributed infrastructures. These themes show that trust is conceptualized across distinct micro-, meso-, and macro-level dimensions depending on the domain, thereby providing the empirical grounding for the integrative socio-technical framework introduced next. These patterns also inform the micro-, meso-, macro-, and cross-level propositions articulated within that framework.

4.2 The multi-level socio-technical framework of trust: analytical structure and testable propositions (RQ5)

Across the four research questions above, this study conceptualizes trust in human–AI collaboration in finance as dynamic, layered, and emergent across interconnected micro-, meso-, and macro-levels. The MMM framework is descriptive and integrative, capturing how trust is conceptualized in the literature rather than specifying deterministic causal mechanisms or predicting outcomes. This framing draws on complex adaptive systems theory, which explains how local interactions among heterogeneous agents can generate system-level coherence over time (Holland 1992). In this view, trust arises through recursive feedback loops linking individual experience, organizational and technical design, and institutional and infrastructural arrangements, rather than from any single technical or psychological factor in isolation. The framework further extends complexity-based perspectives on cooperation and coordination to mixed human–AI settings, where participation, commitment, and trust evolve through adaptive interaction processes embedded within changing technical and governance environments (Lee et al. 2019; Dafoe et al. 2020; Reuel et al. 2025).

To render the MMM framework analytically tractable, we derive a set of level-specific propositions grounded in recurring trust mechanisms (see Tables 7, 8, 9 and 10).

At the micro-level, trust is grounded in users' subjective perceptions of AI-supported decision-making, shaped by assessments of competence, reliability, fairness, and decision authority, as well as by affective responses such as comfort, reassurance, or perceived social presence. These dynamics are particularly salient in social and advisory financial applications, including robo-advisory and conversational agents. This pattern aligns with socio-cognitive models of trust and the CASA paradigm (Nass et al. 1994; Castelfranchi and Falcone 2010; Pelau et al. 2021), which emphasize that trust is co-produced through analytical evaluation and affective experience. Across thematic clusters, however, the literature consistently distinguishes trust from observable reliance: high trust does not necessarily translate into improved decision quality when confidence exceeds actual system capability or when decision authority boundaries are unclear, highlighting the importance of calibration rather than trust maximization. These dynamics motivate the micro-level propositions in Table 7.

At the meso-level, trust becomes embedded in organizational procedures, interface design choices, and corporate governance arrangements that structure how humans and AI collaborate in practice. Explainability tools (e.g., SHAP, LIME), workflow design, and escalation mechanisms operationalize transparency, understandability, and recoverability, translating internal development practices and control structures into user-facing assurances. From this perspective, transparency is not a purely technical attribute but a governance-mediated property whose trust effects depend on institutional embedding rather than informational disclosure alone. This perspective is formalized in the meso-level propositions in Table 8.

At the macro-level, trust is framed as an institutional and infrastructural property shaped by regulatory regimes, standards, and verification technologies. Regulatory systems encode expectations regarding accountability, auditability, and responsibility attribution, while infrastructural mechanisms, such as standardized metrics, audits, assurance

Table 7 Micro-level propositions on perception and trust calibration in human–AI collaboration in finance

Proposition	Statement
P1: Trust-reliance calibration	Higher subjective trust in financial AI systems does not predict higher quality decision outcomes unless user trust is calibrated to the system's actual competence, uncertainty, and decision authority
P2: Anthropomorphism and overreliance risk	Anthropomorphic design cues (e.g., conversational tone or human-like avatars) increase initial trust in financial AI systems, but also increase the likelihood of overreliance when organizational interfaces fail to surface model uncertainty, scope limits, or escalation pathways

Table 8 Meso-level propositions on organizational design and corporate AI governance in finance

Proposition	Statement
P3: Governance-moderated transparency	In high-stakes financial decision contexts, transparency and explainability features increase user trust only when embedded within governance mechanisms that ensure auditability, accountability, and enforceable oversight
P4: Corporate governance maturity and reliance discipline	Organizations with mature corporate AI governance arrangements (e.g., human-in-the-loop controls, escalation protocols, and model risk oversight) exhibit lower variance in human reliance behavior than organizations with comparable technical performance but weaker governance structures

regimes, or distributed verification technologies, can embed trust into shared institutional processes that reduce reliance on individual judgment. In this interpretative layer, trust is less a purely subjective belief and more a property of the socio-technical environment itself. This interpretation is formalized in the macro-level propositions presented in Table 9.

Within the MMM framework, cross-level dynamics are analytically distinct and interact dynamically rather than hierarchically, as illustrated in Fig. 9. Individual trust perceptions and reliance behaviors emerge within organizationally structured interaction environments and institutionally conditioned expectations, while simultaneously feeding back into organizational design choices, governance practices, and legitimacy assessments. Meso-level arrangements, such as interface design, explainability practices, and corporate governance structures, both shape and are reshaped by user trust calibration and by macro-level regulatory, normative, and infrastructural constraints. At the same time, macro-level standards, regulatory regimes, and trust infrastructures directly condition individual trust expectations and organizational strategies, even as they evolve in response to aggregated micro-level experiences and meso-level governance practices. Trust thus, emerges as a distributed, co-produced property of the socio-technical ecosystem, continuously reconstituted through reciprocal interactions across analytical levels rather than stabilized at any single locus.

Trust dynamics also unfold across levels over time, particularly in response to failures, shocks, or large-scale adoption patterns. Trust repair, institutional accountability, and infrastructural design determine whether disruptions remain localized or propagate systemically across the financial ecosystem. These temporal and cross-level dynamics are formalized in the propositions presented in Table 10.

To exemplify this dynamic interplay, Table 11 outlines four finance-oriented contexts, i.e., wire transfer verification,

Table 9 Macro-level propositions on regulation and institutional trust in financial AI systems

Proposition	Statement
P5: Institutional embedding of trust	In regulated financial settings, trust in human–AI collaboration is primarily shaped by confidence in institutional and infrastructural governance mechanisms (e.g., audits, assurance regimes, regulatory oversight), which condition how individual-level trust in AI systems translates into decision acceptance
P6: Standardization as a condition for regulatory trust	Sustained regulatory trust in AI systems in finance depends on the availability of standardized benchmarks for explainability, validation, and governance compliance, rather than on firm-specific transparency claims alone

