JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Navigating The Black Box: A Scoping Review of Large Language Models and Explainable Artificial Intelligence for Sovereign Wealth Funds and Public Financial Institutions Under Fiduciary Accountability

[1*]Labaran Usman Yunusa, [2]Rufai Yauri Aliyu, [3]Charles Isah Saidu, [4]Rislana A. Kanya, [5]Amina F. Shehu, [6]Ahmad N. Tambaya

[1,2,3,5]Department of Computer Science, Baze University, Abuja, Nigeria
[4]Cosmopolitan University, Abuja, Nigeria
[6]Cyber Safety Alliance (CSA), Nottingham, United Kingdom
Corresponding Author: usman8194@bazeuniversity.edu.ng

ABSTRACT

Public financial institutions (PFIs) are adopting general-purpose large language models (LLMs) faster than the evidence supports. Sovereign wealth funds (SWFs) operating under fiduciary duties and transparency obligations set out in the Santiago Principles are unable to adapt quickly because generic or unexplained model outputs cannot satisfy these requirements. This review maps the literature on adapting LLMs for institutional finance and the explainability of their outputs with the aim of identifying the evidence gaps critical for SWF deployment. We conducted a corpus-bounded scoping review following PRISMAScR. The corpus comprised a 496-record library assembled during a longitudinal study of an African SWF (2023-2026). After deduplication, 470 unique records were screened against predefined criteria, yielding 230 included studies synthesized across three themes: LLM adaptation, explainable artificial intelligence (xAI), and deployment risk and evaluation. Theme mappings comprised 77 studies on adaptation, 108 on explainability, and 74 on deployment risk and evaluation with some studies spanning multiple themes. Adaptation splits into two opposing approaches as parametric fine-tuning injects institutional knowledge but obscures its provenance while retrieval-augmented generation (RAG) exposes its sources but cannot guarantee their use. xAI techniques developed for classifiers degrade significantly on free-text generative output. No study evaluates an LLM deployed in an SWF within the defined search boundaries. High performance on generic benchmarks does not certify a model for institutional deployment, and African SWFs remain an empirical void. This review establishes the research agenda required to address this gap.

KEYWORDS: Large Language Models; Explainable Artificial Intelligence; Public Financial Institutions; Sovereign Wealth Funds; Retrieval-Augmented Generation; Scoping Review

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

I. INTRODUCTION

Large language models are artificial intelligence systems trained on extensive text corpora to generate fluent human-like language, and they are being deployed across financial services at a rapid pace (Fatemi and Hu, 2023; Fatemi et al., 2024). Since the advent of ChatGPT, generative AI (GenAI) has become a strategic priority across the sector with surveys cataloguing applications ranging from document summarization to investment research (Chui et al., 2023; Li et al., 2023; Nie et al., 2024). However, sovereign wealth funds which are state -owned investment vehicles managing national assets exhibit considerable caution. These institutions operate under strict governance, transparency, and fiduciary requirements codified internationally in the Santiago Principles which mandate accountable and well-documented decision-making (IFSWF, 2025). For such institutions, an unexplainable algorithmic output is inadmissible, and as generic models proliferate across the finance industry, the divergence between model capability and the operational constraints of PFIs continues to widen. The financial ramification of this divergence is significant. By 2024, SWF assets under management (AuM) reached a record US$13.2 trillion (Megginson et al., 2023; SWFI, 2024), a 14% annual increase on top of the two decades of growth at ~11% per year (Aguilar and Johnson, 2024). These funds are highly knowledge-intensive yet rely more on paper-based internal documentation such as investment mandates, memos, risk reports, and audited accounts. Industry analysis suggests that a large fund can realize net annual gains exceeding US$100 million through improved institutional knowledge management practices (BCG et al., 2025), and leading funds are consequently developing internal analysis and summarization AI tools (Burroughes, 2024). Therefore, the primary research question has shifted from whether SWFs will adopt LLMs to how SWFs can adopt LLMs without breaching their governance frameworks.

Two barriers dominate this question. The first is hallucination because models produce confident outputs regardless of whether the underlying facts are true (Shuster et al., 2021). The second but much subtler barrier is parametric override where a model tends to prefer its own memorized weights when they conflict with supplied context (Longpre et al., 2021). Both barriers are manifestations of opacity and have been long identified as the main obstacle to trust in AI systems (Adadi and Berrada, 2018; Eschenbach, 2021). This problem is more intense in generative systems because conventional explainability techniques like LIME and SHAP were built to attribute a single prediction to its inputs and transfer poorly to open-ended text (Schneider, 2024; Zhao et al., 2024). Financial regulators now expect institutions to justify algorithmic decisions, and commercial banks already struggle to meet that expectation (Kuiper et al., 2021; OECD, 2023). Nevertheless, a public fund accountable to a national legislature is faced with a similar demand but with lower risk tolerance. To address these challenges, this review synthesizes the literature around three primary objectives. (i) It examines domain specific adaptation especially how general-purpose LLMs are specialized for finance through fine-tuning (Hu et al., 2021; Liu et al., 2023) or RAG (Lewis et al., 2020). (ii) It evaluates the explainability of LLMs in high-stakes financial contexts by assessing the available explanation techniques, their theoretical validity, and their empirical validation (Weber et al., 2024; Černevičienė and Kabašinskas, 2024). (iii) It analyses deployment risks and evaluation methodologies by focusing on the measurement of hallucination, knowledge conflicts, and security vulnerabilities, as well as the benchmarking of systems prior to institutional reliance (Chen et al., 2024; Laskar et al., 2024).

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

This review is guided by three research questions:

RQ1. How do domain-adaptation methods like fine-tuning, RAG, and hybrid architectures integrate institutional knowledge into models?

RQ2. Can existing xAI techniques turn open-ended text generation into an auditable trail that satisfies fiduciary and regulatory accountability?

RQ3. Do standard evaluation practices measure the attributes that matter to high-accountability organizations?

The rationale for integrating these domains is a notable absence in the evidence base. Empirical literature on LLMs in finance concentrates on commercial banking and capital markets in advanced economies. PFIs are rarely examined with the notable exception of a government finance assistant utilizing Indonesian fiscal data (Febrian and Figueredo, 2024). African institutions are absent from the literature as they remain focused on fintech prospects and regulatory readiness rather than deployed systems (Katterbauer et al., 2024; Makore and Makore, 2024). This omission matters because Africa hosts a growing number of SWFs including the Nigeria Sovereign Investment Authority (NSIA) which is a Santiago Principles signatory managing stabilization, savings, and infrastructure funds for the continent's largest economy (NSIA, 2025; Johnson et al., 2025). These institutions would benefit from document grounded AI but remain underrepresented in the research required for safe adoption. Thus, this review aims to establish the foundational requirements for designing and evaluating such a system for an African SWF. Our review makes three distinct contributions with the first being the integration of three distinct bodies of literature (domain adaptation of LLMs, xAI in finance, and LLM evaluation and risk). These literature bodies are typically reviewed in isolation, but we analyze them through the operational requirements of PFIs rather than commercial entities. Secondly, we identify critical evidence gaps particularly concerning explanation methods validated for generative models, evaluations grounded in institutional documents, and empirical research on emerging-market PFIs. Thirdly, we derive practical implementation considerations for governance-constrained institutions with the aim of bridging the gap between model capabilities and institutional compliance requirements.

II. RELATED WORKS

The reviewed literature relevant to this study can be categorized into three distinct strands and this section outlines the contributions of each strand. It also identifies their limitations in respect to the responsible deployment of LLMs and xAI within PFIs. The first strand covers broad surveys of LLM architectures and capabilities. Zhao et al. (2023) and Minaee et al. (2024) document the training methodologies and the emergent properties of current models. Chen et al. (2024) further extends this analysis to high-stakes societal domains like finance. Looking towards model evaluation, Zhuang et al. (2023) categorize evaluation by core competencies while Laskar et al. (2024) identify the limitations of prevailing benchmark. These surveys comprehensively establish model capabilities and evaluation metrics, but they treat finance as just another application domain, among others. Also, the specific institutional contexts required to adopt, govern, and take accountability for these models remain unaddressed.

