

Responsible Large Language Models in Finance: A Descriptive Bibliometric Overview and Taxonomy of Responsibility

Chong Hui Tan and Qinxu Ding * 

School of Business, Singapore University of Social Sciences, 463 Clementi Road, Singapore 599494, Singapore

* Correspondence: qinxuding@suss.edu.sg

Abstract

The rapid adoption of large language models (LLMs) in financial services has generated a growing literature on “responsible AI” in domains such as investment analysis, credit assessment, risk management, compliance, and financial advisory systems. Unlike earlier AI systems, LLMs introduce responsibility challenges that differ in important ways from those addressed by earlier responsible AI frameworks, including hallucinations, prompt injection and manipulation, generative opacity, instruction-following failures, context sensitivity, output inconsistency, and emergent capabilities that arise only at larger model scales. While responsible AI concerns extend across AI systems more broadly, this review focuses specifically on LLMs, which since 2022 have become a major technological force reshaping financial services and have generated a distinctive body of responsibility discourse that existing frameworks are only beginning to address. However, the meaning of responsibility in this literature remains heterogeneous and inconsistently operationalized across studies. This paper provides a structured review of research on responsible LLMs in finance using a broad-to-narrow screening process across Web of Science and Scopus, reported with reference to the applicable PRISMA 2020 items. We combine a descriptive bibliometric overview with qualitative content analysis to examine the corpus. The descriptive overview covers publication trends, publication venues, disciplinary orientations, geographic distributions, and institutional affiliations, while the qualitative analysis codes the corpus on two dimensions: responsibility depth—whether responsibility is constitutive of the paper’s contribution, operative within its design, or peripheral to its framing—and responsibility mode—whether that engagement is normative, technical, evaluative, or mixed. The findings reveal a field concentrated in the Operative–Technical mode. Constitutive contributions cluster in Normative and Mixed modes, while a purely Evaluative orientation remains comparatively uncommon. Addressing these limitations requires moving beyond principle-level discourse toward institution-specific accountability mechanisms, empirical evaluation across clearly documented research and operational settings, and systematic alignment with relevant financial regulatory requirements. These efforts must address the distinctive challenges posed by LLMs, rather than only concerns inherited from the broader responsible AI literature.



Academic Editor: Jungchieh Lee

Received: 9 July 2026

Revised: 12 August 2026

Accepted: 16 August 2026

Published: 18 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.**Keywords:** large language models; responsible AI; finance; systematic review; responsibility depth; responsibility mode**JEL Classification:** G20; G28; O33

1. Introduction

Large language models (LLMs) are rapidly reshaping financial services, with applications spanning investment analysis, credit assessment, fraud detection, financial advisory systems, risk management, regulatory compliance, and customer interaction [1–3]. As these systems increasingly generate financial narratives, interpret disclosures, and automate analytical tasks, concerns surrounding fairness, transparency, explainability, robustness, privacy, governance, and accountability have moved to the forefront of both academic and policy discussions [4]. This review focuses specifically on LLMs, which since 2022 have become a major technological force in financial services and which introduce responsibility challenges that differ in important ways from those addressed by earlier responsible AI frameworks. These challenges include hallucinations [5], prompt injection and manipulation, generative opacity [6], instruction-following failures and alignment instability [7], context sensitivity, output inconsistency, and emergent capabilities that arise only at larger model scales [8,9]. These challenges arise not merely because LLMs are deployed in high-stakes financial contexts, but because of the nature of the technology itself: generative systems whose outputs are probabilistic, whose internal decision-making is not directly inspectable, and whose behaviour may change in ways that developers and deployers cannot fully anticipate. Because many relevant studies use broader terms such as “AI” or “machine learning” rather than “large language model”, we adopt a broad-to-narrow review strategy: broad responsible AI in finance terms construct the initial retrieval pool, while LLM relevance is applied during screening and final inclusion.

Existing reviews on AI and finance have focused on broader themes such as machine learning applications, bibliometric trends, or explainable AI (XAI) techniques [1,4,10], without treating the distinction between general responsible AI concerns and LLM-specific generative challenges as an explicit object of analysis. The meaning of responsibility in this emerging literature remains heterogeneous and inconsistently operationalized, with some papers producing governance frameworks as their primary output, others embedding responsibility within system design, and still others invoking it only as framing [1,11,12].

This paper addresses this gap through a structured review combining a descriptive bibliometric overview with qualitative content analysis [13], applied to 109 corpus papers via a two-dimensional coding structure. The descriptive bibliometric overview characterises the literature in terms of publication trends, publication venues, disciplinary orientations, geographic distributions, and institutional affiliations. This paper contributes to the literature in three ways. First, it provides a structured synthesis of the emerging literature on responsible LLMs in finance. Second, it develops a two-dimensional coding structure—responsibility depth and responsibility mode—that clarifies the different ways responsibility is conceptualised across studies, distinguishing between concerns inherited from the broader responsible AI literature and those arising from LLMs’ generative and emergent properties. Third, it identifies major gaps between high-level responsible LLM principles and the practical requirements of accountability, governance, and deployment in financial settings.

2. Methods and Data

2.1. Data Collection

Academic records were retrieved from Scopus and Web of Science using a search strategy capturing the intersection of three conceptual dimensions: responsibility-related considerations, AI technologies, and financial application domains [14]. The following search string was applied to both databases (WoS: TS; Scopus: TITLE-ABS-KEY):

(fairness OR bias OR transparency OR explainab* OR interpretab* OR accountab* OR auditab* OR privacy OR secur* OR robust* OR governance OR “human oversight” OR

ethic* OR trust*) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “language model*” OR “large language model*” OR LLM* OR “generative AI” OR “foundation model*”) AND (financ* OR bank* OR “financial service*” OR credit OR lending OR trading OR “asset management” OR investment OR “capital market*” OR fintech)

Searches were restricted to titles, abstracts, and keywords and limited to English-language journal articles and conference proceedings. The bibliographic search and systematic review covered records published from 1 January 2023 to 20 April 2026. The year 2023 was selected as the starting point because the review focuses on the literature emerging after the public release and widespread diffusion of ChatGPT (GPT-3.5) and related LLM applications in late 2022. This period captures the stage in which LLMs became a substantive technological and research focus in financial services. The same temporal boundary was applied consistently to the initial retrieval, eligibility assessment, final review corpus, descriptive bibliometric overview, and qualitative coding. Because the bibliographic searches were conducted on 20 April 2026, records from 2026 represent only a partial publication year and should not be interpreted as a complete annual publication count.

2.2. Screening and Selection

This review was guided by the applicable reporting items in the PRISMA 2020 statement [14]. Specifically, we report the eligibility criteria (Item 5), information sources and date of the last search (Item 6), complete database search strategy (Item 7), record-screening and study-selection process, including the role of semantic prioritisation (Item 8), data-extraction and coding procedures (Items 9 and 10), methods used to tabulate and synthesise the evidence (Item 13), and the number of records identified, screened, excluded, and included (Item 16a). The review’s retrospective OSF registration is also reported in accordance with Item 24. Figure 1 presents the study-selection process using the PRISMA 2020 flow-diagram structure. PRISMA 2020 was used as a reporting framework for search and study selection; it did not determine the descriptive bibliometric or qualitative analytical frameworks.

The initial merged dataset contained 32,900 records. Following DOI- and title-based deduplication, 23,934 unique records remained. Records were then subjected to a semi-automated semantic-screening procedure using sentence-level embeddings [15] derived from titles, abstracts, and keywords. Screening was conducted in two stages: first for relevance to responsible AI in financial contexts and subsequently for substantive LLM engagement. A record was retained only if it satisfied both stages: scoring in the top 5% of responsible-AI-in-finance relevance across the full deduplicated pool, and separately meeting a semantic LLM-proximity threshold together with an explicit LLM-specific term (“large language model”, “LLM”, “generative AI”, “foundation model”, or equivalent) in its text. Broader AI and machine learning terms in the search string (Section 2.1) served only to construct a sufficiently broad retrieval pool, since much of the literature uses these broader terms even when LLMs are the object of study; they did not by themselves satisfy the inclusion bar. Appendix A reports the semantic-screening procedure in detail, including a recall-validation audit empirically justifying the 95th percentile threshold and the 0.08 penalty term. Semantic ranking was used only to prioritise records for screening and did not make final inclusion or exclusion decisions. Final eligibility decisions were made through manual screening. The numbers of records identified, deduplicated, screened, excluded, and included, together with the principal reasons for full-text exclusion, are reported in Figure 1. The final corpus comprises 109 peer-reviewed journal articles and conference papers.

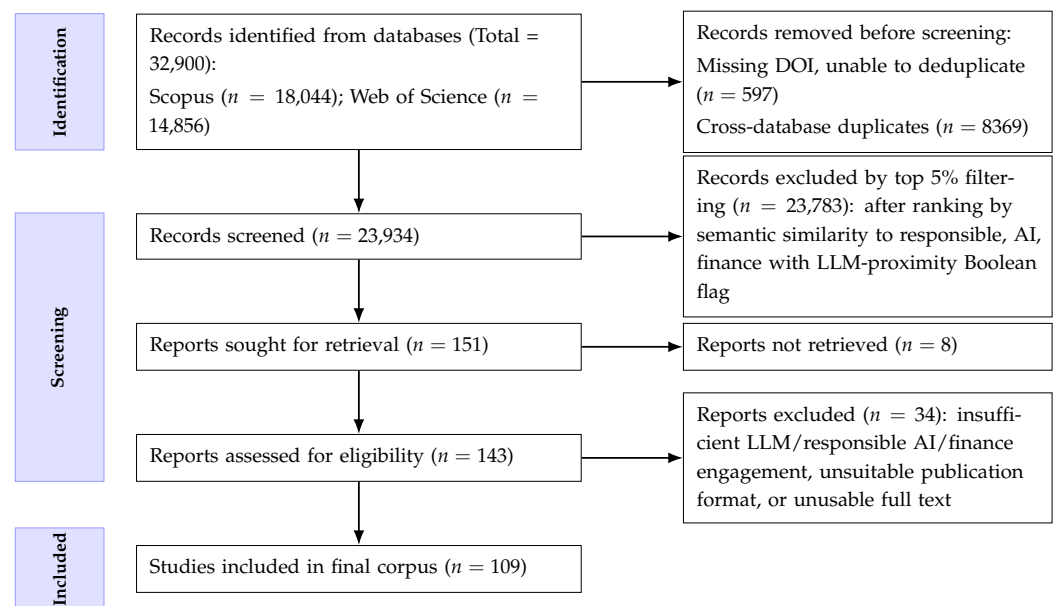


Figure 1. Study-selection process presented using the PRISMA 2020 flow-diagram structure.