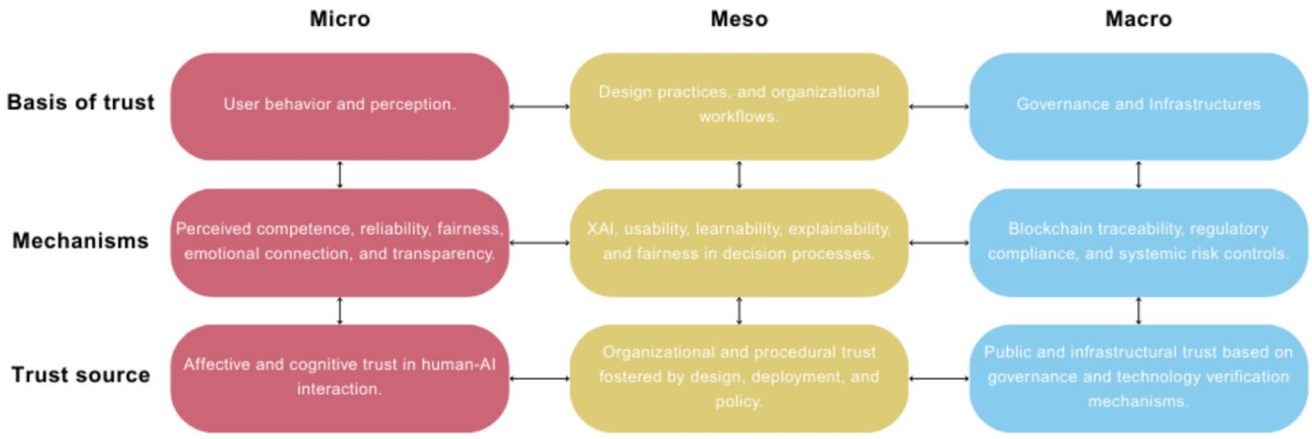

Fig. 9 Basis of trust, mechanisms, and trust sources in the micro–meso–macro (MMM) socio-technical framework of trust in human–AI collaboration in finance. Columns represent analytical levels (micro ↔ meso ↔ macro), while rows denote thematic dimensions. Horizontal and vertical bidirectional arrows indicate reciprocal influences, positioned within the gaps between cells for clarity

Table 10 Cross-level propositions on trust dynamics, repair, and systemic risk in AI in finance

Proposition	Statement
P7: Trust repair through responsibility attribution	Following AI-related failures, trust recovery is faster and more durable when responsibility attribution across human decision-makers, AI systems, and institutions is explicit and contestable rather than diffuse or ambiguous
P8: Systemic trust fragility under correlated adoption	As financial institutions converge on similar AI models, vendors, or governance templates, trust becomes more systemically fragile, increasing the risk that localized failures propagate into sector-wide trust breakdowns

financial education, stock market participation, and financial reporting, illustrating how trust is distributed across micro-, meso-, and macro-mechanisms rather than localized in any single actor or technology. Figure 10 provides a simplified visual synthesis of these interactions.

To delimit the scope of the framework and enhance its analytical clarity, Table 13 in the Appendix operationalizes the proposed multi-level socio-technical framework by specifying its domains of application, explicating the mechanisms linking micro-, meso-, and macro-level trust processes, and outlining indicative, empirically observable measures at each level.

4.3 Trust as a corporate governance problem

The MMM framework introduced above highlights that trust in human–AI collaboration in finance is discussed in the literature in relation to interactions among individual perceptions, organizational practices, and institutional infrastructures. Within this configuration, corporate governance occupies a central meso-level position in how trust is framed and organized. In particular, trust is not discussed solely in relation to the technical properties of AI systems or to individual user attitudes, but rather in relation to how AI-enabled decision-making is situated within governance structures that allocate responsibility, enable oversight, and provide assurance (Hilb 2020; Papagiannidis et al. 2023; Sharma 2024).

This interpretation aligns with governance-oriented research that conceptualizes AI not merely as a passive decision-support tool, but as an actor-like participant within organizational decision processes. As AI systems are increasingly involved in corporate activities (e.g., investment advice, financial reporting, and compliance), they are discussed as materially relevant to outcomes affecting multiple stakeholders. In this context, AI-supported decision-making is associated with reconfigurations of informational asymmetries within organizations and with increased complexity in traditional lines of fiduciary accountability, particularly

Table 11 Four examples of dynamic interactions across micro-, meso-, and macro-levels in financial contexts

Context	Micro	Meso	Macro	Dynamic interactions
Verification of Payee (VoP)	Payers verify that beneficiary names match account identifiers before confirming transfers	Payment providers perform real-time name verification between payer and payee institutions	The European Payments Council defines common rules to enhance security and trust in credit transfers	Coordination among users, providers, and scheme governance reinforces reliability and confidence in payments
Financial education	Individuals seek to understand and manage financial choices	Organizations use digital tools and adaptive learning systems to deliver guidance	National strategies and public policies promote inclusion and literacy	Feedback between learners, providers, and policy design improves effectiveness and trust
Stock market participation	Investors interpret market signals and assess risk	Trading platforms employ algorithmic analytics and advisory systems	Regulators and exchanges oversee standards, transparency, and stability	Investor actions influence platform design and regulatory responses over time
Financial reporting	Users interpret and rely on disclosed financial data	Firms employ automated reporting and analysis systems	Standards and oversight bodies ensure consistency and accountability	Mutual adaptation between users, firms, and regulators sustains transparency and trust

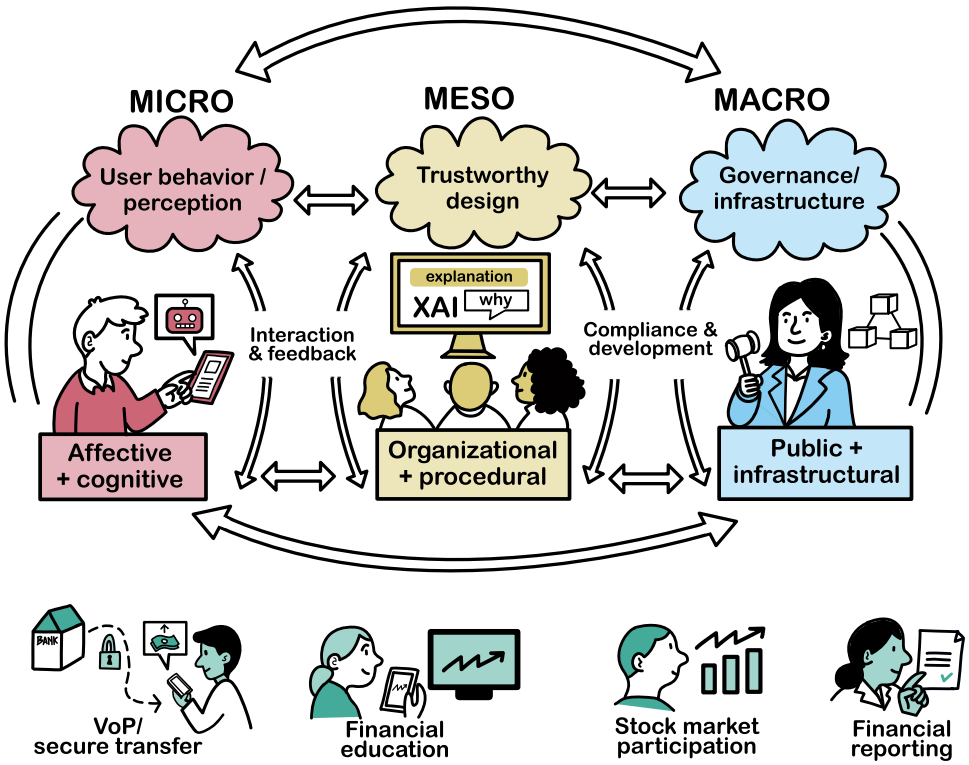

Fig. 10 Visual sketch of the MMM socio-technical framework of trust in human–AI collaboration in finance. The illustration simplifies the conceptual relationships between micro-, meso-, and macro-level dynamics, emphasizing feedback loops and adaptive interactions across individuals, organizations, and institutional infrastructures. It serves as an accessible visual complement to the theoretical model presented in Fig. 9 and to the examples provided in Table 11

in regulated financial settings (Hilb 2020; Ustahaliloğlu 2025). Within this perspective, AI systems are increasingly described as functioning as stakeholding or actor-like participants within corporate governance arrangements, even though they lack legal personhood and moral agency. Across the literature, AI is consistently framed as non-moral yet governance-relevant: while algorithms cannot bear responsibility in a philosophical or legal sense, their outputs materially affect stakeholder interests, redistribute informational power, and reshape organizational risk exposure (Véliz 2021; Camilleri 2024). This distinction is critical, as it prevents the misattribution of moral agency to AI while simultaneously acknowledging its capacity to exert structural influence over decision-making processes.