A second strand of literature surveys xAI methodologies with numerous foundational reviews by Adadi and Berrada (2018), Guidotti et al. (2018), and Arrieta et al. (2020) establishing standard taxonomies to distinguish between intrinsic and post-hoc explanations. They also clearly identified opacity as the primary barrier to trust (Gilpin et al., 2018). Subsequent research pivoted to evaluation frameworks with Vilone and Longo (2021) and Nauta et al. (2023) systematizing evaluation properties, Kadir et al. (2023) categorizing metrics, and expanding on human-centered studies to expand upon the evaluation logic (Kim et al., 2024;

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

Doshi-Velez & Kim, 2017). These reviews simply prove that explanation quality is inherently multidimensional and notoriously difficult to measure. However, they fail to anchor these properties within specific accountability regimes and the intended user of an explanation in these reviews is usually an abstract stakeholder instead of an auditor, regulator, or fiduciary that is legally bound. Another limitation is that most of the surveyed techniques were built for traditional classifier models, and the recent reviews looking specifically at generative LLMs (Zhao et al., 2024; Schneider, 2024) show that legacy attribution methods transfer poorly to open-ended text generation.

The third strand mostly examines AI applications and governance in the financial sector. Recent adaptation surveys are encouraging as they show rapid domain adoption across forecasting, advisory, and risk management (Li et al., 2023; Nie et al., 2024; Borovkova, 2024). In contrast, the parallel xAI reviews tell a narrower story as explanation techniques are mostly concentrated in credit scoring and fraud detection using mostly LIME and SHAP (Černevičienė and Kabašinskas, 2024; Weber et al., 2024; Chen et al., 2023). On governance, the regulatory analyses agree that supervisors now routinely demand explainability before deployment (Gasser & Almeida, 2017; Kuiper et al., 2021; Crisanto et al., 2024; Leitner et al., 2024). Nevertheless, the clear blind spot is that this review strand is majorly present in commercial banking, insurance, and capital markets while institutions like PFIs which manage public capital under strict fiduciary mandates rather than for profit fall completely outside the scope.

This gap becomes undeniable when focusing on SWFs. This is because even though there is a mature literature on their governance and political economy (Alhashel, 2015; Clark and Monk, 2010; Boubakri et al., 2023), it simply predates modern LLMs and ignores AI entirely. The nearest technical work is thin: a government-finance LLM built for Indonesia (Febrian and Figueredo, 2024), some high-level policy analyses of sovereign LLM strategies (Bondarenko et al., 2025), and a regional assessment of generative AI in African fintech (Katterbauer et al., 2024). Apart from these, no existing review connects explainability techniques to the strict fiduciary accountability these institutions carry, and nobody has engaged with SWFs directly.

We therefore position our review against this literature in three specific ways. Firstly, we use the institution as the primary unit of analysis rather than the technical method. Secondly, we evaluate both LLM and xAI literatures through a governance and accountability framework rather than a purely technical capability lens. Thirdly, we bound our corpus using an active longitudinal study of an African SWF. This ensures the synthesis remains focused on the real-world practical requirements of a deploying institution.

III. METHODOLOGY

This paper follows the PRISMA-ScR structured scoping review approach which is suitable for computing and intelligent systems research. This approach was adopted to make the identification, screening, eligibility assessment, and inclusion of studies processes transparent and reproducible (Tricco et al., 2018). A search and source log, screening spreadsheet, reviewer-agreement sheet, and PRISMA flow diagram were developed to document the process and to support the credibility of the final synthesis. In addition, the search log documents the iterative search path of the underlying longitudinal study rather than a single protocol-driven sweep because the design is corpus-bounded. This is disclosed as a limitation.

A. Review Corpus and Search Strategy

This review employs a corpus-bounded methodology rather than a fresh protocol-driven database search. It utilizes the complete reference library assembled during a longitudinal research project on an African SWF conducted between January 2023 and December 2025,

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

with the library consolidated in May 2026. Due to the corpus-bounded design of this review, the search log records the iterative search trail of the underlying longitudinal study rather than a single protocol-driven sweep. This is disclosed in Section III.G as a limitation.

A.1 Search Terms and Sources

The various research phases ran iterative literature searches through Google Scholar and in other publisher repositories like ACM, IEEE (IEEE Xplore), Elsevier (ScienceDirect), Springer (SpringerLink), the ACL, MDPI, Wiley, SSRN and arXiv. Each phase combined three concept blocks: (i) language models ("large language model*", "LLM", "generative AI", "GenAI", "GPT", "foundation model*"); (ii) institutional finance ("sovereign wealth fund*", "public financial institution*", "development finance institution*", "state-owned investor", financ*); (iii) accountability and risk (explainab*, interpretab*, "XAI", "SHAP", "LIME", counterfactual*, hallucinat*, "knowledge conflict*", "retrieval-augmented generation", "RAG", benchmark*, evaluation, abstention). An illustrative full string is: ("large language model*" OR "LLM" OR "generative AI") AND ("sovereign wealth fund*" OR "public financial institution*" OR financ*) AND (explainab* OR hallucinat* OR "retrieval-augmented generation" OR benchmark*). These searches were supplemented in every phase through snowballing from included studies and citation tracking.

A.2 Record Accumulation

The library developed through every phase as all relevant records in any phase were retained rather than filtered to a single question: adaptation and fine-tuning searches (2023), retrieval and knowledge-conflict searches (2023-2024), explainability and evaluation searches (20242025), and deployment-risk and governance searches (2025). This accumulation produced the 496-record library with no records published after 2025 because it is bounded by the conclusion of the longitudinal study. The final consolidation date of the corpus was May 2026.

A.3 Source Verification and Composition

The publication source of each record was verified from its digital object identifier (DOI) or uniform resource locator (URL) with Table 1 detailing the distribution of the 496 records. Notably, peer-reviewed and preprint sources dominate the portion of the corpus with arXiv as the single largest source, reflecting the rapid publication cycle of this domain. The corpus consists of ~34% grey literature (institutional reports, regulatory publications, and web sources) because a significant portion of authoritative materials on PFIs are not published in traditional academic journals. The supplementary OpenAlex search described in Section III.F is an external validation check and is explicitly not part of this primary corpus or synthesis.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

Table 1: Publication Sources of Records in the Review Corpus

Source	Records	%
arXiv	77	15.5
ScienceDirect (Elsevier)	51	10.3
ACM Digital Library	24	4.8
SpringerLink	23	4.6
IEEE Xplore	15	3.0
MDPI	15	3.0
ACL Anthology	12	2.4
SSRN	11	2.2
Wiley	6	1.2
Other Academic Publishers	94	19.0
Institutional Reports and Web Sources (Grey Literature)	168	33.9
Total	**496**	**100**

B. Inclusion and Exclusion Criteria

We established clear boundaries for this review during the screening process through the specific inclusion and exclusion criteria detailed in Table 2.

Table 2: Inclusion and Exclusion Criteria for Selecting Studies

Decision	Criterion
Include	Addresses at least one review theme (domain adaptation, explainability and trust, or deployment risk and evaluation)
Include	Is a research study or a substantive technical report with reported methods or evidence
Include	Falls within the corpus timeline (up to 2025)
Exclude	Not research: news coverage, vendor marketing, or product pages
Exclude	Off topic: no bearing on language models or their institutional use
Contextual (Cited but not Synthesized)	Background on SWF governance, knowledge management theory, or research methodology

C. Study Selection

The selection process we employed followed the standard phases illustrated in Figure 1: identification, screening, eligibility, and inclusion. 26 duplicates were removed from the initial 496 records to provide 470 unique records for the title and abstract screening phase. In this next phase, we removed 77 more records because 68 were non-research materials (news articles, vendor announcements, and product documentation) while 9 did not align with our review's focus. The remaining 393 records went through full-text assessment against the review themes, and 163 were categorized as purely contextual literature that cover SWF governance, knowledge management theory, and broad methodology. It is worth noting that these 163 records were not synthesized as primary evidence even though they helped in theoretically formulating the review. This final filter left us with 230 studies for final inclusion.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