2.3. Analytical Approach

The review combines a descriptive bibliometric overview with qualitative content analysis [16,17]. The descriptive bibliometric overview summarises publication trends, publication venues, disciplinary orientations, geographic distributions, and institutional affiliations. It is intended to characterise the composition and development of the corpus rather than to conduct citation-based science mapping, such as co-citation analysis, bibliographic coupling, or collaboration-network analysis.

Geographic and institutional classifications were produced through a hybrid procedure combining manual verification with programmatic aggregation. First, the author-affiliation strings obtained from Scopus and Web of Science were manually checked and reconciled to address incomplete, inconsistent, or ambiguous institutional and geographic information. Country and territory labels were then standardised, and countries were assigned programmatically to their corresponding geographic regions using a predefined country–region mapping. Hong Kong was retained as a separate affiliation territory in accordance with the labels contained in the bibliographic records.

Geographic distributions were calculated using full counting: a paper was counted once for every country and region represented in its author affiliations. A paper was classified as an international collaboration when its authors reported affiliations from at least two countries or territories.

Institutional types were coded from the verified affiliation information. Universities and academic research institutes were classified as academic; commercial companies as industry; government agencies, regulators, central banks, and other public authorities as public sector; and authors explicitly identified as independent researchers without an organisational affiliation as independent. These affiliation-level classifications were subsequently aggregated programmatically into the paper-level categories reported in Table 1, including academic only, industry only, academic–industry, academic–independent researcher, academic–public-sector, and independent researcher only. Ambiguous affiliations were reviewed manually before the final counts were generated.

Table 1. Geographic distribution, international collaboration, and institutional composition.

Dimension	Category	Papers	Share of Corpus (%)
Panel A: Leading countries			
United States	Country affiliation present	39	35.8
India	Country affiliation present	20	18.3
China	Country affiliation present	15	13.8
United Kingdom	Country affiliation present	11	10.1
Spain	Country affiliation present	6	5.5
Germany	Country affiliation present	4	3.7
Australia	Country affiliation present	4	3.7
Canada	Country affiliation present	4	3.7
Italy	Country affiliation present	4	3.7
Switzerland	Country affiliation present	3	2.8
Panel B: International collaboration			
Collaboration	Single-country paper	73	67.0
Collaboration	International collaboration	36	33.0
Panel C: Institutional composition			
Institution	Academic only	74	67.9
Institution	Industry only	18	16.5
Institution	Academic–industry	10	9.2
Institution	Academic–independent researcher	3	2.8
Institution	Academic–public sector	2	1.8
Institution	Independent researcher only	2	1.8

Notes: Country and regional distributions use full counting. A paper is counted once for each country or region represented in its author affiliations; therefore, the geographic percentages do not sum to 100%. International collaboration refers to papers containing affiliations from at least two countries. Hong Kong is retained as a separate affiliation territory in accordance with the geographic labels used in the bibliographic records. An “independent researcher” refers to an author who is explicitly identified in the bibliographic record as an independent researcher and does not report an affiliation with a university or research institution, commercial organisation, or government or public-sector body. “Academic–independent researcher” indicates collaboration between at least one academically affiliated author and at least one such independent researcher, whereas “independent researcher only” indicates that no author reports an organisational affiliation.

The coding structure applies two dimensions to each article. The first, responsibility depth, captures how substantively responsibility is engaged: as Constitutive (the paper’s primary output is a framework, principles, or governance structure), Operative (responsibility concepts are built into the research design, methods, or evaluation criteria), or Peripheral (responsibility appears in framing or motivation but is not operationalised in methods or findings). The second, responsibility mode, applies to Constitutive and Operative papers and captures how responsibility is operationalised: through normative argument or governance principles (Normative), system design or algorithmic constraint (Technical), measurement or empirical assessment (Evaluative), or a substantial combination (Mixed).

Thirteen concept flags—fairness, transparency, explainability, accountability, auditability, governance, bias, robustness, privacy, security, trustworthiness, human oversight, and ethics—were scored on explicit textual presence. Each concept flag was coded as a binary indicator of whether the concept was explicitly referenced in the paper, regardless of whether it was introduced as contextual background, discussed conceptually, operationalised within the study design, or empirically evaluated. Consequently, the concept flags capture thematic coverage across the corpus rather than the depth or maturity of engagement with individual responsibility concepts. Studies were coded through structured extraction applied to full-text PDFs, with all codings reviewed and reconciled by the authors through iterative discussion. Structured extraction was performed with LLM assistance (Claude Sonnet 5), applying coding rules mirrored in the human-rater coding manual in Appendix B, with both authors independently reviewing and reconciling

every extracted value against the source text. Inter-coder reliability was subsequently assessed on a 10-paper subsample, independently double-coded by both authors: Cohen's $\kappa = 0.38$ for responsibility depth (60% raw agreement) and $\kappa = 0.67$ for responsibility mode (75% agreement, $n = 4$ jointly-coded papers—mode is coded only where a paper's depth is Constitutive or Operative, and since the two authors did not always agree on depth, the subset where both independently classified a paper as non-Peripheral is smaller than the full 10-paper sample), with the thirteen concept flags and six LLM-challenge flags showing 80–100% raw agreement across the subsample. The coding scheme was piloted on an earlier batch of approximately 10 papers, which surfaced insufficiently sharp category boundaries and led to a revised prompt version before the full corpus was coded. Full reliability results, including worked disagreement cases, are reported in Appendix B.8.

Unit of analysis and evidential interpretation. Each included paper was treated as one analytical unit for the descriptive counts and qualitative coding, regardless of whether it was published as a journal article or conference paper. This equal-counting approach supports analysis of the composition and prevalence of concepts within the literature; it does not imply that all studies provide evidence of equivalent quality or strength. In particular, the responsibility-depth and responsibility-mode categories describe how centrally and in what form responsibility is addressed, rather than the methodological quality of the study. A formal quality or risk-of-bias appraisal was not conducted because the review does not synthesise intervention effects or estimate pooled empirical outcomes, and the corpus contains heterogeneous contribution types for which a single appraisal instrument would not be appropriate. The coding framework also did not systematically distinguish conceptual frameworks, benchmark studies, prototypes, case reports, pilots, internal institutional applications, and independently evaluated operational deployments. Consequently, all reported frequencies should be interpreted as measures of the prevalence of concepts or forms of engagement in the literature, not as measures of the reliability, effectiveness, or evidential strength of the underlying claims.

Registration: This systematic review was retrospectively registered with the Open Science Framework (OSF): <https://doi.org/10.17605/OSF.IO/C4BVH> (accessed on 15 August 2026). All review stages—search, screening, full-text eligibility assessment, data extraction, coding, and synthesis—together with manuscript preparation, were complete before the registration was created; no methodological changes were made afterward. The registration therefore documents a completed review process rather than constraining a prospective one, and should be interpreted accordingly.

3. Descriptive Bibliometric Overview

3.1. Temporal Concentration and Field Emergence

Within the final systematic review corpus, publication activity concentrates heavily in the two most recent complete years, consistent with the wider diffusion of generative AI following the release of ChatGPT in late 2022. Three of the 109 included studies were published in 2023, the corpus's starting year; this rose to 19 in 2024 and to 71 in 2025—nearly two-thirds of the entire corpus (71 of 109, 65.1%)—with a further 16 already identified in the partial 2026 year (search conducted 20 April 2026, capturing under four months of that year's output). Figure 2 presents this distribution.

Because the corpus begins in 2023 and its earliest annual counts are small in absolute terms, this pattern is better read as evidence that responsible-LLM-in-finance research only became a substantively active focus of study from 2024 onward, rather than as a precisely estimated growth rate. The concentration of nearly two-thirds of all included studies in 2025 alone indicates a genuinely recent and, as of the partial 2026 count, still-accelerating field.

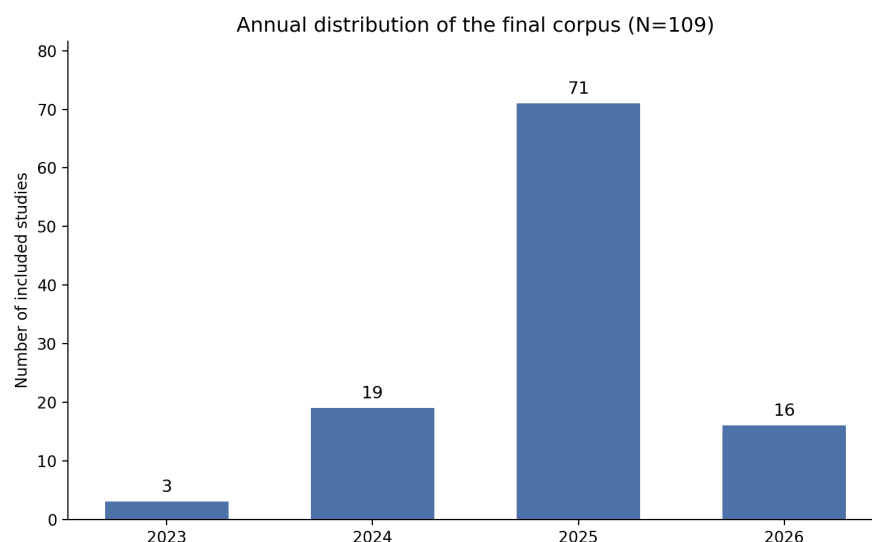


Figure 2. Annual distribution of the 109 studies included in the systematic review, by publication year (2023–2026). The 2026 count is partial, as the search was conducted on 20 April 2026.