Agentic AI intensifies these concerns by introducing autonomous, adaptive, and coordinated decision-making. In contrast to static or rule-bound automation, such systems diffuse accountability across interacting components, endogenize systemic risk, and shift governance attention away from moral attribution toward the design of behavioral constraints and oversight mechanisms (Aldridge et al. 2025). Further governance-oriented contributions sharpen this interpretation. Moro-Visconti (2025) frames AI as an autonomous governance agent whose defining role is to mediate, accelerate, and reconfigure interactions among existing human stakeholders by occupying a high-centrality interlayer position within multilayer governance networks. In this configuration, AI dynamically links boards, managers, regulators, and financial institutions by accelerating information flows, reducing governance latency, and intensifying monitoring capacity. In this role, AI does not replace human decision-makers but reconfigures the conditions under which governance is exercised, amplifying both the effectiveness of oversight and the consequences of governance failure. Further research similarly emphasizes that AI systems possess functional agency in the sense that they guide, steer, or constrain human judgment, often without allowing meaningful intervention at the point of decision (Miller 2022). This capacity to shape outcomes, particularly for passive or weakly represented stakeholders, elevates AI from a neutral tool to a governance-relevant actor whose influence must be institutionally managed. Conversely, Véliz (2021) explicitly rejects the notion that algorithms can be moral agents, arguing that their lack of sentience precludes responsibility, intentionality, or ethical understanding. Rather than weakening the governance argument, this position strengthens it: if AI cannot be morally accountable, then trust cannot rationally be placed in the system itself, but must instead be grounded in the institutional arrangements that supervise, constrain, and assume responsibility for AI use. Legal and governance scholarship echoes this logic by showing that trust deficits emerge most acutely where responsibility attribution is ambiguous, oversight is fragmented, or accountability chains are poorly specified (Armour and Eidenmüller 2019; Camilleri 2024). Accordingly, trust is consistently conceptualized not as a psychological belief

in AI's benevolence or competence, but as a governance-structural property. Stakeholders are described as needing to trust that AI-supported decisions are embedded within robust governance architectures, encompassing board oversight, auditable processes, stakeholder representation, and escalation mechanisms, capable of rendering outcomes contestable and accountable ex post (Moro-Visconti 2025; Papagiannidis et al. 2025). In this framing, AI's growing actor-like role intensifies, rather than diminishes, the importance of corporate governance: trust in human–AI collaboration ultimately depends less on the system's technical sophistication than on the institutional capacity to govern an influential but non-moral decision-making participant.

The patterns identified in Cluster 1 (AI Governance in Finance) and Cluster 2 (XAI for Finance) are consistent with this interpretation. Across these clusters, trust is repeatedly associated with accountability, auditability, compliance, and alignment with organizational and regulatory expectations, rather than with model accuracy alone. Explainability and transparency are treated less as ends in themselves than as elements discussed in relation to governance processes, such as model validation, internal audit, and regulatory reporting (Sharma 2024). Conversely, in contexts where organizational AI governance remains underdeveloped and responsibility for AI-enabled decisions is insufficiently specified, trust-related risks are more accurately framed in terms of over-reliance on automated outputs, uncritical automation acceptance, and ex post diffusion of accountability, rather than as mere deficits in user understanding (Camilleri 2024).

Integrating this governance lens into the multi-level framework clarifies the interpretive role of each analytical level. At the micro-level, trust concerns how clients, employees, and decision-makers calibrate their reliance on AI-supported outputs, balancing confidence with appropriate skepticism. At the macro-level, trust is discussed in relation to regulatory regimes, standards, and infrastructural arrangements, including supervisory expectations and compliance architectures (Hickman and Petrin 2021). The meso-level of corporate governance is described as a translation layer between these domains, where regulatory norms are converted into internal controls and where individual trust experiences are discussed in relation to organizational design choices (Hilb 2020; Sharma 2024).

Across the literature, three recurring governance-trust linkages (L1, L2, and L3) can be identified, which collectively support a structural interpretation of trust as a governance-mediated outcome rather than a purely psychological or technical attribute. First, accountability and auditability (L1) are consistently grounded in formal governance practices that specify roles, responsibilities, and decision rights across the AI lifecycle, supported by documentation, traceability, and monitoring mechanisms that enable ex post scrutiny of AI-enabled outcomes (Armour and Eidenmüller 2019; Camilleri 2024; Papagiannidis et al. 2025). In both ethical-governance and corporate-law frameworks, accountability is not treated as a property of AI systems themselves, but as an organizational obligation requiring board-level oversight, auditable processes, and governance structures capable of demonstrating compliance with legal, ethical, and fiduciary standards. Sharma (2024) reinforces this position by explicitly framing AI governance as a board and C-suite responsibility, emphasizing that trust in AI-enabled decisions depends on clearly assigned ownership, lifecycle governance, and the ability of directors to evidence oversight through structured policies, monitoring, and assurance mechanisms.

Second, overreliance and miscalibrated trust (L2) are repeatedly associated with weak governance safeguards, particularly the absence of meaningful human agency, escalation protocols, and procedural controls during deployment and use. In the board governance context, Ahdadou et al. (2024) explicitly identify overtrust and overreliance as central risks of augmented intelligence when AI-supported decision-making expands without corresponding mechanisms for human oversight and accountability. Papagiannidis et al. (2025) show that when responsible AI principles remain decoupled from structural, procedural, and relational governance practices, organizations risk normalizing uncritical automation and moral disengagement. From a corporate-law perspective, Armour and Eidenmüller (2019) conceptualize this risk as algorithmic failure, whereby AI-enabled decisions generate harmful or unlawful outcomes without clear points of human responsibility. Sharma (2024) complements this analysis by warning that organizations frequently overinvest in technical capability while underinvesting in governance maturity, creating conditions in which AI systems are operationalized without sufficient human-in-the-loop controls, escalation pathways, or accountability mapping, thereby amplifying the risk of blind reliance rather than calibrated trust. Third, transparency and explainability (L3) acquire trust-relevant meaning only when embedded in assurance-oriented governance practices, such as systematic documentation, logging, validation, and disclosure mechanisms. Across the reviewed literature, transparency is not framed as informational openness to end users alone, but as an institutional capability that enables auditability, responsibility attribution, and regulatory scrutiny (Camilleri 2024; Papagiannidis et al. 2025). Sharma (2024) explicitly aligns transparency with board-level assurance, arguing that explainability supports trust only insofar as it is operationalized through governance artifacts, such as model documentation, risk assessments, and reporting structures, that allow boards, regulators, and other stakeholders to evaluate AI behavior over time rather than relying on ad hoc explanations.