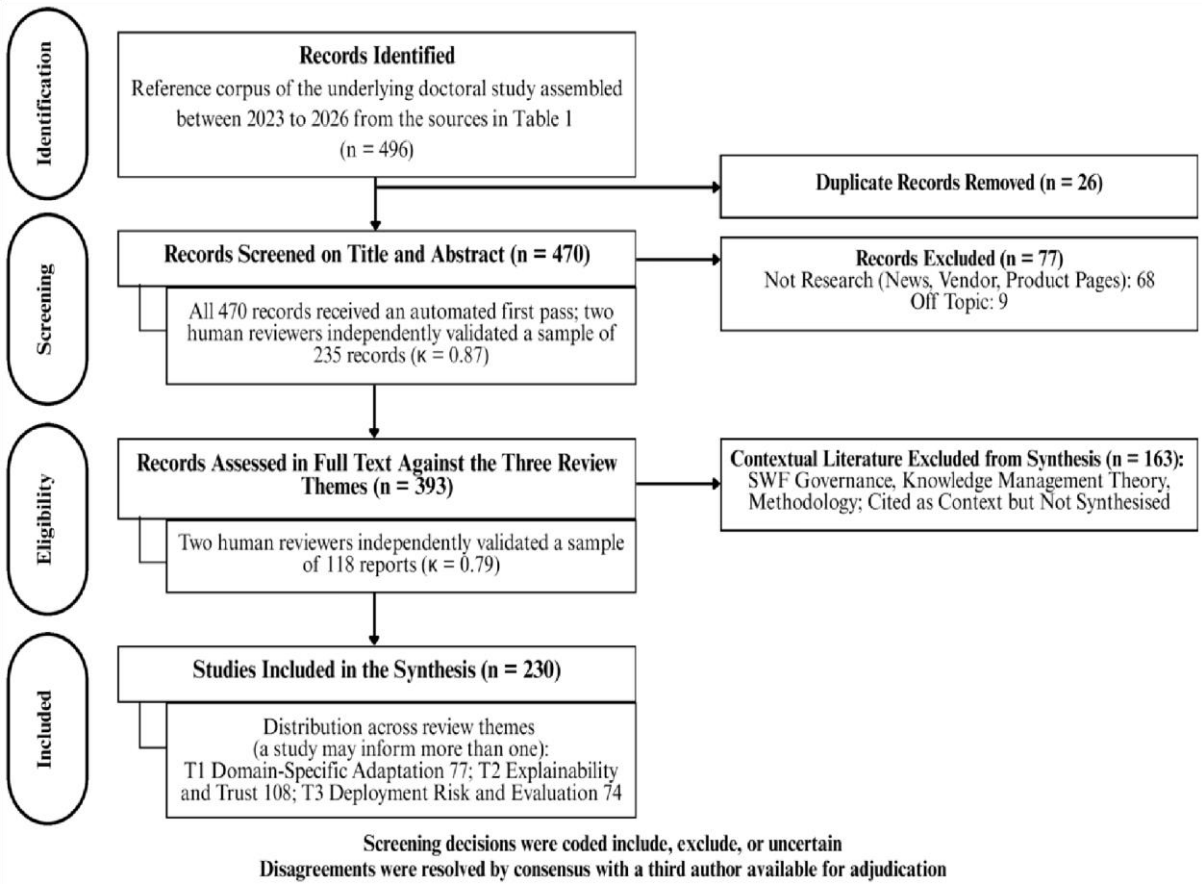

Figure 1: PRISMA-ScR Flow Diagram of the Study Selection Process

D. Thematic Mapping

We mapped the included studies into three predefined themes we derived from our primary research questions.

1. Theme 1 (T1) covers Domain-Specific Adaptation of Language Models (77 studies)
2. Theme 2 (T2) covers Explainability and Trust (108 studies)
3. Theme 3 (T3) covers Deployment Risk and Evaluation (74 studies)

The total count of these studies when classified into themes is 259, and it deliberately exceeds the 230 included studies because individual studies frequently cross thematic boundaries. For example, a paper on RAG speaks to both adaptation and evaluation.

E. Screening Procedure

We conducted the screening in two stages. In stage one, we adopted a machine pass using an automated zero-shot text classifier configured with the Table 2 criteria as class descriptors to run over all 470 deduplicated records and map each one against the three review themes and the exclusion criteria. This stage produced our preliminary inclusion list. In stage two, to verify the reliability of this automated process and ensure accuracy, we drew a validation sample of records for independent human review.

E.1 Validation-Sample Selection

The samples were drawn by stratified random sampling with the strata defined by source type (peer-reviewed, preprint, or grey literature). To ensure the samples reflected the heterogeneity of the corpus, we crossed them with the provisional theme labels. This yielded 235 records

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

(50% of 470) at the title and abstract stage and 118 records (30% of 393) at the full-text eligibility stage. Two co-authors independently screened the samples using the predefined criteria in Table 2 with screening decisions coded as include, exclude, or uncertain. Disagreements between the human reviewers were discussed against the eligibility criteria and resolved by consensus with a third author available for adjudication where consensus could not be reached. Inter-rater agreement between the two human reviewers on the validation samples were assessed using Cohen's kappa (κ) coefficient and the results are summarized in Table 3.

At the title and abstract screening stage, the reviewers assessed a sample of 235 records, agreed on 226 records, and disagreed on 9 records, producing a κ value of 0.87. At the full-text eligibility stage, the reviewers assessed 118 papers, agreed on 106 papers, and disagreed on 12 papers, producing a κ value of 0.79. These values indicate strong agreement at the title and abstract screening stage and substantial agreement at the full-text eligibility assessment. Following independent verification, consensus resulted in three of the 9 disputed title and abstract records being retained for full-text assessment and six being excluded. At the full-text stage, five of the 12 disputed reports were included and seven were excluded. The final consensus decisions from the validation sample confirmed the reliability of the automated tool.

Table 3: Reviewer Agreement During Study Selection (Validation Sample)

Screening Stage	Title and Abstract Screening	Full-Text Eligibility
Records Assessed (Sample)	235 of 470	118 of 393
Agreements	226	106
Disagreements	9	12
Percentage Agreement	96.2%	89.8%
Cohen's Kappa	0.87	0.79
95% Confidence Interval	0.79 to 0.95	0.68 to 0.90
Interpretation	Strong Agreement	Substantial Agreement
Resolution Procedure	Disagreements were resolved through discussion and consensus	Disagreements were resolved through discussion and consensus

E.2 Error Propagation

The human consensus decisions were compared against the automated tool's classifications to estimate category-wise error rates focusing on false inclusions and false exclusions per exclusion criterion and per theme. This error propagation followed three distinct rules:

(a). Categories with an error rate over 5% caused manual re-screening of all remaining records carrying an automated label.

(b). Categories below the 5% threshold had their automated labels retained and theme counts adjusted by the estimated error rate.

(c). All adjustments were recomputed holistically to construct the PRISMA flow diagram in Figure 1 and determine the final 230 studies included in the review.

The validation samples served both to certify the tool and to quantify residual classification error in the unscreened remainder.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

F. Supplementary Validation Search

Two features of the primary design required an external check. The corpus is bounded by the search trail of a single longitudinal study, and the central empirical finding of the review is an absence where no included study evaluates a language model within an SWF. To mitigate the selection bias inherent in a corpus-bounded approach and to test this absence against an open catalogue, a supplementary targeted search was conducted in July 2026 using OpenAlex, an open bibliographic index covering more than 250 million scholarly works across disciplines and publishers. The search intersected three concept blocks, namely PFIs, GenAI technologies, and deployment or evaluation contexts, through the expression ("sovereign wealth fund" OR "public financial institution") AND ("large language model" OR "generative AI") AND ("deployment" OR "evaluation"), translated into OpenAlex query syntax. The three intersections returned 42, 25, and 5 records respectively, yielding 72 records for screening with records appearing under more than one intersection screened once.

Consequently, the two reviewers screened all 72 records at title and abstract level against the criteria in Table 2 with abstracts consulted where a title alone was insufficient. No record documented the deployment or evaluation of a language model with or without an xAI component inside an SWF. The closest matches mentioned SWFs only in passing or appeared as policy and market commentary which is consistent with the neighbouring categories discussed in Section IV.D. The supplementary protocol therefore independently confirms the empirical gap identified within the primary corpus. These 72 records were screened only to test the gap claim against an open catalogue but none of these records were added to the included review corpus and the synthesis of the 230 selected studies comes from the primary corpus only. Thus, the supplementary search is methodologically distinct from and subordinate to the primary review procedure.