3.2. Publication Sources and Disciplinary Structure

The 109 studies are distributed across 90 distinct publication venues, indicating a highly dispersed and bibliometrically unconsolidated field. The five most frequently represented venues account for only 18.3% of the corpus, and 81 of the 90 venues contribute only one paper. The ACM International Conference on AI in Finance (ICAIF) is the most concentrated venue, contributing eight papers (14.3% of conference output), signalling the emergence of a specialist research community at the finance–AI intersection. Among journals, *AI* leads with five articles, followed by *IT Professional*, *Digital Finance*, and *Expert Systems* with two each; finance-specialist outlets such as *Finance Research Letters* and *Financial Innovation* are represented but none is dominant. Table 2 summarises the leading publication venues. Figure 3 presents the disciplinary distribution of venues, confirming that the literature is shaped strongly by AI and computer science communities while gradually extending into finance, management, legal, governance, and interdisciplinary outlets—a pattern consistent with the field’s early stage of disciplinary consolidation [16].

Table 2. Leading publication venues in the final corpus.

Publication Type	Publication Venue	Papers	Share Within Type (%)
Conference	ACM International Conference on AI in Finance (ICAIF)	8	14.3
Conference	Joint FinNLP–FNP–LLMFinLegal Workshop	3	5.4
Conference	World Conference on Explainable Artificial Intelligence (xAI)	2	3.6
Conference	ACM Conference on Fairness, Accountability, and Transparency (FAccT)	2	3.6
Conference	International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)	2	3.6
Journal	<i>AI</i>	5	9.4
Journal	<i>IT Professional</i>	2	3.8
Journal	<i>Digital Finance</i>	2	3.8
Journal	<i>Expert Systems</i>	2	3.8
Journal	<i>Finance Research Letters</i>	1	1.9

Notes: “Share within type” is calculated as the number of papers published in a given venue divided by the total number of papers belonging to the same publication type, multiplied by 100. The corpus contains 56 conference papers and 53 journal articles. For example, the eight ICAIF papers represent 14.3% of all conference papers ($8/56 \times 100$). The database title “AI (Switzerland)” is shortened to “AI” for presentation.

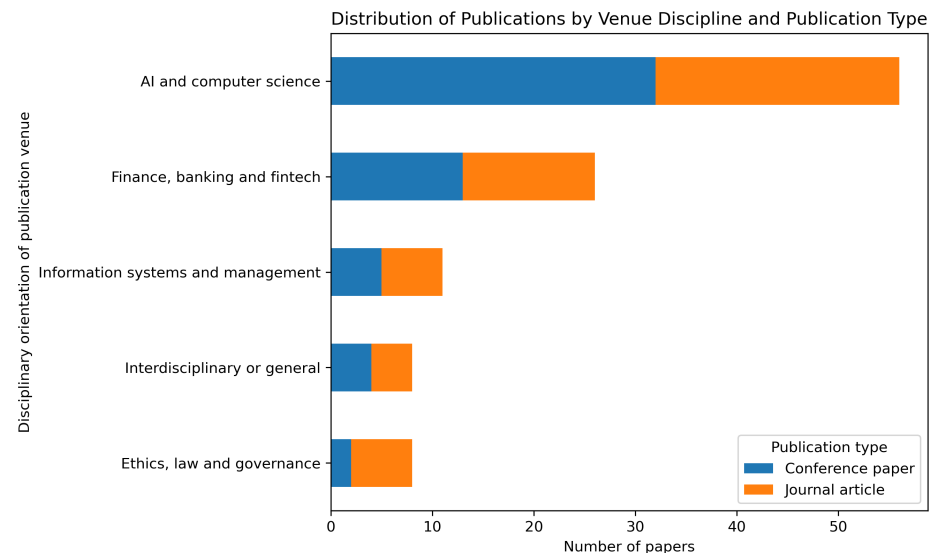


Figure 3. Distribution of publications by disciplinary orientation of the venue and publication type. Venue disciplines are assigned according to the primary scope of each journal or conference rather than the subject of the individual paper.

Publication venues were classified into five disciplinary categories according to the stated or commonly recognised scope of each journal or conference: artificial intelligence and computer science; finance, banking, and fintech; information systems and management; ethics, law, and governance; and interdisciplinary or general outlets. This venue-level classification follows the bibliometric practice of using source categories to characterise the disciplinary structure of a research field [18]. The classification was conducted at the venue level rather than the article level; therefore, the disciplinary label captures the primary orientation of the publication outlet rather than the substantive topic of each individual paper. Ambiguous venues were assigned to the category that best reflected their dominant publication scope.

As Figure 3 shows, the corpus is dominated by AI and computer science venues, while finance, management, legal, and governance outlets account for a smaller but growing share—a pattern that underscores both the technical origins of the field and its gradual extension toward institutional and regulatory concerns.

3.3. Geographic Distribution and Institutional Participation

The United States is represented in 39 papers (35.8%), followed by India with 20 papers (18.3%), China with 15 (13.8%), and the United Kingdom with 11 (10.1%). Despite this apparent US dominance, South Asia and East Asia together form a substantial component of the corpus, while Africa and Latin America remain markedly underrepresented [19]. Figure 4 presents the regional distribution. Of the 109 papers, 73 (67.0%) involve affiliations from a single country, while 36 (33.0%) involve international collaboration. Institutionally, 89 papers (81.7%) include at least one academic-institution affiliation; industry-only research accounts for 18 papers (16.5%); and explicit public-sector participation appears in only two papers (1.8%), both academic collaborations. This near-absence is substantively important: financial authorities are responsible for defining governance expectations, monitoring model and operational risk, and assessing systemic consequences of AI use [20], and their limited presence in the corpus is consistent with a literature that is richer in conceptual frameworks and technical prototypes than in regulatory-aligned evaluation and evidence from live financial deployment. Table 1 summarises the geographic and institutional characteristics.

As Figure 4 illustrates, North America, Europe, South Asia, and East Asia account for the bulk of the corpus; the limited representation of Africa, Latin America, and Southeast Asia raises questions about whether the resulting literature adequately captures the regulatory environments and institutional constraints of lower-represented financial systems.

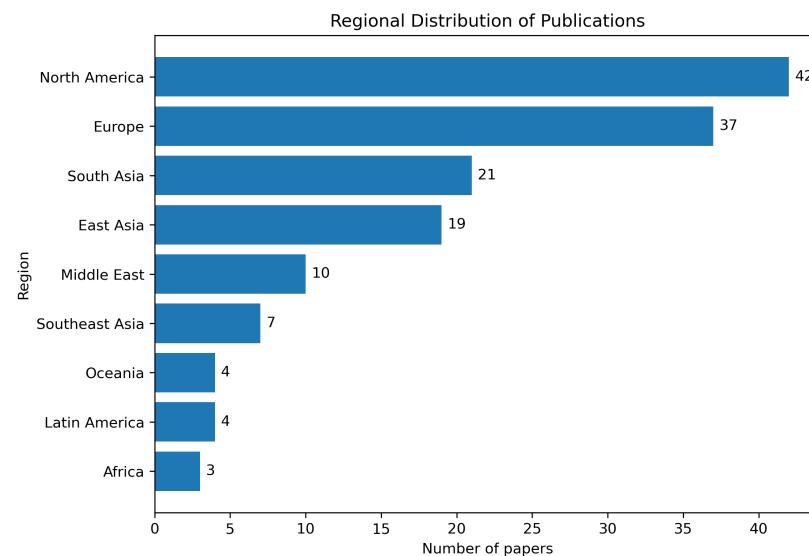


Figure 4. Regional distribution of publications based on author affiliations. Regional counts use full counting, so internationally collaborative papers may appear in more than one region.

As Table 1 shows, the near-absence of public-sector participation (1.8%) is the most policy-relevant institutional finding: it is consistent with, though it does not by itself establish, a gap between academic responsible AI frameworks and the governance requirements of financial regulators.

3.4. Thematic Structure

The corpus exhibits thematic heterogeneity rather than convergence around a unified responsible AI framework, as evidenced by the uneven distribution of responsibility concepts across thirteen dimensions. The responsibility concept flags capture the explicit textual presence of responsibility-related concepts within each paper rather than the depth of their treatment. A paper was flagged when a concept was explicitly referenced, regardless of whether it was introduced as contextual background, discussed conceptually, operationalized within the study, or empirically evaluated. Consequently, the frequencies reported in Table 3 should be interpreted as indicators of thematic salience across the literature rather than measures of the extent to which individual concepts are substantively developed.

Transparency (89.9%), bias (81.7%), and explainability (76.1%) are the most frequently referenced responsibility concepts, indicating their prominence within the responsible AI discourse on LLMs in finance. Trustworthiness (76.1%), ethics (72.5%), privacy (67.9%), and robustness (66.1%) also appear in a majority of the corpus, suggesting broad recognition of these concerns across diverse research contexts.

Institutional and procedural concepts—including fairness, accountability, governance, human oversight, security, and auditability—appear in between 46.8% and 63.3% of papers. Their comparatively lower prevalence suggests that these dimensions receive less consistent attention across the literature. However, these frequencies should not be interpreted as evidence that the concepts are equally well defined, operationalized, or empirically examined within individual studies. Rather, they indicate the relative visibility of responsibility themes across the corpus.

Taken together, the findings suggest that while a common vocabulary surrounding responsible AI has emerged in the LLM-for-finance literature, the extent to which individual responsibility concepts are translated into methodological frameworks, evaluation protocols, or governance mechanisms varies considerably. The thematic analysis therefore identifies the breadth of responsibility-related discourse rather than the maturity of research on each responsibility dimension.

Table 3. Prevalence of responsibility-concept flags across the 109-paper corpus. Flags indicate explicit textual presence and are counted without quality weighting; prevalence does not indicate the strength or validity of the supporting evidence.