Accordingly, this governance-oriented interpretation is operationalized in the Supplementary Materials, where we operationalize this governance perspective by translating abstract trust concepts into concrete organizational responsibilities and auditable practices: for each trust mechanism identified, it specifies the governance body accountable for its oversight and the evidentiary artifacts through which trust is demonstrated, monitored, and contested in practice (e.g., validation reports, audit trails, escalation logs, and board-level documentation).

5 Concluding remarks and future work

5.1 Main conclusions

This study presents, to the best of our knowledge, the first Bibliometric–Systematic Literature Review of trust in human–AI collaboration in finance, synthesizing 114 studies published between 2018 and 2025. The evidence reveals a rapidly expanding yet conceptually fragmented field structured into six clusters, in which trust consistently functions as a multidimensional bridge linking socio-cognitive mechanisms, technical properties and organizational practices, and institutional and infrastructural assurances. Despite notable advances in XAI and adoption research, integration across trust conceptualizations remains limited, constraining cumulative theory building and practical translation in high-stakes financial settings.

To address this fragmentation, the review develops a multi-level socio-technical framework that organizes trust mechanisms across micro-level user perceptions and calibration, meso-level organizational design and corporate governance arrangements, and macro-level regulatory and infrastructural conditions. Importantly, the framework goes beyond descriptive consolidation by articulating analytically distinct and testable propositions that specify conditions under which trust becomes calibrated, misaligned, repaired, or institutionally sustained in human–AI collaboration in finance. In doing so, the review provides a structured foundation for future experimental, longitudinal, and field-based research and offers an actionable analytical basis for the design, deployment, and governance of trust-aware, human-, corporate-, and societally aligned AI in finance.

Within this framework, trust is explicitly conceptualized as a latent evaluative belief held by a human trustor toward an AI system or AI-enabled socio-technical arrangement under conditions of uncertainty. Trust reflects expectations regarding competence, integrity, and role-conformity, including alignment with fiduciary, organizational, and regulatory obligations. This definition analytically distinguishes trust from observable behaviors such as reliance or adoption, and from system capabilities themselves, which are treated as antecedents or outcomes rather than trust per se. Clarifying the object and basis of trust provides a coherent conceptual anchor for integrating the fragmented literature and for interpreting empirical findings across levels of analysis.

5.2 Contributions to theory and practice

Building on this synthesis, the review advances the literature in three main ways. First, it provides, to the best of our knowledge, the first Bibliometric–Systematic Literature Review of trust in human–AI collaboration in finance. By combining bibliometric mapping with systematic qualitative analysis, the study offers a comprehensive and reproducible overview of the intellectual structure, dominant streams, and emerging themes in the field.

Second, the review clarifies how trust is conceptualized and empirically studied across fragmented research streams. It identifies dominant methodological approaches, recurring conceptual ambiguities, and persistent blind spots, thereby highlighting where existing research converges and where it remains theoretically underdeveloped. This clarification is essential for cumulative theory building.

Third, the study introduces the MMM multi-level socio-technical framework, which explains how trust in human–AI collaboration is formed, calibrated, and sustained. By systematically linking micro-level trust perceptions to meso-level design practices and governance arrangements, and further to macro-level technological and regulatory assurances, the framework integrates previously fragmented perspectives and offers a coherent analytical lens for researchers, institutions, and practitioners concerned with trust-aware human–AI collaboration in financial contexts.

5.3 Limitations and boundary conditions

Although we have reported results from a methodologically solid and grounded Bibliometric–Systematic Literature Review, several boundary conditions must be acknowledged.

First, the retrieval is Scopus-based and time-bounded. Thus, this review paper relies exclusively on Scopus as the source database, and as a result, relevant contributions indexed in other databases or preprint repositories may be omitted. This choice is appropriate for a first bibliometric–systematic review, given Scopus's broad multidisciplinary coverage and standardized metadata, which support reproducible bibliometric analyses. However, it necessarily excludes preprints (e.g., arXiv, SSRN), legal and policy documents, and practitioner or industry reports, potentially underrepresenting fast-moving technical and regulatory debates. This reflects a deliberate boundary choice to establish a coherent, peer-reviewed academic baseline and highlights opportunities for future extensions using complementary sources. Furthermore, backward and forward

snowballing was not conducted, which may limit the inclusion of relevant literature. Additionally, potential conceptual, cultural, or linguistic biases may arise from the specific query strategy employed. While the primary retrieval focused on *trust* as the focal construct, we systematically examined closely related alternative terms (i.e., acceptance, adoption, assurance, automation bias, automation complacency, confidence, reliance, safety), which are frequently used in adjacent literature to describe human–AI interactions. As documented in the Supplementary Materials, these terms were deliberately excluded from the primary query because they capture distinct phenomena, with corresponding implications for coverage.

Second, bibliometric structures are intrinsically sensitive to database scope, citation density, and temporal dynamics. Changing the threshold for bibliographic coupling would reduce the number of papers included in the coupling network, since higher thresholds exclude documents with weaker reference overlaps, thereby decreasing connectivity and potentially biasing cluster identification toward densely cited works. In our case, this threshold was set to the minimum (one shared reference) to preserve network connectivity and avoid excluding relevant but less densely connected papers. Additionally, citation-based metrics tend to favor already well-cited or older works, potentially underrepresenting novel or interdisciplinary studies (van den Besselaar and Sandström 2019). To mitigate such biases, we adopted eigenvector centrality rather than raw citation counts as the primary measure of influence, as it emphasizes connection to other influential works rather than sheer citation volume. This approach aligns with best practices in contemporary bibliometric research (Belter 2015) and provides a more balanced representation of intellectual influence.

Third, the interpretation of clusters and thematic boundaries reflects both algorithmic and human judgment. Although the weighted Leiden algorithm and full-text review enhance robustness, the resulting themes depend on parameter choices and coding interpretations. Hence, cluster delineations should be viewed as indicative rather than absolute, providing a conceptual rather than taxonomic map of the field.

Finally, the results are temporally bound and context-dependent. Rapid advances in AI, emerging regulatory frameworks, and evolving financial technologies may reshape the research landscape beyond the period covered by this study. Accordingly, while the present synthesis offers a robust baseline, it should be revisited and updated to capture the dynamic evolution of research on trust in human–AI collaboration in finance.