G. Limitations of the Design

This adopted methodology presents three inherent limitations with the first highlighting that the review is bounded by a single corpus developed during one longitudinal study. This absence of a formal protocol-driven database search implies that relevant studies outside the initial search parameters may have been omitted. However, the supplementary OpenAlex validation search described in Section III.E mitigates this concern with respect to the central gap claim but studies outside both search trails may still have been omitted. Secondly, while grey literature is included in the corpus, it is analytically separated from peer-reviewed evidence and any claim relying exclusively on grey sources is explicitly identified. Thirdly, works published in other languages are underrepresented as the corpus is restricted to English language publications only.

IV. RESULTS AND THEMATIC SYNTHESIS

The synthesis encompasses 230 included studies that map against the three review themes to give 77 studies contributing to domain-specific adaptation (T1), 108 to explainability and trust (T2), and 74 to deployment risk and evaluation (T3). However, because individual studies may address multiple themes, the combined total exceeds the number of included studies. The evidence base is strongly concentrated in recent years as more than half of all included studies were published between 2023 and 2024. This indicates rapid growth and increasing scholarly attention in this area.

Consequently, one structural finding frames the discussion that follows and although explainability and trust (T2) is the largest theme by volume, the literature provides surprisingly limited insight into the systems that financial institutions are now deploying in practice. This is

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

evident as most explainability research was developed in the context of traditional classifiers and predictive models while the technologies increasingly entering financial institutions today are generative language models whose outputs are dynamic, context-dependent, and often difficult to interpret using established explainability methods.

Across the 108 studies mapped to T2, only a small minority address explainability for generative models. Notably, not a single study demonstrates a production-ready explainability framework for LLMs operating within a financial institution. This highlights a fundamental mismatch between the focus of present explainability literature and the realities of real-world deployment. The following sections examine how this gap emerged by tracing how domain specific adaptation methods evolved in Section IV.A. The next section shows how explainability became a regulatory mandate even though existing explainability toolkits are limited in Section IV.B. Furthermore, we identify the unresolved deployment risks and evaluation challenges for financial GenAI in Section IV.C. Finally, Section IV.D synthesizes these findings to articulate the resulting research gap.

A. The Trajectory of Domain-Specific Adaptation

The literature on adaptation documents a clear story of the steady reduction in generic model specialization cost along with a fundamental shift in the place models store domain knowledge. Early transformer architectures absorbed factual knowledge directly into their internal parameters during massive pre-training runs (Vaswani et al., 2017; Brown et al., 2020; Kaplan et al., 2020). Subsequent advances in instruction tuning and reinforcement learning from human feedback (RLHF) aligned model behaviour with user intent without necessarily adding new facts (Ziegler et al., 2019; Ouyang et al., 2022; Wu et al., 2023). More recently, parameter efficient fine-tuning (PEFT) methods like LoRA and quantization techniques such as QLoRA have drastically lowered the computational barrier to allow institutions to adapt multi-billion parameter models on standard hardware (Hu et al., 2021; Dettmers et al., 2022; Dettmers et al., 2023). Collectively with the release of capable open-weight models (Touvron et al., 2023; Bai et al., 2023; Grattafiori et al., 2024), these developments enable organizations to adapt models entirely within their internal infrastructure. This is a capability increasingly discussed in the context of sovereign AI strategies (Bondarenko et al., 2025).

RAG took an alternative paradigm from fine-tuning by giving systems the ability to retrieve relevant text from an external document store at inference time and supplying them in the prompt (Lewis et al., 2020; Karpukhin et al., 2020). This approach keeps knowledge in versioned and citable documents while measurably reducing fabricated content (Shuster et al., 2021; Es et al., 2023). The finance industry followed this general trajectory with early domain adaptation relying on continued pre-training of encoder models for tasks like sentiment analysis (Araci, 2019; Shah et al., 2022). More recent efforts have produced large-scale financial models ranging from proprietary systems trained from scratch to open alternatives fine-tuned on financial instruction data (Liu et al., 2023; Xie et al., 2023; Fatemi and Hu, 2023; Fatemi et al., 2024). Similarly, parallel developments in other specialized domains confirm the viability of this approach (Singhal et al., 2022; Luo et al., 2022; Taylor et al., 2022).

For PFIs, the one distinction that matters above all the others is where does the knowledge end up and can it be traceable. Fine-tuning embeds institutional facts into model weights, and once this is done, mechanistic studies show that these facts cannot subsequently be isolated or cited (Jain et al., 2023; Wang et al., 2024). Retrieval does the opposite by keeping knowledge outside the model. In principle, this means that every answer can carry its own citation but in practice this depends on how the model defers to the supplied context, and this deference is not guaranteed as shown in Section IV.C. This is the main reason hybrid designs that combine

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

domain fluency in model weights and retrieval at inference are emerging as a practical compromise (Zhang et al., 2023; Efeoglu and Paschke, 2024; Febrian and Figueredo, 2024; O'Leary, 2024; Lee et al., 2024). Table 4 consolidates the discussed trade-offs with an inverse relationship between parametric depth and auditability. The approaches that inject the most institutional knowledge leave the least evidence while RAG which injects nothing has the strongest per-answer audit trail. This inverse relationship rather than cost of deployment is the decisive consideration for PFIs. It also explains the corpus's conjunction toward hybrid designs that pair domain fluency with retrievable sources.

Table 4: Principal Adaptation Approaches and Their Traceability Consequences

Approach	What It Installs	Source Traceability	Cost Profile	Representative Works
Domain pretraining from scratch	Broad domain knowledge in parameters	None: facts are diffused across weights	Extreme; feasible for few institutions	Liu et al. (2023); Li et al. (2023)
Continued domain pretraining of encoders	Domain vocabulary and style	None	Moderate	Araci (2019); Shah et al. (2022)
Instruction tuning/RLHF	Behaviour and preference alignment instead of facts	Not applicable: shapes conduct, not content	Moderate; needs curated instruction data	Ziegler et al. (2019); Ouyang et al. (2022); Taori et al. (2023)
PEFT (LoRA and QLoRA)	Domain knowledge and style through small adapters	None: adapted facts still live in weights	Low; single-workstation easible	Hu et al. (2021); Dettmers et al. (2023)
Quantisation	Nothing new; compresses an existing model	Neutral; preserves whatever the model had	Extremely low; enables on-premises deployment	Dettmers et al. (2022)
RAG	Nothing; knowledge supplied per query from documents	High by design: every answer can cite retrieved passages	Low training cost; ongoing index upkeep	Lewis et al. (2020); Karpukhin et al. (2020); Shuster et al. (2021)
Hybrid finetuning plus retrieval	Domain fluency in weights; facts from documents	Partial: depends on model deference to retrieved sources	Low to moderate	Zhang et al. (2023); Efeoglu and Paschke (2024); Febrian and Figueredo (2024)

B. The Imperative for Explainability

Explainability dominates the corpus because it is not optional in finance. Data protection law constrains automated decision-making and grounds a right to meaningful explanation (Wachter et al., 2017; EDPS, 2023). Regulators now expect an audit trail behind any model-driven decision (Crisanto et al., 2024; Global and Scholz, 2025), and the literature evidence shows that banks are already treating explainability as a condition of deployment (Kuiper et al., 2021). Underneath all the rules sits a simple fiduciary fact that any institution managing public funds must be able to justify a system's decision. However, an opaque system breaks the chain of justification required to trust such decisions (Eschenbach, 2021; Langer et al., 2021). The field that answers this demand is mature but for the wrong generation of models. The foundational surveys establish the standard taxonomies (Adadi and Berrada, 2018; Guidotti et al., 2018; Arrieta et al., 2020; Gilpin et al., 2018), and the finance-specific reviews show the techniques are majorly concentrated within credit scoring, risk management, and forecasting (Yeo et al., 2023; Černevičienė and Kabašinskas, 2024; Weber et al., 2024; Chen et al., 2023).

The most adopted technique families are feature attribution with SHAP and LIME, saliency maps, counterfactual explanations, and rule-based surrogates. Further applied studies

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

demonstrate their effectiveness in classifier settings (Demajo et al., 2020; Bussmann et al., 2021; Misheva et al., 2021; Bianchi and Brì, 2021; Saves et al., 2025). Table 5 details why each evaluation family's fiduciary suitability diverges from its classifier-era reputation once outputs become open-ended text. Provenance and source attribution are the only ones that retain strong suitability for generative systems while all post-hoc families either scales poorly (feature attribution), addresses the wrong audience (saliency), lacks theoretical reliability (attention), is inaccurate for open text (counterfactuals and surrogates), or cannot be trusted as sole evidence (self-explanation). This asymmetry is the empirical basis for the provenance-first recommendation.