Responsibility Concept	N	%
Transparency	98	89.9
Bias	89	81.7
Explainability	83	76.1
Trustworthiness	83	76.1
Ethics	79	72.5
Privacy	74	67.9
Robustness	72	66.1
Fairness	69	63.3
Accountability	69	63.3
Governance	62	56.9
Human Oversight	60	55.0
Security	58	53.2
Auditability	51	46.8

4. Responsibility in the Literature: Depth and Mode

The depth and mode categories were developed inductively, grounded in the observed structure of the corpus rather than adapted from a single pre-existing taxonomy. LLMs and finance originated as largely separate fields; as they converged, three distinct publishing patterns emerged that motivated the two-dimensional scheme. First, because finance is a heavily regulated industry, regulators and policy-oriented authors began publishing normative commentary on how LLMs should be governed in financial contexts—output for which responsibility is the entire contribution, motivating the Constitutive depth category and the Normative mode. Second, the broader literature follows a common, if not always sharp, academic/practitioner and theory/applied division, and much of it originates from authors primarily interested in applying LLMs to financial tasks, who then engage with responsibility only because the industry’s regulatory environment demands it—producing papers where responsibility is built into technical design (Operative depth, Technical mode) or is only briefly acknowledged (Peripheral depth). Third, a smaller set of papers go further to empirically test or audit these mechanisms, motivating the Evaluative mode, with the Mixed category capturing papers that substantially combine more than one mode. The categories were thus derived from recurring patterns in how responsibility actually appears across the corpus, rather than imposed from an external framework.

The 109 studies were coded by responsibility depth, resulting in 29 Constitutive, 46 Operative, and 34 Peripheral papers. Responsibility mode was assigned only to the 75 Constitutive and Operative papers because mode captures the form through which responsibility is substantively operationalised. Table 4 reports the resulting joint distribution, separately presenting the mode subtotal ($n = 75$) and the full-corpus depth total ($N = 109$).

Table 4. Joint distribution of responsibility depth and mode. Mode is assigned only to Constitutive and Operative papers ($n = 75$); Peripheral papers ($n = 34$) are classified by depth but not by mode. The full corpus contains 109 papers.

Responsibility Depth	Normative	Technical	Evaluative	Mixed	Depth Total
Constitutive	11	1	2	15	29
Operative	5	21	4	16	46
Mode subtotal	16	22	6	31	75
Peripheral (mode not assigned)	–	–	–	–	34
Full corpus depth total	Not applicable				109

4.1. Responsibility Depth: Constitutive, Operative, and Peripheral

Illustrative examples in this section were selected to represent the dominant pattern within each depth/mode cell, prioritising papers with the clearest textual grounding for their classification.

Across the 109 papers, 29 are Constitutive, 46 Operative, and 34 Peripheral.

Constitutive papers are those whose primary contribution is a responsibility-related output: a governance framework, a set of principles, an evaluation methodology, or a taxonomy of responsible AI concerns. Responsibility is not an objective that the research design serves; it is the research design's output. A representative example is Mandal et al. [21], who develop a six-pillar responsible AI governance framework for LLMs in banking operationalised across three deployment lifecycle stages, with specific benchmarks and evaluation metrics mapped to each pillar.

Operative papers incorporate responsibility concepts into the research design—as technical objectives, evaluation criteria, or analytical constructs—without responsibility itself being the paper's primary output. Operative papers are the most numerous in the corpus (46; 42%), reflecting the field's current centre of gravity in technical and applied work. Their concept-flag profiles are somewhat narrower than those of Constitutive papers (mean 8.3 flags vs. 10.9), though transparency (91%), trustworthiness (87%), and explainability (78%) dominate. One Operative example combining multiple challenge mitigations within a single architecture is Okpala et al. [22], who design a human-centred agentic framework for financial ML modelling in which hallucination, instruction failure, and output inconsistency are named and directly motivate four distinct guardrail mechanisms—temperature control, agent-task separation, human-in-the-loop orchestration, and a judge agent.

Peripheral papers invoke responsibility vocabulary only in framing or motivation without influencing methods or findings. These papers are not necessarily weak contributions—many advance LLM capabilities substantively—but their presence in a responsible AI review rests on framing rather than substance, illustrating responsibility discourse functioning as a legitimating device. Ali and Aysan [10] provide a wide-ranging overview of generative AI in financial services, invoking fairness, privacy, accountability, and ethics in framing sections without developing these concerns analytically.

4.2. Responsibility Mode: Normative, Technical, and Evaluative

For the 75 non-Peripheral papers, mode is distributed as follows: 16 Normative, 22 Technical, 6 Evaluative, and 31 Mixed.

Normative papers operationalise responsibility primarily through argument, principles, or governance framing. Aldasoro et al. [12] (BIS) articulate an eight-principle governance framework for AI in the financial system, drawing on US, EU, and Chinese regulatory positions, with the governance structure itself constituting the primary scholarly contribution.

Technical papers operationalise responsibility through system design, architectural choice, or algorithmic constraint. A critical analytical question is whether mechanisms are designed for LLM-specific generative properties—hallucination, prompt sensitivity, chain-of-thought instability—or whether pre-LLM frameworks are applied without modification (examined in Section 4.4). Zöller et al. [23] describe a live deployment at a German Landesbank whose responsible AI pillars are implemented as architectural mechanisms directly addressing hallucination, prompt manipulation, and data leakage as LLM-specific risks.

Evaluative papers operationalise responsibility through measurement, audit, or empirical assessment—the mode most directly concerned with the gap between stated responsibility intentions and actual model behaviour. This mode is underrepresented relative to Technical (six pure Evaluative papers). Lee et al. [24] evaluate sector, size, and momentum bias across six major LLMs in investment analysis tasks, finding that latent biases escalate into confirmation bias when models are presented with conflicting evidence. Biancotti et al. [25] (Banca d'Italia) take a different evaluative approach, prompting ten LLMs to impersonate the CEO of a financial institution facing an ethical dilemma, finding substantial variation in models' willingness to misappropriate customer assets—an empirical assessment of LLM alignment with financial ethical and legal standards.

Mixed papers combine two or more modes substantially. Mandal et al. [21] exemplify the Mixed mode, combining a normative governance framework with technical operationalisation and a prescribed evaluative component developed by practitioners at BNY Mellon.

4.3. Distribution and Cross-Dimensional Patterns

Several patterns in the joint distribution are analytically significant. First, Constitutive papers cluster strongly in Mixed and Normative modes: of 29 Constitutive papers, 15 are Mixed and 11 are Normative, with only 2 Evaluative and 1 Technical. This reflects the field's tendency to produce governance frameworks through argument or combined normative-technical design rather than empirical evaluation alone.

Second, Operative depth is dominated by Technical and Mixed papers: of 46 Operative papers, 21 are Technical—the largest single cell in Table 4—and 16 are Mixed, while only 5 are Normative and 4 Evaluative. This reflects the predominant mode of operationalising responsibility through system architecture and engineered guardrails.

Third, only 6 of the 75 papers assigned a responsibility mode are classified as purely Evaluative, making this the least common mode. This finding indicates that evaluation is comparatively uncommon as the principal form through which responsibility is operationalised. It does not imply that papers classified as Technical or Mixed contain no empirical elements.

Fourth, the Peripheral category accounts for nearly one-third of the corpus (34 of 109; 31%), indicating that a substantial share of the literature invokes responsibility as discursive framing rather than substantive operationalisation. Responsibility depth and LLM relevance are separate axes of classification: Peripheral describes how substantively a paper engages with responsibility concepts, not how substantively it engages with LLMs. All 109 included studies, Peripheral or not, satisfied the same LLM-relevance inclusion criterion described in Section 2.2 before being coded on this dimension.

4.4. LLM-Specific Challenges Across the Corpus

A central analytical question is whether the literature addresses challenges distinctive to LLMs as generative systems, or primarily applies frameworks developed for earlier AI systems. Table 5 reports both the mention and substantive addressing of six defined challenge categories.

Table 5. LLM-specific challenge flags across 109 corpus papers. “Mentioned” = explicit naming in the paper body; “Addressed” = substantive engagement in research design, methods, or findings. $N = 109$.

LLM-Specific Challenge	Mentioned	%	Addressed	%
Hallucination	51	46.8	44	40.4
Output inconsistency	27	24.8	22	20.2
Prompt manipulation	18	16.5	17	15.6
Instruction-following failure	7	6.4	7	6.4
Generative opacity	6	5.5	6	5.5
Emergent capabilities	6	5.5	4	3.7

Hallucination is the most widely engaged challenge, mentioned in 46.8% of papers and substantively addressed in 40.4%, reflecting both its practical salience in financial advisory contexts and the relative tractability of mitigation approaches—notably retrieval-augmented generation (RAG) and human-in-the-loop validation. Output inconsistency is second (24.8% mentioned; 20.2% addressed), treated primarily through architectural responses such as temperature control and ensembling.

The remaining four challenges are markedly less engaged. Prompt manipulation appears in only 16.5% of papers, generative opacity in 5.5%, instruction-following failures in 6.4%, and emergent capabilities in 5.5%. This concentration is analytically significant: the challenges receiving most attention are precisely those with accessible technical mitigations, while challenges that are comparatively harder to engineer around—generative opacity, emergent capabilities, and instruction-following failures—receive markedly less attention, though whether this reflects their technical difficulty, limited awareness, or the field’s early stage cannot be determined from the corpus alone. Generative opacity refers to the fundamental difficulty of explaining why a generative system produced a specific output rather than another: unlike discriminative models, LLMs generate outputs through processes whose internal logic resists direct inspection even in principle. Emergent capabilities refer to behaviours that appear at scale without being programmed or anticipated, creating responsibility challenges not captured by evaluation frameworks designed for pre-specified task performance. One plausible reading of the near-absence of both concepts from the corpus is that the literature’s responsible AI frameworks are currently calibrated to more tractable manifestations of LLM distinctiveness; terminological variation and publication lag given the field’s recency (Section 2.1) are equally plausible explanations, and the corpus data cannot adjudicate between these readings.