5.4 Future research directions

Beyond consolidation, the review reveals the need for a focused and prioritized research agenda. We suggest five research directions that are both theoretically and practically decisive for trustworthy human–AI collaboration in finance:

1. **Longitudinal trust calibration:** future research should move beyond cross-sectional attitudes toward longitudinal designs that track how trust, overreliance, and repair evolve across repeated interactions, system errors, and changing risk conditions. Such studies are essential to distinguish justified trust from blind reliance and to understand when trust becomes miscalibrated in high-stakes financial decisions.
2. **Field experiments in real financial platforms:** laboratory and vignette-based studies should be complemented by field experiments embedded in real-world settings such as robo-advisory platforms, payment systems, reporting tools, and compliance analytics. Field evidence is necessary to assess trust under genuine incentives, regulatory constraints, and accountability pressures that cannot be replicated in controlled environments.
3. **Governance and accountability measurement:** a critical gap concerns the operationalization of governance mechanisms as measurable trust safeguards. Future work should develop and validate indicators for auditability, responsibility attribution, human override protocols, and escalation procedures, linking organizational controls to individual trust calibration and decision quality.
4. **Standardized XAI benchmarks for finance:** explainability research would benefit from shared, finance-specific benchmarks that evaluate not only technical fidelity but also usability, decision-support value, and regulatory adequacy. Standardized benchmarks are a necessary condition for comparing XAI methods across tasks such as credit scoring, trading, and financial reporting.
5. **Cross-cultural effects in advice-taking:** trust in financial AI is shaped by cultural norms, regulatory expectations, and institutional trust. Comparative and cross-cultural studies are needed to examine how advice-taking, reliance, and responsibility attribution vary across jurisdictions, informing the design of globally deployable yet locally legitimate AI systems.

Appendix

See Tables 12 and 13.

Table 12 Top five nodes per cluster ranked by raw eigenvector on each cluster's induced subgraph H_c

Cluster	Node ID	Eigenvector (raw; local norm.)	Degree (raw; local norm.)	Betweenness (raw; local norm.)	Closeness (raw; local norm.)
1	30	0.174 (0.27)	0.63 (0.92)	0.03 (0.14)	1.22 (0.88)
1	36	0.256 (0.39)	0.53 (0.75)	0.04 (0.18)	1.30 (0.99)
1	**38**	**0.640** (**1.00**)	**0.68** (**1.00**)	**0.17** (**0.70**)	**1.31** (**1.00**)
1	69	0.614 (0.96)	0.47 (0.67)	0.04 (0.18)	1.27 (0.95)
1	99	0.169 (0.26)	0.53 (0.75)	0.03 (0.15)	1.13 (0.75)
2	9	0.327 (0.61)	0.50 (1.00)	0.45 (1.00)	0.74 (0.84)
2	29	0.489 (0.93)	0.38 (0.71)	0.04 (0.09)	0.81 (1.00)
2	41	0.295 (0.55)	0.38 (0.71)	0.13 (0.30)	0.64 (0.62)
2	**63**	**0.524** (**1.00**)	**0.44** (**0.86**)	**0.18** (**0.40**)	**0.81** (**0.98**)
2	78	0.402 (0.76)	0.38 (0.71)	0.12 (0.28)	0.69 (0.72)
3	43	0.353 (0.70)	0.36 (0.67)	0.55 (0.86)	0.62 (0.97)
3	79	0.326 (0.65)	0.21 (0.33)	0.04 (0.06)	0.62 (1.00)
3	100	0.376 (0.75)	0.50 (1.00)	0.64 (1.00)	0.61 (0.96)
3	**107**	**0.501** (**1.00**)	**0.29** (**0.50**)	**0.14** (**0.22**)	**0.45** (**0.51**)
3	111	0.494 (0.99)	0.21 (0.33)	0.00 (0.00)	0.45 (0.50)
4	**12**	**0.393** (**1.00**)	**0.72** (**0.77**)	**0.07** (**0.40**)	**2.30** (**0.88**)
4	23	0.391 (0.99)	0.83 (0.92)	0.01 (0.08)	2.46 (1.00)
4	27	0.333 (0.83)	0.89 (1.00)	0.16 (1.00)	2.29 (0.88)
4	31	0.383 (0.97)	0.78 (0.85)	0.08 (0.51)	2.29 (0.88)
4	75	0.332 (0.82)	0.67 (0.69)	0.04 (0.23)	2.20 (0.81)
5	56	0.375 (0.91)	0.50 (1.00)	0.07 (0.30)	1.04 (1.00)
5	68	0.368 (0.89)	0.20 (0.33)	0.02 (0.06)	0.79 (0.59)
5	**73**	**0.411** (**1.00**)	**0.50** (**1.00**)	**0.08** (**0.32**)	**1.05** (**1.00**)
5	89	0.328 (0.80)	0.40 (0.78)	0.04 (0.15)	1.04 (0.99)
5	97	0.373 (0.91)	0.25 (0.44)	0.03 (0.12)	0.84 (0.68)
6	55	0.149 (0.23)	0.50 (0.83)	0.30 (0.71)	1.09 (0.49)
6	90	0.366 (0.60)	0.58 (1.00)	0.42 (1.00)	1.39 (0.82)
6	92	0.307 (0.50)	0.58 (1.00)	0.17 (0.39)	1.56 (1.00)
6	108	0.602 (1.00)	0.50 (0.83)	0.09 (0.20)	1.51 (0.95)
6	**109**	**0.603** (**1.00**)	**0.58** (**1.00**)	**0.08** (**0.18**)	**1.50** (**0.93**)

Raw centrality values (eigenvector with three decimals; others with two) and, in parentheses, the corresponding cluster-normalized scores (min–max within cluster to [0, 1], two decimals)

Bold text denotes the maximum observed values

Table 13 Operational specification of the multi-level socio-technical framework of trust in human–AI collaboration in finance

Analytical level	Boundary conditions	Trust-shaping mechanisms	Indicators
Micro (Individual perception and calibration)	High-stakes financial decisions involving uncertainty and partial automation (e.g., credit approval, investment advice, compliance assessment) affecting clients, professionals, or supervisors	Cognitive evaluation of AI competence, fairness, and reliability; affective responses; calibration between perceived trustworthiness and actual system capability or authority	Subjective trust measures; trust-reliance mismatch; over-/under-reliance rates; confidence-accuracy gaps; frequency of human escalation or override
Meso (Organizational design and corporate governance)	AI systems embedded in organizational workflows with defined decision rights, accountability allocation, and internal oversight (e.g., banks, insurers, asset managers, regulated fintechs)	Translation of governance and risk controls into user-facing assurances via XAI, UX design, human-in-the-loop arrangements, escalation protocols, and accountability mapping	Presence of override protocols; audit trails; explainability artifacts (e.g., SHAP/LIME documentation); model governance documentation; incident and override logs; governance maturity indicators
Macro (Regulatory and technological infrastructures)	AI deployment subject to regulatory supervision and/or reliance on shared technological infrastructures encoding verification, traceability, or compliance (e.g., reporting regimes, payment rails, blockchain-based provenance systems)	Institutionalization of trust through regulation, standardization, and infrastructural substitution of interpersonal trust via automated verification, traceability, immutability, and certification	Regulatory approval outcomes; compliance audit results; adoption of standardized validation or XAI benchmarks; reliance on infrastructural assurance (e.g., blockchain traceability, automated reconciliation, certification systems)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-026-03049-y.

Acknowledgements Jose M. Alonso-Moral recognizes the support of the Galician Ministry of Education, Science, Universities and Professional Training (grant CiTIUS 2024-2027 ED431G2023/04) and the European Union (European Regional Development Fund - ERDF). This work is also supported by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033/) through MAIXAI4TRUST PID2024-157680NB-I00 grant and the Fundación Ramón Areces (Spain) through the CONFIA project.