Table 5: Explainability Techniques and Their Fit to Generative Models in Fiduciary Settings

Technique Family	What it Explains	Behaviour on Long Generative Output	Suitability for Fiduciary Settings
Feature Attribution (SHAP/LIME)	Each input feature's share in one prediction	Cost grows with prompt and output length; Unstable per query	Strong for classifiers; Impractical per-answer for LLM output
Saliency/ Gradient Methods	Input regions the model was sensitive towards	Requires model internals; Token-level maps hard for non-experts to act on	Limited: informative to developers not to auditors
Attention Inspection	Where the model "looked"	Cheap but unreliable as explanation	Weak: risks false assurance
Counterfactuals	Smallest input change that alters the output	Poor for open-ended text; No single decision boundary	Strong for lending decisions; Poor for generated text
Model Self-Explanation	The model's own stated rationale	Fluent but not dependably faithful	Unsafe as sole evidence; Useful only alongside checks
Provenance/ Source Attribution	Which documents an answer drew upon	Native to retrieval systems; Low cost per-answer citations	Strong: aligns with audit and record-keeping
Rule-Based Surrogates	Model logic's global approximation	Cannot faithfully summarize an open-ended generator	Useful for scoped sub decisions not free text

The central problem documented in the corpus is that this explainability toolkit was designed for models with fixed input features and single outputs and generative language models possess neither as their inputs are lengthy prompts and their outputs are sequential token choices. Thus, attribution methods scale poorly under these conditions while perturbation-based methods require computationally prohibitive numbers of model evaluations per query (Zhao et al., 2024; Schneider, 2024). Additionally, inspecting attention weights is computationally cheap but theoretically unreliable as an explanation (Jain and Wallace, 2019) while asking models to generate self-explanations yields fluent text that does not dependably reflect the underlying computation (Huang et al., 2023; Luo et al., 2021). The one approach that survives the transition to generative models intact is provenance tracing. In retrieval-based systems, each answer can be explicitly linked to the source text it utilised and while this does not explain the internal computation of the model, it provides auditors with a verifiable path from the generated answer back to the source document. In a fiduciary context, this system-level property is often more valuable than post-hoc feature attribution (Lewis et al., 2020; Es et al., 2023).

The evaluation of explanation quality presents further challenges as the literature organises evaluation into functionality-grounded, human-grounded, and application-grounded levels (Doshi-Velez and Kim, 2017; Mohseni et al., 2021). Further reviews note a heavy reliance on automated proxy measures while expert evaluations remain rare (Nauta et al., 2023; Kim et al.,

https://doi.org/10.5281/zenodo.22952993

2024; Vilone and Longo, 2021; Kadir et al., 2023). Crucially, these evaluation levels frequently disagree as the explanations that users find satisfying are not necessarily faithful to the model's actual reasoning. This implies that evaluations based on user satisfaction can inadvertently reward persuasiveness over accuracy (Sokol and Vogt, 2024; Amarasinghe et al., 2024; MiróNicolau et al., 2024). For a PFI, the implication is that an explanation highly rated by users is not evidence that the system is trustworthy.

C. Deployment Risks: Hallucination, Parametric Override, and The Evaluation Gap

The risk literature keeps returning to three major problems. The first is hallucination. Generative models produce fluent statements that are still wrong, and in finance the costliest and hardest to detect involve numerical figures, dates, and rates (Cao et al., 2024; Borovkova, 2024; Khan and Umer, 2024). Financial question answering makes it worse because the answer depends on chained numerical reasoning, and that chain is a known weakness even in the strongest models (Chen et al., 2022; Xie et al., 2023; Xie et al., 2024). The second problem runs deeper because when what the model memorized conflicts with the provided context, it will often just ignore the context. Further research on knowledge conflicts by Longpre et al. (2021) shows that models prefer their memorized weights over retrieved context. This limitation is also why retrieval only reduces fabrication without being able to eliminate it completely (Shuster et al., 2021). Additionally, the frequency at which a model defers to supplied context depends on instruction alignment (Ouyang et al., 2022; Wu et al., 2023), and the mechanistic work shows fine-tuning reshapes internal knowledge in unpredictable ways (Jain et al., 2023; Wang et al., 2024). This simply implies that combining fine-tuning and retrieval gives you two independent safeguards. This is a trade-off rather than a guarantee because the two can interact and fail simultaneously while producing confident but stale answers. Therefore, while fine-tuning can reduce hallucination on domain-specific queries (Xu et al., 2023), it can also embed the wrong parametric belief that is difficult to change during inference. This tension remains unresolved in current financial deployments.

In response to these risks, general AI ethics and governance frameworks have evolved across multiple layers (Gasser and Almeida, 2017; Floridi et al., 2018; UNESCO, 2022). Additionally, financial sector analyses from regulatory bodies (Boukherouaa et al., 2021; Shabsigh, 2023; OECD, 2023; Crisanto et al., 2024; Leitner et al., 2024; Global and Scholz, 2025) and security assessments for sensitive data (Yao et al., 2024) have also evolved. Further empirical studies of deployed behaviour which are scarce indicate significant risks including simulated trading agents acting unlawfully (Biancotti et al., 2024) and measurable bias in chatbot financial advice (Lakkaraju et al., 2023). The third and final problem is that current evaluation instruments do not measure institutional requirements. General-purpose benchmarks (Wang et al., 2018; Wang et al., 2019; Hendrycks et al., 2020; Srivastava et al., 2022; Liang et al., 2022) and finance specific benchmarks (Chen et al., 2022; Shah et al., 2022; Xie et al., 2023) both evaluate models on public open-domain information while no public benchmark tests performance against an institution-specific document corpus comprising internal policies, mandates, and confidential reports. Similarly, evaluation metrics are misaligned as reference-overlap metrics like BLEU, ROUGE, and METEOR reward surface-level similarity to a reference text (Papineni et al., 2001; Lin, 2004; Banerjee and Lavie, 2005; Zhang et al., 2019) as numerous methodological reviews highlight how poorly these track generation quality (Gehrmann et al., 2022; Goyal et al., 2022; Laskar et al., 2024). Crucially, overlap metrics penalize abstention and a system that correctly declines to answer when sources are insufficient receives the same penalty as a system that fabricates an answer. This dynamic directly contradicts the operational requirements of any

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

compliance-bound institution and it is also driving research interest in selective prediction methods that evaluate accuracy alongside coverage (Kamath et al., 2020). Attempting to solve this issue by adopting another LLM as an evaluator only produces more reliability concerns (Yoo, 2024; Zhao et al., 2025).

Table 6: Deployment Risks: Mechanisms, Mitigations, and Unresolved Matters

Risk	Documented Mechanism	Corpus Mitigations	Unresolved Matters
Hallucination	Fluent generation ungrounded in evidence	Retrieval grounding (Shuster et al., 2021); Domain fine-tuning (Xu et al., 2023)	Residual fabrication persists; No deployed setting error rates for institutional corpora
Numeric Infidelity	Weak chained numerical reasoning conceals errors in fluent text	Finance benchmarks expose the weakness (Chen et al., 2022; Xie et al., 2023)	Overlap metrics barely detect wrong numbers in fluent text
Parametric Override	Model prefers memorized knowledge over supplied context (Longpre et al., 2021)	Instruction alignment increases deference (Ouyang et al., 2022)	Interaction of finetuning and retrieval untested in production finance
Opacity in Audit	Post-hoc techniques degrade on generative output (Zhao et al., 2024)	Provenance-first system design (Lewis et al., 2020; Es et al., 2023)	Faithfulness of cited sources to generated claims rarely verified
Evaluation Misfit	Benchmarks and metrics test public and open domain competence	Selective prediction rewards honest abstention (Kamath et al., 2020)	No benchmark reflects an institution-specific document estate
Data Sovereignty and Privacy	Sensitive data exposed through external model interfaces	Local open-weight deployment (Touvron et al., 2023; Bondarenko et al., 2025)	Governance templates for deployment are prescriptive and not evidenced
Misaligned or Biased Behaviour	Unlawful agent actions and uneven advice quality documented in simulation (Biancotti et al., 2024; Lakkaraju et al., 2023)	Regulatory frameworks and oversight layers (Crisanto et al., 2024)	No study audits these behaviours in a live institutional deployment

Table 6 maps each deployment risks to its documented mechanism, the mitigations available in the corpus, and the matters that remain unresolved. The four-pillar certification framework proposed below responds directly to these unresolved items. The risks mapped in Table 6 force a fundamental shift in how PFIs validate generative systems because current overlap metrics penalise a model for saying it does not know and this directly conflicts with fiduciary accountability. Therefore, certifying an institutional LLM needs a different type of evaluation framework, and we argue that it should stand on four distinct pillars.