5. Structural Gaps

The coding reveals a literature whose current centre of gravity lies in Operative-Technical work: papers incorporating responsibility as a design constraint through explainability mechanisms, bias mitigation, and robustness testing. This distribution is consistent with the developmental trajectory of an emergent field, where conceptual groundwork and technical prototyping naturally precede institutional deployment and empirical validation.

5.1. Governance Without Operational Metrics

Transparency—the most prevalent concept in the corpus (90%)—is also among the least operationally specified, invoked to describe model interpretability, regulatory disclosure, output explainability, and process auditability interchangeably. Of the 29 Constitutive papers, 11 are Normative in mode: they articulate principles and governance structures without developing the operational metrics by which compliance could be assessed [12,26]. The industry-authored papers partially address this gap: Mandal et al. [21] specify eval-

uation benchmarks and monitoring cadences for each of their six responsible AI pillars, demonstrating that operational specification is achievable. Future research should develop institution-specific metrics specifying which transparency or accountability requirement applies, to whom, and under what conditions.

5.2. *The Explainability-Accountability Gap*

Explainability is the third most prevalent concept flag in the corpus (76%), and the majority of its operationalisations are technical: post hoc attribution methods such as SHAP [27] and LIME [28], counterfactual explanation [29], attention visualisation [30], prompted chain-of-thought rationales [31], and RAG-based reference trails [32]. These mechanisms, however, have different explanatory properties. Chain-of-thought prompting can produce a human-readable account of intermediate reasoning and may support user scrutiny, error detection, and interaction with the model. Nevertheless, the generated reasoning trace is itself a model output rather than a direct observation of the model's internal computational process. It may omit factors that influenced the answer, provide a post hoc rationalisation, or present a plausible but unfaithful explanation [33]. Chain-of-thought should therefore be treated as a user-facing rationale or diagnostic aid rather than as inherently faithful model explainability unless its faithfulness is independently evaluated.

These approaches can improve the transparency or traceability of model outputs, but they do not straightforwardly satisfy the forms of explanation required by financial regulation. Vilone et al. [32] demonstrate that current LLM-based robo-advisory systems do not generate MiFID II-compliant explanations under a framework of seven compliance questions derived from regulatory text, even when prompted with explicit explainability objectives. A related gap concerns auditability (46.8% of papers): in regulatory contexts, auditability requires retrospective verification of model behaviour—a requirement creating particular challenges for LLMs whose probabilistic generation means exact reconstruction of a prior output may be impossible. Future research should develop regulatory-aligned evaluation frameworks connecting XAI outputs to MiFID II, SR 11-7, and equivalent financial compliance requirements.

5.3. *The Need for Regulatory Alignment*

Financial applications of LLMs operate within regulatory environments that extend beyond general responsible AI principles. Relevant external instruments include the EU AI Act [34], the Digital Operational Resilience Act (DORA) [35], the Markets in Financial Instruments Directive (MiFID II) [36], the model-risk-management guidance in SR 11-7 [37], and the Monetary Authority of Singapore's FEAT principles [38].

These instruments serve distinct regulatory and governance purposes and should not be treated as interchangeable. The EU AI Act establishes a horizontal, risk-based framework governing the development, provision, and use of AI systems across sectors. DORA focuses specifically on the management of information and communication technology risk and the digital operational resilience of financial entities. MiFID II regulates investment services and markets, with particular emphasis on investor protection, suitability, conduct, transparency, and disclosure. SR 11-7 establishes supervisory expectations for model development, implementation, validation, governance, and model-risk management. MAS FEAT provides financial institutions with principles for the fair, ethical, accountable, and transparent use of AI and data analytics. Although these instruments may overlap in areas such as governance, accountability, risk management, and oversight, they address different objects, institutions, and regulatory objectives.

The present coding framework did not include separate flags for references to individual regulatory instruments. Accordingly, this review does not estimate how frequently the

EU AI Act, DORA, MiFID II, SR 11-7, MAS FEAT, or other specific frameworks appear in the corpus, nor does it classify whether a paper merely cites an instrument, maps its requirements, proposes operational implementation, or conducts empirical compliance evaluation. We therefore do not present regulatory convergence or divergence as a quantitatively established corpus-level finding.

Instead, systematic comparison between responsible LLM principles and established financial regulatory requirements is identified as an important direction for future research. Such research could distinguish among citation, conceptual alignment, compliance mapping, operational implementation, and empirical evaluation, while examining how responsible AI principles correspond to, complement, or conflict with the distinct requirements already applicable to financial institutions.

5.4. *Limited Evaluative Orientation*

The responsibility-mode distribution indicates that a purely Evaluative orientation remains uncommon. Of the 75 Constitutive and Operative papers assigned a responsibility mode, 6 are classified as Evaluative, compared with 16 Normative, 22 Technical, and 31 Mixed papers. Evaluation is therefore less frequently the principal form through which responsibility is operationalised than normative argument, technical design, or a combination of approaches.

The coding framework did not separately record whether a study evaluated a benchmark, prototype, pilot application, or live institutional deployment. Consequently, this review does not make quantitative claims regarding the prevalence of these evaluation settings or the number of real-world deployments in the corpus. Future reviews could extend the present framework by coding evaluation setting, deployment status, institutional context, evaluation metrics, and evidence of operational performance. These variables would enable a more systematic assessment of how responsible LLM mechanisms are evaluated across research and operational settings.

5.5. *The LLM-Specificity Gap*

The most conceptually significant structural gap is the limited engagement with challenges distinctive to LLMs as generative systems. The dominant responsibility vocabulary—transparency, bias, fairness, accountability—was established in the pre-LLM responsible AI literature and has been carried into the LLM-in-finance context without systematic examination of whether the same concepts and mechanisms apply. Hallucination is largely treated as a reliability problem addressed through RAG and human oversight: the distinctive epistemological character of hallucination in LLMs—that outputs can be syntactically fluent, contextually plausible, and factually wrong in ways difficult for users to detect—is rarely engaged analytically, despite the potential for material harm in financial contexts. Generative opacity and emergent capabilities are each mentioned in only 5.5% of papers (6 of 109), with emergent capabilities substantively addressed in just 3.7% (4 of 109); Section 4.4 discusses plausible, non-exclusive explanations for this low prevalence. Governance concepts adequate to LLM-specific challenges—generative opacity, emergent capabilities—that are genuinely distinct from pre-LLM responsible AI frameworks represent the field's most significant conceptual frontier. Table 6 summarises all five structural gaps, their evidential basis, and the research directions they motivate.

Table 6. Key analytical gaps and research needs identified in the responsible LLM-in-finance literature, their basis in the present review, and associated research directions. $N = 109$.

Analytical Gap or Research Need	Basis in the Present Review	Research Direction
Governance without operational metrics	Eleven papers are classified as Constitutive–Normative, indicating that governance principles and normative argument constitute their principal mode of engagement rather than evaluation.	Develop institution-specific metrics specifying which transparency, accountability, and oversight requirements apply, to whom, and under what conditions.
Explainability–accountability gap	Explainability appears in 83 papers (76.1%), accountability in 69 (63.3%), and auditability in 51 (46.8%). The presence of these concepts does not, however, establish alignment with financial explanation and accountability requirements.	Develop regulatory-aligned evaluation frameworks connecting technical explanation mechanisms to applicable financial accountability, disclosure, suitability, and governance requirements.
Need for regulatory alignment	The EU AI Act, DORA, MiFID II, SR 11-7, and MAS FEAT address distinct regulatory purposes. Because references to individual instruments and forms of regulatory engagement were not separately coded, their prevalence and use in the corpus are not quantified.	Systematically distinguish among citation, conceptual alignment, compliance mapping, operational implementation, and empirical evaluation when examining the relationship between responsible LLM principles and financial regulation.
Limited Evaluative orientation	Only 6 of the 75 Constitutive and Operative papers assigned a responsibility mode are classified as purely Evaluative. Evaluation setting and deployment status were not coded as separate variables.	Future studies and reviews should distinguish among benchmark evaluation, prototypes, pilots, internal institutional use, customer-facing production systems, and independently evaluated operational deployments.
LLM-specificity gap	Generative opacity is mentioned in only 5.5% of corpus papers; emergent capabilities are mentioned in 5.5% but substantively addressed in only 3.7%.	Develop governance concepts and evaluation methods addressing challenges arising specifically from the generative and emergent properties of LLMs, rather than relying only on frameworks inherited from pre-LLM responsible AI.

5.6. Limitations of Evidence Appraisal

The present review characterises how responsibility is represented and operationalised in the literature, but it does not provide a formal appraisal of methodological quality or evidential strength. Journal articles and conference papers, as well as conceptual, technical, evaluative, and mixed contributions, are counted as individual analytical units. Equal counting is appropriate for describing the distribution of concepts and modes of engagement, but it does not establish that these contributions provide equivalent evidence.

Accordingly, the prevalence of a responsibility concept should not be interpreted as evidence that the associated principle or mechanism has been validated, implemented successfully, or shown to improve outcomes. A concept may be widely discussed while being supported primarily by normative argument or technical proposals rather than independent empirical evaluation. Conversely, a less frequently discussed concept may be supported by a smaller but methodologically stronger body of evidence.

Future evidence syntheses should extend the coding framework by classifying research design and evidence setting, including conceptual frameworks, benchmark evaluations, prototypes, industry case reports, pilots, internal institutional applications, customer-facing production systems, and independently evaluated operational deployments. These contribution types could then be assessed using design-appropriate quality criteria rather than a single quality scale applied across the heterogeneous corpus.

A related limitation concerns the precision implied by Table 4's cell-level counts. These counts derive from the single, LLM-assisted coding pass described in Section 2.3, not from the reliability subsample. The reliability exercise reported in Appendix B.8 found only

moderate agreement on responsibility depth (Cohen's $\kappa = 0.38$ on a 10-paper subsample), indicating genuine interpretive uncertainty in depth classification at the level of individual papers. This uncertainty should temper how precisely individual cells in Table 4 are read, though it does not necessarily imply a comparable bias in the corpus-wide proportions, since the observed disagreements ran in both directions rather than systematically toward one category.