Author contributions M.M. and J.M.A.M. contributed to the conceptualization of the study. M.M. developed the methodology, wrote the code, performed the formal analysis, and carried out validation, investigation, and data curation. M.M. wrote the original draft and prepared the visualizations. M.M., G.E.C., and J.M.A.M. reviewed and edited the manuscript. G.E.C. and J.M.A.M. supervised the work. All authors read and approved the final manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The dataset used in this study was deposited in the Zenodo repository under the DOI https://doi.org/10.5281/zenodo.17922549 and is available at the following URL: https://doi.org/10.5281/zenodo.17922549.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abou Ali M, Dornaika F, Charafeddine J (2026) Agentic AI: a comprehensive survey of architectures, applications, and future directions. Artif Intell Rev 59(1):11. https://doi.org/10.1007/s10462-025-11422-4

Agnihotri A, Bhattacharya S (2024) Chatbots' effectiveness in service recovery. Int J Inf Manag 76:102679. https://doi.org/10.1016/j.ijinfomgt.2023.102679

Ahdadou M, Aajly A, Tahrouch M (2024) Unlocking the potential of augmented intelligence: a discussion on its role in boardroom decision-making. Int J Discl Gov 21(3):433–446. https://doi.org/10.1057/s41310-023-00207-2

Aldasoro I, Gambacorta L, Korinek A et al (2025) Intelligent financial system: how AI is transforming finance. J Financ Stab 81:101472. https://doi.org/10.1016/j.jfs.2025.101472

Aldridge I, An J, Burke R, Cao M, Chien C-Y, Deng K, Deng R, Gao Y, Guo O, He S, Li Z, Lin G, Lin W, Lyu F, Ng K, Wang Q, Xiao

H, Xu D, Xue Y, Zhang S, Zhang S, Zhang Y, Zhao S, Zhao X, Zhao Y, Zheng W (2025) Agentic artificial intelligence in finance: a comprehensive survey. SSRN Electron J. https://doi.org/10.2139/ssrn.5803628

Ali S, Abuhmed T, El-Sappagh S et al (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Inf Fusion 99:101805. https://doi.org/10.1016/j.inffus.2023.101805

Alkhalil Z, Hewage C, Nawaf L et al (2021) Phishing attacks: a recent comprehensive study and a new anatomy. Front Comput Sci. https://doi.org/10.3389/fcomp.2021.563060

Armour J, Eidenmüller HGM (2019) Self-driving corporations? Harv Bus Law Rev. https://doi.org/10.2139/ssrn.3442447

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bawack RE, Wamba SF, Carillo KDA et al (2022) Artificial intelligence in E-Commerce: a bibliometric study and literature review. Electron Mark 32(1):297–338. https://doi.org/10.1007/s12525-022-00537-z

Behera RK, Bala PK, Rana NP (2023) Creation of sustainable growth with explainable artificial intelligence: an empirical insight from consumer packaged goods retailers. J Clean Prod 399:136605. https://doi.org/10.1016/j.jclepro.2023.136605

Behl A, Sampat B, Pereira V et al (2023) The role played by responsible artificial intelligence (RAI) in improving supply chain performance in the MSME sector: an empirical inquiry. Ann Oper Res. 354:5–34. https://doi.org/10.1007/s10479-023-05624-8

Belanche D, Casaló LV, Flavián C (2019) Artificial Intelligence in FinTech: understanding robo-advisors adoption among customers. Ind Manag Data Syst 119(7):1411–1430. https://doi.org/10.1108/IMDS-08-2018-0368

Belter CW (2015) Bibliometric indicators: opportunities and limits. J Med Libr Assoc 103(4):219–221. https://doi.org/10.3163/1536-5050.103.4.014

Bhatia A, Chandani A, Chhateja J (2020) Robo advisory and its potential in addressing the behavioral biases of investors—a qualitative study in Indian context. J Behav Exp Finance 25:100281. https://doi.org/10.1016/j.jbef.2020.100281

Bhatia A, Chandani A, Atiq R et al (2021) Artificial intelligence in financial services: a qualitative research to discover robo-advisory services. Qual Res Financ Mark 13(5):632–654. https://doi.org/10.1108/QRFM-10-2020-0199

Bonacich P (2007) Some unique properties of eigenvector centrality. Soc Netw 29(4):555–564. https://doi.org/10.1016/j.socnet.2007.04.002

Breiman L (2001) Statistical modeling: the two cultures. Stat Sci Rev J Inst Math Stat 16(3):199–231. https://doi.org/10.1214/ss/1009213726

Brenner L, Meyll T (2020) Robo-advisors: a substitute for human financial advice? J Behav Exp Finance 25:100275. https://doi.org/10.1016/j.jbef.2020.100275

Bussmann N, Giudici P, Marinelli D et al (2021) Explainable machine learning in credit risk management. Comput Econ 57(1):203–216. https://doi.org/10.1007/s10614-020-10042-0

Camilleri MA (2024) Artificial intelligence governance: ethical considerations and implications for social responsibility. Expert Syst 41(7):e13406. https://doi.org/10.1111/exsy.13406

Castelfranchi C, Falcone R (2010) Socio-cognitive model of trust: basic ingredients. In: Trust theory: a socio-cognitive and computational model. Wiley, pp 35–94. https://doi.org/10.1002/9780470519851.ch2

Cecchetti SG, Lumsdaine RL, Peltonen T, Sánchez Serrano A (2025) Artificial intelligence and systemic risk. SSRN Electron J. https://doi.org/10.2139/ssrn.5866223

Černevičienė J, Kabašinskas A (2024) Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artif Intell Rev 57(8):216. https://doi.org/10.1007/s10462-024-10854-8

Chen J, Guo F, Ren Z et al (2024) Effects of anthropomorphic design cues of chatbots on users' perception and visual behaviors. Int J Hum-Comput Interact 40(14):3636–3654. https://doi.org/10.1080/10447318.2023.2193514

Chen Y, Aw ECX, Tan GWH (2025) Financial empowerment through robo-advisors: understanding the keys to trust and loyalty. Ind Manag Data Syst 125(6):2178–2205. https://doi.org/10.1108/IMDS-05-2024-0461

Chowdhury S, Joel-Edgar S, Dey PK et al (2023) Embedding transparency in artificial intelligence machine learning models: managerial implications on predicting and explaining employee turnover. Int J Hum Resour Manag 34(14):2732–2764. https://doi.org/10.1080/09585192.2022.2066981

Cramer H, Garcia-Gathright J, Springer A et al (2018) Assessing and addressing algorithmic bias in practice. Interact 25(6):58–63. https://doi.org/10.1145/3278156

Dafoe A, Hughes E, Bachrach Y et al (2020) Open problems in cooperative AI. arxiv:2012.08630

Daníelsson J, Macrae R, Uthemann A (2022) Artificial intelligence and systemic risk. J Bank Financ 140:106290. https://doi.org/10.1016/j.jbankfin.2021.106290

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–340

Dellermann D, Calma A, Lipusch N et al (2019) The future of human–AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. In: Bui TX (ed) Proceedings of the 52nd annual Hawaii international conference on system sciences: January 8–11, 2019, Maui, Hawaii. University of Hawaii at Manoa Hamilton Library ScholarSpace, Honolulu, HI, pp 274–283. http://hdl.handle.net/10125/59468