(a). Technical Performance (Selective Prediction and Numerical Fidelity)
(b). Explainability Grounded in Verifiable Provenance
(c). Operational Efficiency
(d). Institutional Trustworthiness

D. The Review Gap

Our review findings combine into one unavoidable conclusion: a good benchmark score does not certify a model for institutional use. This is because existing benchmarks evaluate public knowledge with metrics that reward fluency, tolerate numerical errors, and penalize

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

abstentions. This leads to an institution learning almost nothing if it tries to validate an assistant against its own internal mandates. Therefore, public benchmarks provide a weak proxy for institutional readiness because they measure the wrong capabilities. This misalignment in evaluation reflects a deeper empirical gap where no study out of the 230 included studies evaluates an LLM deployed within an SWF. The closest approximations are a prototype for a finance ministry (Febrian and Figueredo, 2024), a private sector knowledge management pilot (Lee et al., 2024), and a strategic analysis of sovereign models (Bondarenko et al., 2025). Notably, African PFIs are entirely absent from the empirical literature as they appear only in regulatory and policy commentary (Katata, 2021; Makore and Makore, 2024). Relevant literature had to be compiled across adjacent fields to assemble the review corpus used for this review due to the absence of a dedicated research stream.

Existing literature reviews support the dimensions of this gap from their respective domains. Financial xAI surveys admit that the field remains concentrated on classifiers and predictive models (Yeo et al., 2023; Černevičienė and Kabašinskas, 2024; Weber et al., 2024). Further financial LLM reviews catalogue adaptation methods without addressing institutional deployment evidence (Nie et al., 2024; Chen et al., 2024). On evaluation, Nauta et al. (2023) and Kim et al. (2024) both note the scarcity of expert task-grounded studies. These reviews identify a pattern where each captures a distinct part of the problem, but none tackles the intersection where generative capability meets institutional accountability. Addressing this gap will require institution-specific and governance-aware evaluation frameworks that employ benchmarks built from internal documents, metrics that reward grounded answers and smart refusal, and systems that treat provenance as a primary output. The future work associated with the underlying longitudinal study is designed around these requirements. We acknowledge several limitations of the evidence base but none invalidates the main conclusions. A significant proportion of the included studies are preprints, which reflects the rapid pace of research in this domain relative to traditional peer-review cycles. Furthermore, the heavy concentration of publications between 2023 and 2024 suggests that specific technical findings may become outdated quickly. Additionally, the corpus-bounded methodology of this review means that we might have omitted outlier studies. None of these limitations would completely obscure a published and evaluated LLM deployment within a PFI if one existed. The empirical gap remains substantial and establishes the necessary direction for further research.

E. Implications and Research Agenda

Based on the synthesized evidence, institutional systems should be grounded in retrieval architectures rather than relying primarily on memorized parametric knowledge. These systems must be designed around provenance to ensure generated outputs maintain verifiable links to source documents while they also support local or institution-controlled deployment to protect sensitive information. Furthermore, abstention must be prioritized as a system capability because a model that explicitly acknowledges uncertainty or declines to answer in the absence of supporting evidence should be regarded as exhibiting a critical governance feature rather than deficient performance. Several research priorities emerge from this analysis with the first being the critical need for the development of institution-specific benchmarks derived from the documentary environments of PFIs because generic financial datasets only provide limited evidence of performance on mandate-bound corpora. Secondly, empirical research is required to evaluate the faithfulness of provenance-based explanations especially to determine whether cited sources genuinely support the content of generated responses. Additionally, future research must formalize a multidimensional evaluation methodology validated against the internal documentary estates of emerging-market SWFs to assess technical performance,

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.22952993

explainability, operational efficiency, and trustworthiness. Thirdly, research must be expanded to include African SWFs and comparable PFIs whose governance contexts, documentary practices, and operational constraints remain absent from the current literature. The development of this framework represents the immediate next step in addressing these gaps.

V. CONCLUSION

This review investigated the current state of knowledge on the deployment of LLMs within PFIs, assessing whether existing explainability techniques can guarantee accountability. The synthesis points to three principal findings. Firstly, adaptation methods handle knowledge in distinct ways that heavily impact traceability. Fine-tuning injects information directly into model weights and makes it difficult to isolate, verify, or cite during deployment. In contrast, retrieval architectures preserve knowledge in external documents that allow generated outputs to be explicitly linked to specific sources. Secondly, older generation explainability techniques fail on generative text. Rigorous provenance which shows the documentary basis of an answer currently represents the most realistic way to enforce accountability in institutional deployments. Thirdly, evaluation practices are not in line with PFI governance requirements because standard benchmarks typically reward fluency and general task performance. They also provide limited insight into source grounding, numerical fidelity, smart refusal, or real institutional reliability. This review makes two primary contributions. Firstly, it represents the first scoping review to examine LLM deployments from the perspective of PFIs, SWFs, and other development finance institutions. Secondly, the review adopts a corpus-bounded and reproducible methodology where all claims are derived from a clearly defined body of included studies. Finally, this review reframes the central challenge of LLM deployments in public finance. The main objective shifts from the pursuit of more capable models to the design of systems with outputs that can be traceable, auditable, and governable in fiduciary settings.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aguilar, J. C., & Johnson, D. (2024). Sovereign Wealth Funds 2024 Resilience and Growth in a New Global Landscape [Report]. Sovereign Wealth Fund Institute.

Alhashel, B. (2015). Sovereign wealth funds: A literature review. Journal of Economics and Business, 78, 1-13. https://doi.org/10.1016/j.jeconbus.2014.10.001

Amarasinghe, K., Rodolfa, K. T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., ... Ghani, R. (2024). On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods. Proceedings of the AAAI Conference on Artificial Intelligence, 38(19), 20921-20929. https://doi.org/10.1609/aaai.v38i19.30082

Araci, D. T. (2019). FinBERT: Financial sentiment analysis with pre-trained language models (arXiv:1908.10063) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1908.10063

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... Zhu, T. (2023). Qwen Technical Report (arXiv:2309.16609) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.16609

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on

Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarisation (pp. 65-72).

BCG, Massi, M., & Shah, P. (2025). Knowledge Management in the Private Equity Industry | BCG [Web publication]. Boston Consulting Group. https://www.bcg.com/industries/principal-investors-private-equity/knowledgemanagement-is-power

Bianchi, M., & Brì, M. (2021). Robo-Advising: Less AI and More XAI? Augmenting algorithms with humans-in-the-loop (SSRN Working Paper No. 3825110) [Preprint]. SSRN. https://doi.org/10.2139/ssrn.3825110

Biancotti, C., Camassa, C., Coletta, A., Giudice, O., & Glielmo, A. (2024). Chat BankmanFried: An Exploration of LLM Alignment in Finance (arXiv:2411.11853) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.11853

Bondarenko, M., Lushnei, S., Paniv, Y., Molchanovsky, O., Romanyshyn, M., & Kiulian, A. (2025). Sovereign Large Language Models: Advantages, Strategy and Regulations (arXiv:2503.04745) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.04745

Borovkova, S. (2024). Large Language Models in Finance [Report]. Probability & Partners.