6. Conclusions

The gaps identified in Section 5 converge on research directions organised across three levels. At the empirical level, future research should evaluate responsible LLM mechanisms across clearly documented settings, including benchmarks, prototypes, pilot applications, and operational financial environments. Such studies should report their institutional context, evaluation criteria, governance mechanisms, operational performance, and observed failure modes. At the conceptual level, the field requires rigorous development of the distinction between general responsible AI and challenges specific to LLMs as generative systems—with governance frameworks adequate to generative opacity and emergent capabilities, and interpretability methods connected to the explanation standards required by financial regulation [32]. At the institutional level, the financial regulatory concept of model risk management (MRM) provides a structured framework already operationalised in practice; research connecting MRM concepts to LLM-specific challenges and comparative analysis across regulatory jurisdictions (EU AI Act, MAS FEAT, US SR 11-7, UK FCA guidance) would close an important gap between academic and regulatory discourse.

This paper has provided a structured review of 109 studies on responsible large language models in finance, combining a descriptive bibliometric overview with qualitative content analysis and a two-dimensional coding framework that distinguishes responsibility depth (Constitutive, Operative, Peripheral) from responsibility mode (Normative, Technical, Evaluative, Mixed). The literature's centre of gravity lies in Operative-Technical work, incorporating responsibility as a design constraint through explainability mechanisms, bias mitigation, and robustness testing. Constitutive contributions are concentrated in Normative and Mixed modes rather than Evaluative ones, indicating a field stronger at articulating governance principles than at specifying the metrics by which those principles can be assessed. A purely Evaluative orientation is comparatively uncommon in the coded corpus. However, because evaluation setting and deployment status were not coded as separate variables, the review does not quantify the distribution of benchmark, prototype, pilot, and operational evaluations. The LLM-specificity analysis reveals a further limitation: hallucination and output inconsistency receive meaningful attention through architectural mitigations that treat them as reliability problems, while generative opacity, emergent capabilities, and instruction-following failures—the challenges most directly implicating the distinctive properties of LLMs as generative systems—are mentioned, respectively, in only 5.5%, 5.5%, and 6.4% of corpus papers, with emergent capabilities substantively addressed in just 3.7%. As discussed in Section 4.4, this may reflect the field's current calibration to more tractable manifestations of LLM distinctiveness, though terminological variation and publication lag are equally plausible explanations that the corpus data cannot distinguish between.

These findings have implications for researchers and practitioners alike. For researchers, the systematic gaps in evaluative and deployment evidence, in operational metrics for normative governance frameworks, and in LLM-specific challenge engagement identify concrete directions for the field's development. For practitioners and regulators, the divergence between academic responsible AI discourse and the model risk management requirements of financial regulation suggests that closer integration between these gover-

nance traditions is needed to produce frameworks that are both intellectually grounded and institutionally operable. Closing the gap between high-level responsible AI principles and the practical requirements of accountability, governance, and deployment in financial applications of LLMs is not a marginal research agenda: it is one of the central challenges facing financial AI scholarship in the years ahead.

These findings concern the prevalence and form of responsibility discourse within the reviewed literature rather than the comparative quality of the evidence supporting individual claims. High concept prevalence should therefore be interpreted as evidence of scholarly attention, not as evidence that the corresponding governance principle, technical mechanism, or evaluation approach has been independently validated or shown to be effective.

The bibliographic searches were completed on 20 April 2026. The 2026 literature is therefore incomplete, and relevant studies published or indexed after this date are not included in the corpus. Consequently, the 2026 publication count should be interpreted as a partial-year observation rather than a complete annual total.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/fintech5030071/s1>, File S1: Data and Code.zip, a reproducibility archive containing: (i) the complete listing of the 109 studies included in the systematic review; (ii) cleaned bibliographic data, venue-discipline classifications, and Jupyter notebooks used for the descriptive bibliometric analysis; (iii) the coding manual, coding prompt, and paper-level coding data; (iv) independent coding sheets and calculation workbooks used for the inter-coder reliability analysis; (v) Python scripts and output files for the parameter- and weight-sensitivity analyses; and (vi) the Web of Science–Scopus data-processing, deduplication, and semantic-screening pipeline.

Author Contributions: The conceptualization, methodology, validation, formal analysis, investigation, and writing of the paper were completed by C.H.T. and Q.D. equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used Claude (Anthropic, Claude Sonnet 5, 2026) to assist with structured data extraction from source literature and with drafting and revising portions of the manuscript text. The authors have reviewed, verified, and edited all AI-assisted content and take full responsibility for the accuracy and integrity of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Semantic Screening Procedure

This appendix provides full technical details of the semantic screening procedure used to prioritise records for manual review, as referenced in Section 2.2 of the main paper.

Appendix A.1. Document Representation

Each document was represented using a unified textual field constructed from its title, abstract, and keywords. To reflect the higher informational content of titles and keywords relative to abstracts, these fields were mildly upweighted by repeating each once in the unified text representation, while the abstract appeared once. The resulting text was encoded into dense vector representations using Sentence-BERT [15], specifically

the all-mpnet-base-v2 model accessed via the sentence-transformers library (version 5.4.1), which provides strong performance for semantic similarity tasks.

Appendix A.2. Composite Relevance Score

Semantic similarity scores were computed for each document using Sentence Transformer embeddings. Documents were compared against eight query prototypes representing responsible AI in finance (e.g., “Responsible artificial intelligence in finance”; “AI governance, explainability, bias, privacy and security in financial services”; “Large language models and generative AI governance in finance”), together with separate concept probes capturing responsibility, artificial intelligence, and finance dimensions individually. Lexical coverage scores were additionally calculated based on the presence of domain-specific keyword lists for each dimension, serving as a robustness check to ensure that semantically relevant documents also contain explicit textual evidence of the corresponding concepts.

The composite relevance score for document i is defined as:

$$S_i = 0.45 \cdot \text{SemanticMax}_i + 0.20 \cdot \text{SemanticMean}_i + 0.15 \cdot \text{SemanticResponsible}_i + 0.05 \cdot \text{SemanticAI}_i + 0.05 \cdot \text{SemanticFinance}_i + 0.05 \cdot \text{LexResponsible}_i + 0.025 \cdot \text{LexAI}_i + 0.025 \cdot \text{LexFinance}_i \quad (\text{A1})$$

where SemanticMax captures the strongest semantic match to any single query prototype, SemanticMean represents the average similarity across all prototypes, Semantic{Responsible, AI, Finance} capture dimension-specific similarity, and Lex{Responsible, AI, Finance} capture lexical coverage across the three conceptual domains.

Appendix A.3. Penalty Term

To ensure that retained documents reflect the intersection of all three conceptual domains, a penalty was applied to records that do not satisfy minimum conceptual coverage:

$$S_i^p = \begin{cases} S_i, & \text{if the document exhibits signals for responsibility, AI, and finance;} \\ S_i - 0.08, & \text{otherwise.} \end{cases} \quad (\text{A2})$$

Appendix A.4. Percentile-Based Threshold

The final filtering threshold was defined as the 95th percentile of the penalised score distribution:

$$\tau = Q_{0.95}(S_i^p) \quad (\text{A3})$$

Documents satisfying $S_i^p \geq \tau$ were retained for manual screening ($n = 151$). The use of a quantile-based criterion improves robustness by adapting to the empirical distribution of similarity scores, reducing sensitivity to scale variations across embedding outputs. The 95th percentile threshold reflects a balance between precision and coverage, retaining a manageable high-priority subset while substantially reducing the manual screening burden. The threshold was validated through targeted inspection of sampled records to confirm alignment with the conceptual scope of the review.

Appendix A.5. Recall Validation

The 95th percentile threshold and the 0.08 penalty term were evaluated empirically against the risk that relevant records were excluded before manual screening began, using three checks.

First, a stratified recall audit of 60 records was conducted, combining two sub-samples: a percentile-stratified sample of 35 records concentrated on the bands immediately below the retention threshold—the population at highest risk of false exclusion—and a targeted sample of 25 records excluded specifically by the conjunctive (AND) LLM-proximity filter described in the next subsection. Each sampled record was manually assessed for eligibility using the same criteria applied during full-text screening, without reference to its screening score. Of 15 records sampled from the 90th–95th percentile band, 1 was subsequently judged eligible; of 10 records from the 70th–90th percentile band, 1 was eligible; of 10 records sampled below the 70th percentile, none were eligible; and of the 25 records excluded by the conjunctive LLM-proximity filter, none were eligible. Pooled across all 60 records, 2 (3.3%) were judged eligible. This indicates that the 95th percentile threshold did not measurably exclude eligible studies immediately below the cutoff. Because this audit deliberately concentrated sampling on the populations at highest risk of false exclusion rather than sampling the full deduplicated pool ($n = 23,934$) uniformly, this pooled rate should not be extrapolated to estimate a population-wide count of missed eligible studies.

Second, a sensitivity check varied the penalty term across the range 0.05–0.12; this altered the final screened set by at most one record, indicating that 0.08 is not a precariously tuned parameter.

Third, a sensitivity check varied the LLM-proximity percentile cutoff described in the next subsection: a looser 60th percentile cutoff preserved every record retained under the 70th percentile cutoff while adding further candidates, and a stricter 80th percentile cutoff would have excluded 8 of the 151 records retained through semantic screening.

Appendix A.6. LLM Proximity Filter

Records prioritised in the first stage were further assessed for LLM engagement. Each record was scored against seven LLM-focused query prototypes (e.g., “large language models in finance”; “responsible use of large language models in finance”; “trustworthy generative AI in banking and financial services”) using the same all-mpnet-base-v2 model. A record was retained for manual screening if its maximum similarity to any LLM prototype met or exceeded the 70th percentile of that distribution, and if its document text contained at least one LLM-specific term (“large language model”, “LLM”, “generative AI”, “foundation model”, or equivalent). This second-stage filter ensured that prioritised records were relevant both to responsible AI in financial contexts and to LLMs specifically, consistent with the broad-to-narrow review strategy described in Section 2.1 of the main paper.