Dixon MF, Halperin I, Bilokon P (2020) Machine learning in finance: from theory to practice. Springer, Cham. https://doi.org/10.1007/978-3-030-41068-1

Donthu N, Kumar S, Mukherjee D et al (2021) How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res 133:285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Dou WW, Goldstein I, Ji Y (2025) AI-powered trading, algorithmic collusion, and price efficiency. Working Paper 34054, National Bureau of Economic Research. https://doi.org/10.3386/w34054

Duong CD, Nguyen TH, Ngo TVN et al (2025) Blockchain technology and consumers' organic food consumption: a moderated mediation model of blockchain-based trust and perceived blockchain-related information transparency. J Asia Bus Stud 19(1):54–78. https://doi.org/10.1108/JABS-07-2024-0387

Financial Stability Board (2025) Monitoring adoption of artificial intelligence and related vulnerabilities in the financial sector. Financial Stability Board, Basel. October 10, 2025. https://www.fsb.org/2025/10/monitoring-adoption-of-artificial-intelligence-and-related-vulnerabilities-in-the-financial-sector/

Floridi L (2019) Establishing the rules for building trustworthy AI. Nat Mach Intell 1(6):261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi L (2025) AI as agency without intelligence: on artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. Philos Technol 38(1):30. https://doi.org/10.1007/s13347-025-00858-9

Foucault T, Gambacorta L, Jiang W, Vives X (2025) Artificial intelligence in finance. CEPR Press, Paris

Fronzetti Colladon A, Naldi M (2020) Distinctiveness centrality in social networks. PLoS ONE 15(5):1–21. https://doi.org/10.1371/journal.pone.0233276

Garcia ACB, Garcia MGP, Rigobon R (2024) Algorithmic discrimination in the credit domain: what do we know about it? AI Soc 39(4):2059–2098. https://doi.org/10.1007/s00146-023-01676-3

Gebru T, Morgenstern J, Vecchione B et al (2021) Datasheets for datasets. Commun ACM 64(12):86–92. https://doi.org/10.1145/3458723

Gefen D, Karahanna E, Straub DW (2003) Trust and TAM in online shopping: an integrated model. MIS Q. https://doi.org/10.2307/30036519

Giudici P (2018) Fintech risk management: a research challenge for artificial intelligence in finance. Front Artif Intell. https://doi.org/10.3389/frai.2018.00001

Graham G, Nisar TM, Prabhakar G et al (2025) Chatbots in customer service within banking and finance: Do chatbots herald the start of an AI revolution in the corporate world? Comput Hum Behav 165:108570. https://doi.org/10.1016/j.chb.2025.108570

Guidotti R, Monreale A, Ruggieri S et al (2018) A survey of methods for explaining black box models. ACM Comput Surv. https://doi.org/10.1145/3236009

Gunning D (2019) DARPA's explainable artificial intelligence (XAI) program. In: Proceedings of the 24th international conference on intelligent user interfaces. Association for Computing Machinery, New York, NY, USA, IUI '19, p ii. https://doi.org/10.1145/3301275.3308446

He X, Liu Y (2024) Knowledge evolutionary process of artificial intelligence in e-commerce: main path analysis and science mapping analysis. Expert Syst Appl 238:121801. https://doi.org/10.1016/j.eswa.2023.121801

Heng YS, Subramanian P (2023) A systematic review of machine learning and explainable artificial intelligence (XAI) in credit risk modelling. In: Arai K (ed) Proceedings of the future technologies conference (FTC) 2022, vol 1. Springer International Publishing, Cham, pp 596–614. https://doi.org/10.1007/978-3-031-18461-1_39

Hickman E, Petrin M (2021) Trustworthy AI and corporate governance: the EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. Eur Bus Organ Law Rev 22(4):593–625. https://doi.org/10.1007/s40804-021-00224-0

Hilb M (2020) Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. J Manag Gov 24(4):851–870. https://doi.org/10.1007/s10997-020-09519-9

Holland JH (1992) Complex adaptive systems. Daedalus 121(1):17–30

Jellason NP, Ambituuni A, Adu DA et al (2024) The potential for blockchain to improve small-scale agri-food business' supply chain resilience: a systematic review. Br Food J 126(5):2061–2083. https://doi.org/10.1108/BFJ-07-2023-0591

Jung D, Dorner V, Weinhardt C et al (2018) Designing a robo-advisor for risk-averse, low-budget consumers. Electron Mark 28(3):367–380. https://doi.org/10.1007/s12525-017-0279-9

Khan FS, Mazhar SS, Mazhar K et al (2025) Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. Artif Intell Rev 58(8):232. https://doi.org/10.1007/s10462-025-11215-9

Khushk A, Zhiying L, Yi X et al (2024) Navigating human–AI dynamics: implications for organizational performance (SLR). Int J Organ Anal 33(5):1293–1311. https://doi.org/10.1108/IJOA-04-2024-4456

Kim T, Song H (2021) How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. Telemat Inform 61:101595. https://doi.org/10.1016/j.tele.2021.101595

Lee MK, Kusbit D, Kahng A et al (2019) WeBuildAI: participatory framework for algorithmic governance. Proc ACM Hum-Comput Interact 3(CSCW):181. https://doi.org/10.1145/3359283

Lehner OM, Ittonen K, Silvola H et al (2022) Artificial intelligence based decision-making in accounting and auditing: ethical challenges and normative thinking. Acc Audit Acc J 35(9):109–135. https://doi.org/10.1108/AAAJ-09-2020-4934

Leitner G, Singh J, van der Kraaij A et al (2024) The rise of artificial intelligence: benefits and risks for financial stability. Finance Stab Rev 1:2

Li J, Wu L, Qi J et al (2023a) Determinants affecting consumer trust in communication with ai chatbots: the moderating effect of privacy concerns. J Organ End User Comput 35(1):1–24. https://doi.org/10.4018/JOEUC.328089

Li Y, Wang S, Ding H et al (2023b) Large language models in finance: a survey. In: Proceedings of the fourth ACM international conference on AI in finance. Association for Computing Machinery, New York, NY, USA, ICAIF '23, pp 374–382. https://doi.org/10.1145/3604237.3626869

Lipton ZC (2018) The mythos of model interpretability. Commun ACM 61(10):36–43. https://doi.org/10.1145/3233231

Loaiza I, Rigobon R (2025) The limits of AI in financial services. SSRN. https://doi.org/10.2139/ssrn.5196350

Machucho R, Ortiz D (2025) The impacts of artificial intelligence on business innovation: a comprehensive review of applications, organizational challenges, and ethical considerations. Systems 13(4):264. https://doi.org/10.3390/systems13040264

Majeed Parry G, Revolidis I, Ellul J et al (2024) Bottling up trust: a review of blockchain adoption in wine supply chain traceability. IEEE Access 12:178320–178344. https://doi.org/10.1109/ACCESS.2024.3505428

Makarius EE, Mukherjee D, Fox JD et al (2020) Rising with the machines: a sociotechnical framework for bringing artificial intelligence into the organization. J Bus Res 120:262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