Boubakri, N., Fotak, V., Guedhami, O., & Yasuda, Y. (2023). The heterogeneous and evolving roles of sovereign wealth funds: Issues, challenges, and research agenda. Journal of International Business Policy, 6(3), 241-252. https://doi.org/10.1057/s42214-023-001632

Boukherouaa, E. B., Shabsigh, G., AlAjmi, K., Deodoro, J., Farias, A., Iskender, E. S., ... Ravikumar, R. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance [Report]. International Monetary Fund. https://policycommons.net/artifacts/1860061/powering-the-digital-economy/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners (arXiv:2005.14165) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Burroughes, T. (2024). What Sovereign Wealth Funds Can Teach The Wider Investment World [Web publication]. WealthBriefingAsia. https://www.wealthbriefingasia.com/article.php/What-Sovereign-Wealth-Funds-CanTeach-The-Wider-Investment-World

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. Computational Economics, 57(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0

Cao, X., Li, S., Katsikis, V., Khan, A. T., He, H., Liu, Z., ... Peng, C. (2024). Empowering financial futures: Large language models in the modern financial landscape. EAI Endorsed Transactions on AI and Robotics, 3. https://doi.org/10.4108/airo.6117

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8). https://doi.org/10.1007/s10462-024-10854-8

Chen, X. Q., Ma, C. Q., Ren, Y. S., Lei, Y. T., Huynh, N. Q. A., & Narayan, S. (2023). Explainable Artificial Intelligence in Finance: A Bibliometric Review. Finance Research Letters, 56. https://doi.org/10.1016/j.frl.2023.104145

Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., ... Wang, W. Y. (2024). A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law (arXiv:2405.01769) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.01769

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., & Wang, W. Y. (2022). ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 6279-6292. https://doi.org/10.18653/v1/2022.emnlpmain.421

Chui, M., Yee, L., Hall, B., Singla, A., & Sukharevsky, A. (2023). The State of AI in 2023: Generative AI's Breakout Year [Report]. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in2023-generative-ais-breakout-year

Clark, G. L., & Monk, A. H. B. (2010). Sovereign Wealth Funds: Form and Function in the 21st Century (SSRN Working Paper No. 1675091) [Preprint]. SSRN. https://doi.org/10.2139/ssrn.1675091

Crisanto, J. C., Leuterio, C. B., Prenio, J., & Yong, J. (2024). FSI Insights on policy implementation No 63 Regulating AI in the financial sector: recent developments and main challenges [Report]. Bank for International Settlements. https://www.bis.org/fsi/publ/insights63.htm

Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable AI for Interpretable Credit Scoring. https://doi.org/10.5121/csit.2020.101516

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. Advances in Neural Information Processing Systems, 35. https://doi.org/10.52202/068431-2198

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems, 36.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning (arXiv:1702.08608) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1702.08608

EDPS (2023). European Data Protection Supervisor (EDPS) TechDispatch on Explainable Artificial Intelligence [Report]. European Data Protection Supervisor. https://doi.org/10.2804/132319

Efeoglu, S., & Paschke, A. (2024). Relation Extraction with Fine-Tuned Large Language Models in Retrieval Augmented Generation Frameworks [Preprint]. ResearchGate. https://www.researchgate.net/publication/381038215

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations, 150-158. https://doi.org/10.48550/arXiv.2309.15217

Eschenbach, W. J. V. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. Philosophy and Technology, 34(4), 1607-1622. https://doi.org/10.1007/s13347-021-00477-0

Fatemi, S., & Hu, Y. (2023). A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis (arXiv:2312.08725) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.08725

Fatemi, S., Hu, Y., & Mousavi, M. (2024). A Comparative Analysis of Instruction Fine-Tuning LLMs for Financial Text Classification (arXiv:2411.02476) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.02476

Febrian, G. F., & Figueredo, G. (2024). KemenkeuGPT: Leveraging a Large Language Model on Indonesia's Government Financial Data and Regulations to Enhance Decision Making (arXiv:2407.21459) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.21459

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks,

Principles, and Recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gasser, U., & Almeida, V. A. (2017). A Layered Model for AI Governance. IEEE Internet Computing, 21(6), 58-62. https://doi.org/10.1109/MIC.2017.4180835

Gehrmann, S., Clark, E., & Sellam, T. (2022). Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. Journal of Artificial Intelligence Research, 77, 103-166. https://doi.org/10.1613/jair.1.13715

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018, 80-89. https://doi.org/10.1109/DSAA.2018.00018

Global, O. L., & Scholz, K. (2025). Artificial Intelligence in the Financial Sector - Regulatory Challenges Between DORA, MiCA, the AI Act, and Banking Supervision [Web publication]. Oracle Law Global. https://oraclelawglobal.com/news/artificialintelligence-in-the-financial-sector/

Goyal, T., Li, J. J., & Durrett, G. (2022). News Summarization and Evaluation in the Era of GPT-3 (arXiv:2209.12356) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2209.12356

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024). The Llama 3 Herd of Models (arXiv:2407.21783) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5). https://doi.org/10.1145/3236009

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. ICLR 2021 - 9th International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2009.03300

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). LoRA: LowRank Adaptation of Large Language Models. ICLR 2022 - 10th International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2106.09685

Huang, S., Mamidanna, S., Jangam, S., Zhou, Y., Gilpin, L. H., Santa, U. C., ... Uc, C. (2023). Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations (arXiv:2310.11207) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.11207

IFSWF (2025). Santiago Principles | IFSWF [Web publication]. International Forum of Sovereign Wealth Funds. https://www.ifswf.org/santiago-principles-landing/santiagoprinciples

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation (arXiv:1902.10186) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1902.10186

Jain, S., Kirk, R., Lubana, E. S., Dick, R. P., Tanaka, H., Grefenstette, E., ... Krueger, D. (2023). Mechanistically analysing the effects of fine-tuning on procedurally defined tasks. 12th International Conference on Learning Representations, ICLR 2024. https://doi.org/10.48550/arXiv.2311.12786

Johnson, D., Aguilar, J. C., Barbary, V., & Dixon, A. (2025). 2025 SOVEREIGN IMPACT REPORT [Report]. Sovereign Wealth Research.

Kadir, A. M., Mosavi, A., & Sonntag, D. (2023). Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. https://doi.org/10.1109/INES59282.2023.10297629

Kamath, A., Jia, R., & Liang, P. (2020). Selective Question Answering under Domain Shift. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 5684-5696. https://doi.org/10.18653/V1/2020.ACL-MAIN.503

Kaplan, J., McCandlish, S., OpenAI, T. H., OpenAI, T. B. B., OpenAI, B. C., OpenAI, R. C., ... OpenAI, D. A. (2020). Scaling Laws for Neural Language Models (arXiv:2001.08361) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Katata, K. S. (2021). Artificial Intelligence in Risk Management and Financial Stability: Overview and Lessons for West African Bank Supervisors (WABS). NDIC Quarterly, 36.

Katterbauer, K., Syed, H., Cleenewerck, L., Özbay, R. D., & Yılmaz, S. (2024). Impact of Generative AI on FINTECH in Africa. Yildiz Social Science Review, 10(1), 43-53. https://doi.org/10.51803/yssr.1440501

Khan, M. S., & Umer, H. (2024). ChatGPT in finance: Applications, challenges, and solutions. Heliyon, 10(2). https://doi.org/10.1016/j.heliyon.2024.e24890

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centred evaluation of explainable AI applications: a systematic review. Frontiers in Artificial Intelligence, 7, 1456486. https://doi.org/10.3389/frai.2024.1456486

Kuiper, O., Berg, M. V. D., Burgt, J. V. D., & Leijnen, S. (2021). Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities (arXiv:2111.02244) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2111.02244

Lakkaraju, K., Jones, S. E., Vuruma, S. K. R., Pallagani, V., Muppasani, B. C., & Srivastava, B. (2023). LLMs for Financial Advisement: A Fairness and Efficacy Study in Personal Decision Making. Association for Computing Machinery, Inc. https://doi.org/10.1145/3604237.3626867

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296. https://doi.org/10.1016/j.artint.2021.103473

Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., ... & Huang, J. X. (2024). A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 13785-13816). Association for Computational Linguistics.

Lee, J., Jung, W., & Baek, S. (2024). In-House Knowledge Management Using a Large Language Model: Focusing on Technical Specification Documents Review. Applied Sciences 2024, Vol. 14, Page 2096, 14(5), 2096. https://doi.org/10.3390/APP14052096

Leitner, G., Singh, J., Kraaij, A. V. D., & Zsámboki, B. (2024). The rise of artificial intelligence: benefits and risks for financial stability. Financial Stability Review [Report]. European Central Bank.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arXiv:2005.11401) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2005.11401 Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large Language Models in Finance: A Survey. 4th ACM International Conference on AI in Finance (ICAIF-23).