Table A1 summarises the key variables used across both stages of the screening procedure.

Table A1. Key variables used in the semantic screening procedure.

Variable	Role	Interpretation
Stage 1: Responsible AI in Finance Scoring		
combined_score_penalized	Main ranking score	Overall relevance score after applying a 0.08 penalty to records that do not cover all three conceptual dimensions.
trriage_label	Screening priority	Classifies records into priority groups based on percentile thresholds (95th, 90th, 70th).
passes_minimum_coverage	Coverage constraint	Indicates whether the document contains lexical signals for responsible AI, AI technologies, and financial applications.
semantic_max	Strongest match	Maximum cosine similarity to any of the eight responsible AI in finance query prototypes.

Table A1. *Cont.*

Variable	Role	Interpretation
semantic_mean	Average fit	Mean cosine similarity across all eight query prototypes.
semantic_responsible	Responsibility relevance	Similarity to a concept probe covering fairness, transparency, governance, and accountability.
semantic_ai	AI relevance	Similarity to a concept probe covering machine learning, deep learning, and LLMs.
semantic_finance	Finance relevance	Similarity to a concept probe covering banking, lending, trading, and fintech.
lex_responsible, lex_ai, lex_finance	Lexical validation	Proportion of dimension-specific keyword list matched in the document text.
Stage 2: LLM Proximity Filter		
semantic_llm_max	LLM match	Maximum cosine similarity to any of the seven LLM-focused query prototypes.
has_llm_signal	LLM term flag	Boolean: document text contains at least one LLM-specific term (“large language model”, “LLM”, “generative AI”, “foundation model”, or equivalent).

Appendix B. Coding Manual for Human Raters (Inter-coder Reliability Assessment)

This appendix sets out the coding rules to be applied when independently rating articles from the review corpus, for the purpose of assessing inter-coder reliability on the two-dimensional responsibility taxonomy—responsibility depth and responsibility mode—together with the supporting concept and LLM-challenge flags described below. The rules mirror those used for the first-pass extraction reported in the main text, so that ratings produced by different coders, and by the original extraction, are directly comparable.

Coders should code each assigned article independently, without consulting other raters or the original extracted values, until instructed otherwise by the research team. Every judgment should be based on the article body only—abstract, introduction, methods, results, discussion, and conclusion—and not on the References section.

Appendix B.1. How to Use This Manual

1. Read the full text of the assigned article.
2. Work through Appendixes B.2–B.5 below in order, recording a value for every field on the coding sheet (Appendix B.6).
3. For *responsibility_depth* and *responsibility_mode* in particular, note a brief supporting quotation or paraphrase and its location (e.g., “Section 4, p. 10”)—this will be used to resolve any disagreement between raters.
4. If genuinely unsure between two categories, choose the one considered more likely and flag it explicitly in the notes so it can be discussed.

Appendix B.2. Bibliographic and Context Fields

Coders record the source type and implementation orientation directly from what the paper states.

- Source—“journal”, “conference”, “SSRN”, or “working paper”, based on where the piece was published or posted.
- Implementation orientation—based on what the paper explicitly describes in its methods or system description:

- Conceptual—no system is built or tested.
- Prototype—a system is built and tested under controlled or laboratory conditions.
- Real-world Deployment—a system is operating in a live institutional setting.

Appendix B.3. Responsibility Concept Flags

For each of the thirteen concepts in Table A2, coders score 1 if the concept is explicitly named in the article—the term itself or one of its listed synonyms must appear—and 0 if it is absent or only implied. A concept is not scored as present because it seems conceptually related; the word or a listed synonym must actually appear in the text.

Table A2. Responsibility concept flags: recognised terms.

Concept	Score 1 If Any of These Appear
Fairness	fairness, fair, equitable treatment, non-discrimination
Transparency	transparency, transparent
Explainability	explainability, explainable, interpretability, interpretable
Accountability	accountability, accountable
Auditability	auditability, auditable, audit trail
Governance	governance, oversight framework, regulatory framework
Bias	bias, biased, demographic bias, algorithmic bias
Robustness	robustness, robust, stability, resilience
Privacy	privacy, data privacy, private data, confidentiality
Security	security, secure, cybersecurity, data security
Trustworthiness	trustworthiness, trustworthy, trust
Human oversight	human oversight, human-in-the-loop, human supervision, human control
Ethics	ethics, ethical, ethical AI

Appendix B.4. LLM-Specific Challenge Flags

Each of the six challenges in Table A3 is scored on two dimensions. First, “mentioned”: score 1 if the article explicitly uses the term or a listed synonym, 0 if absent. Second, “addressed”: score 1 only if the article goes beyond simply naming the challenge to engage with it substantively—for example, discussing it as a concern in some depth, evaluating it empirically, proposing a mitigation, or theorising its implications. If “mentioned” is 0, “addressed” is automatically 0.

Table A3. LLM-specific challenge flags: recognised terms.

Challenge	Score “Mentioned = 1” If Any of These Appear
Hallucination	hallucination, confabulation, factual error, fabrication, factually incorrect output
Prompt manipulation	prompt injection, adversarial prompt, jailbreak, prompt attack, prompt manipulation
Generative opacity	generative opacity, chain-of-thought instability, unverifiable generation, black-box generation. “Explainability” or “interpretability” alone do not qualify—the term must specifically reference the non-inspectability of generative reasoning.
Instruction-following failure	instruction following, alignment failure, misalignment, instruction drift, alignment instability
Output inconsistency	output inconsistency, sensitivity to prompt phrasing, context sensitivity, stochastic output, prompt sensitivity
Emergent capabilities	emergent capabilities, emergent behaviour, unpredictable capabilities, capabilities at scale

As a calibration anchor, an article that lists “hallucination” once among a set of generic risks in its introduction, without returning to it, should be scored mentioned = 1,

addressed = 0. An article that names hallucination and then designs a specific mitigation for it—such as a retrieval step, a verification pass, or a guardrail explicitly motivated by the risk of hallucination—should be scored addressed = 1.

Appendix B.5. Responsibility Depth and Mode

These two fields are the primary focus of the reliability assessment. The decision procedure below is applied in order; mode is not assessed before depth is settled.

Appendix B.5.1. Step 1—Responsibility Depth

Coders ask, in order:

- Is responsibility the article's primary contribution? That is, is the paper's main output itself a responsibility-related artefact—a governance framework, a set of principles, an evaluation methodology, or a taxonomy of responsible AI concerns? If yes, code Constitutive.
- If not Constitutive, are responsibility concepts built into the research design? That is, do concepts such as fairness, transparency, bias, or auditability function as technical objectives, evaluation criteria, or analytical constructs within the paper's methods—even though responsibility itself is not the paper's primary output? If yes, code Operative.
- If neither applies, code Peripheral. Responsibility vocabulary appears only in framing or motivation—for example, in the introduction or discussion—without shaping the methods or findings.

Clarifying note: Where a paper is structured as a general survey but contains one or more dedicated subsections presenting a named governance or risk-mitigation framework, treat that subsection as constituting the paper's primary output (Constitutive) only if it presents a distinct structured contribution—named pillars, stages, or a risk-to-mitigation matrix—rather than a narrative synthesis of frameworks already established elsewhere in the literature. This distinction is illustrated by the worked cases in Appendix B.8.

Worked anchors (see the main text for full citations):

- Constitutive—a paper whose central contribution is a governance framework with defined pillars and evaluation benchmarks mapped to each pillar.
- Operative—a paper that designs an agentic system for a financial task and names hallucination, instruction failure, and output inconsistency as risks that directly motivate specific guardrail mechanisms—responsibility is engineered into the design, but the paper's primary contribution is the system itself.
- Peripheral—a broad survey of generative AI in financial services that invokes fairness, privacy, accountability, and ethics in its framing sections without developing them analytically.

Appendix B.5.2. Step 2—Responsibility Mode

This field is coded only if depth was coded Constitutive or Operative in Step 1. Coders ask which of the following best characterises how responsibility is operationalised:

- Normative—primarily through argument, principles, or governance framing (e.g., an eight-principle governance framework built from regulatory positions).
- Technical—primarily through system design, architecture, or algorithmic constraint (e.g., architectural mechanisms built to directly address hallucination, prompt manipulation, or data leakage).

- Evaluative—primarily through measurement, audit, or empirical assessment (e.g., systematically testing several models for a specific bias or failure mode and reporting the results).
- Mixed—two or more of the above are substantially present, rather than one clearly dominating (e.g., a governance framework that also specifies technical mechanisms and a practitioner-developed evaluation component).

Where two modes can be justified with roughly equal textual support, coders should code Mixed rather than forcing a single-mode choice—but only where both are substantial, not where one is a passing mention.

Appendix B.6. Coding Sheet

Table A4 summarises every field to be completed for each assigned article.

Table A4. Coding sheet template.

Field	Value Options	Notes
paper_id	—	Assigned by the research team; not altered by coders.
source	journal/conference/SSRN/working paper	
implementation_orientation	Conceptual/Prototype/Real-world Deployment	See Appendix B.2.
13 concept flags	0/1 each	See Appendix B.3 and Table A2.
6 challenge-mentioned flags	0/1 each	See Appendix B.4 and Table A3.
6 challenge-addressed flags	0/1 each	Scored 0 automatically if the corresponding mentioned flag is 0.
responsibility_depth	Constitutive/Operative/Peripheral	See Appendix B.5, Step 1. Primary reliability target.
responsibility_mode	Normative/Technical/Evaluative/Mixed/(blank)	Only coded if depth is Constitutive or Operative. Primary reliability target.
Evidence/location note	Free text	One short quotation or paraphrase with location (e.g., “Section 3, p. 6”) per judgment field, for reconciliation.

Appendix B.7. Reconciliation Procedure

Once all assigned articles have been coded independently, ratings are compared across coders. Inter-coder agreement is calculated for each field, with particular attention to *responsibility_depth* and *responsibility_mode*. Where two or more coders disagree, the coders discuss the specific article using their recorded evidence notes and reach a reconciled value through discussion; where agreement cannot be reached, the disagreement and both original codes are retained and reported alongside the reliability statistics.