Maple C, Szpruch L, Epiphaniou G et al (2023) The AI revolution: opportunities and challenges for the finance sector. https://doi.org/10.48550/arXiv.2308.16538

Marzi G, Balzano M, Caputo A et al (2025) Guidelines for bibliometric–systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. Int J Manag Rev 27(1):81–103. https://doi.org/10.1111/ijmr.12381

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller GJ (2022) Stakeholder roles in artificial intelligence projects. Proj Leadersh Soc 3:100068. https://doi.org/10.1016/j.plas.2022.100068

Moro-Visconti R (2025) Is artificial intelligence a new stakeholding agent? Hum-Intell Syst Integr. 7:3–16. https://doi.org/10.1007/s42454-025-00069-9

Nass C, Steuer J, Tauber ER (1994) Computers are social actors. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '94, pp 72–78. https://doi.org/10.1145/191666.191703

NetworkX (2025) Eigenvector_centrality—NetworkX documentation. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.eigenvector_centrality.html

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103

Nourallah M (2023) One size does not fit all: young retail investors' initial trust in financial robo-advisors. J Bus Res 156:113470. https://doi.org/10.1016/j.jbusres.2022.113470

Owens E, Sheehan B, Mullins M et al (2022) Explainable artificial intelligence (XAI) in insurance. Risks. 10(12):230. https://doi.org/10.3390/risks10120230

Page MJ, McKenzie JE, Bossuyt PM et al (2021a) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372. https://doi.org/10.1136/bmj.n71

Page MJ, Moher D, Bossuyt PM et al (2021b) Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. https://doi.org/10.1136/bmj.n160

Papagiannidis E, Enholm IM, Dremel C et al (2023) Toward AI governance: identifying best practices and potential barriers and outcomes. Inf Syst Front 25(1):123–141. https://doi.org/10.1007/s10796-022-10251-y

Papagiannidis E, Mikalef P, Conboy K (2025) Responsible artificial intelligence governance: a review and research framework. J Strateg Inf Syst 34(2):101885. https://doi.org/10.1016/j.jsis.2024.101885

Pelau C, Dabija DC, Ene I (2021) What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Hum Behav 122:106855. https://doi.org/10.1016/j.chb.2021.106855

Perianes-Rodriguez A, Waltman L, van Eck NJ (2016) Constructing bibliometric networks: a comparison between full and fractional counting. J Informetr 10(4):1178–1195. https://doi.org/10.1016/j.joi.2016.10.006

Reuel A, Bucknall B, Casper S et al (2025) Open problems in technical AI governance. arXiv:2407.14981

Riedel A, Mulcahy R, Northey G (2022) Feeling the love? How consumer's political ideology shapes responses to AI financial service delivery. Int J Bank Mark 40(6):1102–1132. https://doi.org/10.1108/IJBM-09-2021-0438

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Saarela M, Podgorelec V (2024) Recent applications of explainable AI (XAI): a systematic literature review. Appl Sci. https://doi.org/10.3390/app14198884

Sabharwal R, Miah SJ, Wamba SF et al (2025) Extending application of explainable artificial intelligence for managers in financial organizations. Ann Oper Res 354: 309–339. https://doi.org/10.1007/s10479-024-05825-9

SankeyMATIC (2025) Home. https://sankeymatic.com/.

Scholes MS (2025) Artificial intelligence and uncertainty. Risk Sci 1:100004. https://doi.org/10.1016/j.risk.2024.100004

Schreibelmayr S, Moradbakhti L, Mara M (2023) First impressions of a financial ai assistant: differences between high trust and low trust users. Front Artif Intell. https://doi.org/10.3389/frai.2023.1241290

Sharma R (2024) AI governance. In: AI and the boardroom. Apress, Berkeley, pp 5–26. https://doi.org/10.1007/979-8-8688-0796-1_2

Sharma A, Lin IW, Miner AS et al (2023) Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 5(1):46–57. https://doi.org/10.1038/s42256-022-00593-2

Sowa K, Przegalinska A, Ciechanowski L (2021) Cobots in knowledge work: human–AI collaboration in managerial professions. J Bus Res 125:135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

Sri Vigna Hema V, Manickavasagan A (2024) Blockchain implementation for food safety in supply chain: a review. Compr Rev Food Sci Food Saf 23(5):e70002. https://doi.org/10.1111/1541-4337.70002

Srinivas Aditya U, Singh R, Singh PK et al (2021) A survey on blockchain in robotics: issues, opportunities, challenges and future directions. J Netw Comput Appl 196:103245. https://doi.org/10.1016/j.jnca.2021.103245

Tol P (2021) Colour schemes. Tech. Rep. SRON/EPS/TN/09-002, SRON Netherlands Institute for Space Research. https://sronpersonalpages.nl/~pault/

Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9(1):5233. https://doi.org/10.1038/s41598-019-41695-z

Ustahaliloğlu MK (2025) Artificial intelligence in corporate governance. Corp Law Gov Rev 7(1):123–134. https://doi.org/10.22495/clgrv7i1p11

Uuk R, Gutierrez CI, Guppy D et al (2024) A taxonomy of systemic risks from general-purpose AI. RAND working paper series. SSRN Electron J. https://doi.org/10.2139/ssrn.5030173 (**forthcoming**)

van den Besselaar P, Sandström U (2019) Measuring researcher independence using bibliometric data: a proposal for a new performance indicator. PLoS ONE. https://doi.org/10.1371/journal.pone.0202712

Véliz C (2021) Moral zombies: why algorithms are not moral agents. AI Soc 36(2):487–497. https://doi.org/10.1007/s00146-021-01189-x

Wang D, Weisz JD, Muller M et al (2019) Human–AI collaboration in data science: exploring data scientists' perceptions of automated AI. Proc ACM Hum-Comput Interact. https://doi.org/10.1145/3359313

Wang D, Churchill E, Maes P et al (2020) From human–human collaboration to human–AI collaboration: designing AI systems that can work together with people. In: Extended abstracts of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI EA '20, pp 1–6. https://doi.org/10.1145/3334480.3381069

Wang Y, Mohsin M, Jamil K (2025) Customer trust and willingness to use shopping assistant humanoid chatbot. Serv Ind J. https://doi.org/10.1080/02642069.2024.2432935

Weber P, Carl KV, Hinz O (2024) Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. Manag Rev Q 74(2):867–907. https://doi.org/10.1007/s11301-023-00320-0

Yan H, Wei Y, Xiong H (2025) How do initial and interactive social cues increase customers' continuance usage intention of chatbots? Int J Hum-Comput Interact 41(8):4700–4717. https://doi.org/10.1080/10447318.2024.2352928

Yavaprabhas K, Pournader M, Seuring S (2023) Blockchain as the trust-building machine for supply chain management. Ann Oper Res 327(1):49–88. https://doi.org/10.1007/s10479-022-04868-0

Zhang J, Luo Y (2017) Degree centrality, betweenness centrality, and closeness centrality in social network. In: Proceedings of the 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017), pp 300–303. https://doi.org/10.2991/msam-17.2017.68

Zupic I, Čater T (2015) Bibliometric methods in management and organization. Organ Res Methods 18(3):429–472. https://doi.org/10.1177/1094428114562629