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... Koreeda, Y. (2022). Holistic Evaluation of Language Models. Annals of the New York Academy of Sciences, 1525(1), 140-146. https://doi.org/10.1111/nyas.15007

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarisation Branches Out (pp. 74-81).

Liu, X., Wang, G., Yang, H., & Zha, D. (2023). FinGPT: Democratizing Internet-scale Data for Financial Large Language Models (arXiv:2307.10485) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.10485

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-Based Knowledge Conflicts in Question Answering. EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings, 7052-7063. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.565

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6). https://doi.org/10.1093/BIB/BBAC409

Luo, S., Ivison, H., Han, C., & Poon, J. (2021). Local Interpretations for Explainable Natural Language Processing: A Survey. ACM Computing Surveys, 56(9). https://doi.org/10.1145/3649450

Makore, S. M., & Makore, S. T. M. (2024). Regulating Artificial Intelligence to Advance Financial Inclusion in South Africa. PER. https://doi.org/10.17159/1727

Megginson, W. L., Malik, A. I., & Zhou, X. Y. (2023). Sovereign wealth funds in the postpandemic era. Journal of International Business Policy, 6(3), 253-275. https://doi.org/10.1057/s42214-023-00155-2

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey (arXiv:2402.06196) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.06196

Miró-Nicolau, M., Jaume-i-Capó, A., & Moyà-Alcover, G. (2024). A Comprehensive Study on Fidelity Metrics for XAI (arXiv:2401.10640) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.10640

Misheva, B. H., Hirsa, A., Osterrieder, J., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in Credit Risk Management (SSRN Working Paper No. 3795322) [Preprint]. SSRN. https://doi.org/10.2139/ssrn.3795322

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3-4). https://doi.org/10.1145/3387166

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys, 55(13). https://doi.org/10.1145/3583558

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges (arXiv:2406.11903) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.11903

NSIA (2025). About Us - NSIA [Web publication]. Nigeria Sovereign Investment Authority. https://nsia.com.ng/about-us/

O'Leary, K. (2024). How Gulf Sovereign Wealth Funds Are Reshaping AI, Sports, and Renewable Energy - Chronograph [Web publication]. Chronograph.

https://www.chronograph.pe/how-gulf-sovereign-wealth-funds-are-reshaping-ai-sportsand-renewable-energy

OECD (2023). GENERATIVE ARTIFICIAL INTELLIGENCE IN FINANCE [Report]. Organisation for Economic Co-operation and Development. https://www.oecd.org/finance/generative-artificial-intelligence-in-finance.htm

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35. https://doi.org/10.52202/068431-2011

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2001). BLEU. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 311. https://doi.org/10.3115/1073083.1073135

Saves, P., Palar, P. S., Robani, M. D., Verstaevel, N., Garouani, M., Aligon, J., ... & Morlier, J. (2025). Surrogate modelling and explainable artificial intelligence for complex systems: A workflow for automated simulation exploration (arXiv:2510.16742) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.16742

Schneider, J. (2024). Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda. https://doi.org/10.1007/s10462-024-10916-x

Shabsigh, G. (2023). Generative Artificial Intelligence in Finance (IMF Fintech Notes No. 2023/006) [Report]. International Monetary Fund. https://doi.org/10.5089/9798400251092.063

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., ... Yang, D. (2022). WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 2322-2335. https://doi.org/10.18653/v1/2022.emnlp-main.148

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, 3784-3803. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Natarajan, V. (2022). Large Language Models Encode Clinical Knowledge. Nature, 620(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2

Sokol, K., & Vogt, J. E. (2024). What Does Evaluation of Explainable Artificial Intelligence Actually Tell Us?: A Case for Compositional and Contextual Validation of XAI Building Blocks. Conference on Human Factors in Computing Systems - Proceedings, 1. https://doi.org/10.1145/3613905.3651047

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... Gray, A. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, 2023. https://doi.org/10.48550/arXiv.2206.04615

SWFI (2024). Nigeria Sovereign Investment Authority (NSIA) - Sovereign Wealth Fund, Nigeria - SWFI [Web publication]. Sovereign Wealth Fund Institute. https://www.swfinstitute.org/profile/598cdaa60124e9fd2d05b968

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... Hashimoto, T. B. (2023). Alpaca: A Strong, Replicable Instruction-Following Model [Web publication]. Stanford Centre for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., ... Stojnic, R. (2022). Galactica: A Large Language Model for Science (arXiv:2211.09085) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2211.09085

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., ... Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2302.13971

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of Internal Medicine, 169(7), 467-473. https://doi.org/10.7326/M180850

UNESCO (2022). Recommendation on the Ethics of Artificial Intelligence [Report]. United Nations Educational, Scientific and Cultural Organization.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89-106. https://doi.org/10.1016/j.inffus.2021.05.009

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR (SSRN Working Paper No. 3063289) [Preprint]. SSRN. https://doi.org/10.2139/ssrn.3063289

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A MultiTask Benchmark and Analysis Platform for Natural Language Understanding. EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analysing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop, 353-355. https://doi.org/10.18653/v1/w18-5446

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. R. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1905.00537

Wang, M., Yao, Y., Xu, Z., Qiao, S., Deng, S., Wang, P., ... Zhang, N. (2024). Knowledge Mechanisms in Large Language Models: A Survey and Perspective (arXiv:2407.15017) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.15017

Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of Explainable Artificial Intelligence in Finance-a systematic review of Finance, Information Systems, and Computer Science literature. Management Review Quarterly, 74(2), 867-907. https://doi.org/10.1007/s11301-023-00320-0

Wu, X., Yao, W., Chen, J., Pan, X., Wang, X., Liu, N., & Yu, D. (2023). From Language Modelling to Instruction Following: Understanding the Behaviour Shift in LLMs after Instruction Tuning. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, 1, 2341-2369. https://doi.org/10.18653/v1/2024.naacl-long.130

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance (arXiv:2306.05443) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.05443 Xie, Q., Huang, J., Li, D., Chen, Z., Xiang, R., Xiao, M., ... & Ananiadou, S. (2024). FinNLPAgentScen-2024 shared task: Financial challenges in large language models -

FinLLMs. In Proceedings of the 6th Financial Narrative Processing Workshop. Association for Computational Linguistics.

Xu, S., Liu, S., Culhane, T., Pertseva, E., Wu, M. H., Semnani, S. J., & Lam, M. S. (2023). Fine-tuned LLMs Know More, Hallucinate Less with Few-Shot Sequence-to-Sequence Semantic Parsing over Wikidata. EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, 5778-5791. https://doi.org/10.18653/v1/2023.emnlp-main.353

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. High-Confidence Computing, 4(2). https://doi.org/10.1016/j.hcc.2024.100211

Yeo, W. J., Heever, W. V. D., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2023). A Comprehensive Review on Financial Explainable AI (arXiv:2309.11960) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.11960

Yoo, M. (2024). How much should we trust large language model-based measures for accounting and finance research? (SSRN Working Paper No. 4983334) [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4983334

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. 8th International Conference on Learning Representations, ICLR 2020. https://doi.org/10.48550/arXiv.1904.09675

Zhang, Y., Chen, Z., Fang, Y., Lu, Y., Li, F., Zhang, W., & Chen, H. (2023). Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering (arXiv:2311.06503) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2311.06503

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... Du, M. (2024). Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 15(2). https://doi.org/10.1145/3639372

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. (2023). A Survey of Large Language Models (arXiv:2303.18223) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zhao, W., Rush, A. M., & Goyal, T. (2025). Findings of the Association for Computational Challenges in Trustworthy Human Evaluation of Chatbots. Findings of the Association for Computational Linguistics, 3359-3365. https://aclanthology.org/2025.findingsnaacl.186/

Zhuang, Z., Chen, Q., Ma, L., Li, M., Han, Y., Qian, Y., ... Liu, T. (2023). Through the Lens of Core Competency: Survey on Evaluation of Large Language Models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.08815

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... Irving, G. (2019). Fine-Tuning Language Models from Human Preferences (arXiv:1909.08593) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1909.08593