Appendix B.8. Reliability Outcomes and Worked Boundary Cases

On a 10-paper subsample independently double-coded by both authors, Cohen’s $\kappa = 0.38$ for responsibility depth (60% raw agreement, $n = 10$) and $\kappa = 0.67$ for responsibility mode (75% agreement, $n = 4$ papers coded non-blank by both raters). The thirteen concept flags and six LLM-challenge flags (mentioned and addressed) showed 80–100% raw agreement across the subsample; several individual flags returned κ near zero despite high raw agreement, a known statistical property of kappa when a category is

scored identically on nearly every paper—chance agreement is then already high, leaving little room for kappa to register genuine agreement—rather than a sign of unreliable coding.

Four of the ten papers produced a depth disagreement, summarised in Table A5. Three of the four (Papers 3, 29, 31) share a common pattern: a paper structured as a general survey of AI or LLMs in finance, containing one or more dedicated subsections presenting a named governance or risk-mitigation framework. One coder weighted the survey framing as dominant; the other weighted the embedded framework section as the paper’s substantive contribution. This recurring ambiguity motivated the clarifying addition to the Step 1 decision rule in Appendix B.5 above.

Table A5. Worked depth-classification disagreements from the reliability subsample.

Paper	Coder A	Coder B	Basis of Disagreement
2	Constitutive	Operative	Coder A’s note cited a location only, without a supporting quotation; Coder B pointed to responsibility concepts built directly into the research design and evaluation criteria (Table 2). The case illustrates why evidence notes must include a quotation or paraphrase, not a location alone.
3	Peripheral	Constitutive	Coder A read the paper as a general survey of AI in finance; Coder B identified a dedicated section, “Operationalizing regulatory principles for AI,” mapping regulatory considerations across the AI value chain (design, training and evaluation, deployment, long-term diffusion) as the paper’s substantive contribution.
29	Peripheral	Constitutive	Coder A read the paper as a broad survey invoking responsibility concepts only in framing; Coder B identified a named governance structure (the “High-Trust Artificial Intelligence Framework” and “AI Governance Pyramid”) presented in a dedicated “Strategic Recommendations” section as the paper’s primary output.
31	Operative	Peripheral	Coder A treated a risk-to-mitigation matrix and five-level adoption framework as responsibility functioning as an analytical and decision construct; Coder B treated the same material as a conceptual catalogue within a broader applications survey, not engineered into an empirical research design.

These cases were resolved through discussion following the procedure in Appendix B.7, applying the clarified Step 1 rule above. Notably, all four disagreements were resolved within the taxonomy’s existing categories—no case required a classification the framework did not already provide—and three of the four traced to a single identifiable boundary condition, addressed directly by the Step 1 clarification above, rather than to diffuse or unexplainable disagreement across the coding scheme.

References

1. Dong, Y.; Wu, F.; Zhang, K.; Dai, Y.; Zhang, S.; Ye, W.; Chen, S.; Cheng, Z.Q. Large Language Model Agents in Finance: A Survey Bridging Research, Practice, and Real-World Deployment. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 5–9 November 2025; pp. 17889–17907. [\[CrossRef\]](#)
2. Lee, D.K.C.; Guan, C.; Yu, Y.; Ding, Q. A comprehensive review of generative AI in finance. *FinTech* **2024**, *3*, 460–478. [\[CrossRef\]](#)
3. Ding, Q.; Ding, D.; Wang, Y.; Guan, C.; Ding, B. Unraveling the landscape of large language models: A systematic review and future perspectives. *J. Electron. Bus. Digit. Econ.* **2024**, *3*, 3–19. [\[CrossRef\]](#)
4. Yeo, W.J.; Van Der Heever, W.; Mao, R.; Cambria, E.; Satapathy, R.; Mengaldo, G. A Comprehensive Review on Financial Explainable AI. *arXiv* **2023**, arXiv:2309.11960. [\[CrossRef\]](#)
5. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [\[CrossRef\]](#)
6. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
7. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2022; Volume 35, pp. 27730–27744.
8. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* **2022**. [\[CrossRef\]](#)
9. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large language models: A survey. *arXiv* **2024**, arXiv:2402.06196.
10. Ali, H.; Aysan, A.F. Navigating the Future of Finance: The Transformative Role of Generative AI. *J. Cent. Bank. Law Inst.* **2026**, *5*, 79–124. [\[CrossRef\]](#)
11. Sadek, M.; Kallina, E.; Bohné, T.; Mougenot, C.; Calvo, R.A.; Cave, S. Challenges of responsible AI in practice: Scoping review and recommended actions. *AI Soc.* **2024**, *40*, 199–215. [\[CrossRef\]](#)
12. Aldasoro, I.; Gambacorta, L.; Korinek, A.; Shreeti, V.; Stein, M. Intelligent Financial System: How AI is Transforming Finance. *J. Financ. Stab.* **2025**, *81*, 101472. [\[CrossRef\]](#)
13. Mayring, P. Qualitative Content Analysis. *Forum Qual. Sozialforschung/Forum Qual. Soc. Res.* **2000**, *1*, 20. [\[CrossRef\]](#)
14. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3973–3983. [\[CrossRef\]](#)
16. Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; Lim, W.M. How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* **2021**, *133*, 285–296. [\[CrossRef\]](#)
17. Mariani, M.M.; Machado, I.; Nambisan, S. Types of innovation and artificial intelligence: A systematic quantitative literature review and research agenda. *J. Bus. Res.* **2023**, *155*, 113364. [\[CrossRef\]](#)
18. Morillo, F.; Bordons, M.; Gomez, I. An approach to interdisciplinarity through bibliometric indicators. *Scientometrics* **2001**, *51*, 203–222. [\[CrossRef\]](#)
19. Maslej, N.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Kariuki, N.; Capstick, E.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; et al. Artificial intelligence index report 2025. *arXiv* **2025**, arXiv:2504.07139.
20. Crisanto, J.C.; Leuterio, C.B.; Prenio, J.; Yong, J. Regulating AI in the Financial Sector: Recent Developments and Main Challenges. In *FSI Insights on Policy Implementation 63, Bank for International Settlements*; Financial Stability Institute: Basel, Switzerland, 2024.
21. Mandal, A.; Ramanayake, R.; O'Neill, O.; Flanagan, W.; Kemal, N.; Yekrang, M.; Chatbri, H.; Martin, C. Governing Framework for the Responsible Democratization of Large-Language Models in Banking. *IEEE Internet Comput.* **2025**, *29*, 25–34. [\[CrossRef\]](#)
22. Okpala, I.; Golgoon, A.; Kannan, A.R. Human-Centered Agentic Framework for Machine Learning Modeling in Finance. In Proceedings of the 2025 IEEE International Conference on Service-Oriented System Engineering (SOSE), Tucson, AZ, USA, 21–24 July 2025; pp. 226–233. [\[CrossRef\]](#)
23. Zöller, M.A.; Iurshina, A.; Röder, I. Trustworthy Generative AI for Financial Services. In *Proceedings of the Symposium on Scaling AI Assessments (SAIA 2024)*, Dagstuhl, Germany; Open Access Series in Informatics (OASIs); Dagstuhl Publishing: Wadern, Germany, 2025; Volume 126, pp. 2:1–2:5. [\[CrossRef\]](#)
24. Lee, H.; Seo, J.; Park, S.; Lee, J.; Ahn, W.; Choi, C.; Lopez-Lira, A.; Lee, Y. Your AI, Not Your View: The Bias of LLMs in Investment Analysis. In Proceedings of the ICAIF '25: 6th ACM International Conference on AI in Finance, Singapore, 15–18 November 2025; pp. 150–158. [\[CrossRef\]](#)

25. Biancotti, C.; Camassa, C.; Coletta, A.; Giudice, O.; Glielmo, A. Chat Bankman-Fried: An Exploration of LLM Alignment in Finance. In Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 1–22.
26. Remolina, N. Generative AI in Finance: Risks and Potential Solutions. *Law Ethics Technol.* **2024**, *1*, 0002. [[CrossRef](#)]
27. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30, pp. 4765–4774.
28. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1135–1144. [[CrossRef](#)]
29. Wang, Y.; Ding, Q.; Wang, K.; Liu, Y.; Wu, X.; Wang, J.; Liu, Y.; Miao, C. The skyline of counterfactual explanations for machine learning decision models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2030–2039.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30, pp. 5998–6008.
31. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2022; Volume 35, pp. 24824–24837.
32. Vilone, G.; Sovrano, F.; Lognoul, M. On the Explainability of Financial Robo-Advice Systems. In *Explainable Artificial Intelligence: Second World Conference, xAI 2024, Valletta, Malta, July 17–19, 2024, Proceedings*; Springer: Cham, Switzerland, 2024; Volume 2154, pp. 219–242. [[CrossRef](#)]
33. Turpin, M.; Michael, J.; Perez, E.; Bowman, S. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 74952–74965.
34. European Parliament; Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). In *Official Journal of the European Union*; EU: Brussels, Belgium, 2024.
35. European Parliament; Council of the European Union. Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on Digital Operational Resilience for the Financial Sector (Digital Operational Resilience Act). In *Official Journal of the European Union*; EU: Brussels, Belgium, 2022.
36. European Parliament; Council of the European Union. Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on Markets in Financial Instruments and Amending Directive 2002/92/EC and Directive 2011/61/EU (Recast). In *Official Journal of the European Union*; Markets in Financial Instruments Directive II (MiFID II); EU: Brussels, Belgium, 2014; Volume L173.
37. Board of Governors of the Federal Reserve System; Office of the Comptroller of the Currency. *Supervisory Guidance on Model Risk Management (SR 11-7)*; Technical Report; Federal Reserve System: Washington, DC, USA, 2011.
38. Monetary Authority of Singapore. *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector*; Technical Report; Monetary Authority of Singapore: Singapore, 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.