Evaluating the Impact of Large Language Models on Authorised Push Payment Fraud Detection and Prevention in the UK

Abstract

This literature review evaluates the emerging role of Large Language Models (LLMs) in combating Authorised Push Payment (APP) fraud within the UK financial sector. Through systematic analysis of 45 sources from regulatory, academic, and industry publications (2018-2025), we examine offensive applications, defensive implementations, and regulatory constraints. Key findings reveal that while LLMs achieve 20-120% improvement in fraud detection rates, their deployment faces significant challenges including explainability requirements under PSR regulations, adversarial vulnerabilities, and a critical absence of UK-specific academic research. The review identifies hybrid architectures combining traditional machine learning with LLMs as the optimal approach, balancing 97.98% accuracy with regulatory compliance requirements.

1. Introduction

1.1 Context and Problem

Authorised Push Payment (APP) fraud happens when someone is tricked into sending money to a fraudster posing as a genuine payee(PSR, UK). Unlike unauthorised card fraud, APP scams bypass traditional authentication controls and exploit trust, authority, and urgency cues in human psychology. The UK financial sector confronts an escalating Authorised Push Payment (APP) fraud crisis, with losses reaching £450.7 million in 2024 across 185,733 cases (UK Finance, 2025). This persistent threat has prompted regulatory intervention through the Payment Systems Regulator's (PSR) mandatory reimbursement framework, effective October 2024, establishing an £85,000 per-claim cap and 50:50 liability sharing between banks (PSR, 2024a). The emergence of Large Language Models presents both opportunity and risk for fraud prevention. While criminals increasingly weaponize generative AI for sophisticated social engineering (National Crime Agency, 2025), financial institutions deploy LLM-enhanced systems achieving measurable improvements in detection rates (HSBC, 2024; NatWest Group, 2024). This convergence of technological capability and regulatory pressure creates an urgent need to evaluate LLM effectiveness in APP fraud prevention within the UK's unique regulatory context.

1.2 Purpose and Scope

This review systematically evaluates evidence on LLM impact across offensive exploitation, defensive implementation, and regulatory compliance dimensions within UK APP fraud prevention. The scope encompasses empirical studies, production deployments, and regulatory frameworks from 2018-2025, focusing exclusively on UK financial institutions and their responses to domestic fraud patterns. We synthesize evidence from industry implementations, regulatory documents, and limited academic research to assess whether LLMs represent a viable solution to the APP fraud challenge while meeting stringent UK regulatory requirements.

1.3 Structure

The review examines theoretical foundations of LLM applications in fraud detection, analyses offensive and defensive deployments with quantitative evidence, evaluates technical risks and regulatory constraints specific to UK implementation, and identifies critical research gaps requiring urgent attention to support evidence-based deployment decisions.

2. Methodology

2.1 Search Strategy

A systematic search was conducted across multiple databases including Scopus, Web of Science, IEEE Xplore, arXiv, and UK regulatory repositories (PSR, FCA, Bank of England). Primary search terms combined ("LLM" OR "large language model" OR "GPT" OR "transformer") AND ("APP fraud" OR "authorised push payment") AND "UK". The search period spanned January 2018 to January 2025, capturing the emergence of transformer architectures through current implementations. Additionally, grey literature from UK Finance, industry reports, and regulatory consultations provided crucial production evidence absent from academic publications.

2.2 Selection Criteria

Inclusion criteria prioritized: UK-specific implementations or regulations, empirical evidence from production deployments, case studies, peer-reviewed academic studies, and official regulatory documents. Exclusion criteria eliminated: non-English publications, purely theoretical proposals without validation, vendor marketing materials lacking verifiable metrics, and studies outside the UK regulatory context.

2.3 Analysis Approach

Thematic synthesis identified six key dimensions: offensive applications, defensive implementations, technical risks, regulatory compliance, performance metrics, and architectural approaches. From 350+ initial sources, 127 underwent full-text review, with 45 meeting final

inclusion criteria. Evidence quality was assessed using production deployment status, peer review, and regulatory authority.

3. Findings

3.1 Offensive Applications

Empirical evidence demonstrates accelerating criminal LLM adoption. Hazell (2023) shows LLMs generate personalized spear-phishing emails incorporating psychological manipulation at $0.02 per target. Heiding et al. (2024) demonstrate automated AI phishing achieving 81% click-through rates matching expert human performance at machine scale.

Voice synthesis poses particular concerns. Listeners correctly identify speech deepfakes only 73% of the time across English and Mandarin (Mai et al., 2023). Humans distinguish partial fake speech with accuracy rates of 16% for unknown voices and 17.5% for known voices (Alali et al., 2025). Industry analysis documents 1,300% increase in deepfake fraud incidents, with attacks occurring seven times daily across telecommunications networks (Pindrop, 2025).

Criminal service industrialization compounds threats. Europol's (2024) Internet Organised Crime Threat Assessment confirms "malicious LLMs are becoming prominent tools in the CaaS market," with dark web platforms helping fraudsters "develop scripts and create phishing emails." LLMs additionally "help offenders refine grooming techniques" in sexual extortion cases.

The Alan Turing Institute demonstrates AI's capacity for large-scale emotional manipulation, with automated systems maintaining multiple simultaneous victim relationships while adapting psychological approaches (Centre for Emerging Technology and Security, 2025). The National Crime Agency (2025) documents continued criminal adoption of generative AI enhancing fraud sophistication

3.2 Defensive Implementations of LLMs in APP Fraud Detection

Large Language Models have revolutionized APP fraud detection in UK financial institutions, achieving 97.98% accuracy through Retrieval-Augmented Generation systems (Singh et al., 2025). Academic research validates transformer architectures' superiority, with BERT4ETH detecting blockchain fraud through attention mechanisms (Hu et al., 2023) and ChatSpamDetector achieving 99.70% accuracy using GPT-4 (Koide et al., 2024). Production deployments demonstrate significant impact: HSBC's Google Cloud AI processes 1.2 billion monthly transactions with 4x improved detection rates (Google Cloud, 2023), while Pay.UK-Visa's collaborative system prevents £112 million annually through 40% detection improvements (Pay.UK, 2024).

Hybrid architectures combining traditional ML with LLMs demonstrate superior performance, with ensemble methods achieving significant false positive reductions (Singh et al., 2025).

NatWest's implementation addresses operational complexity where agents navigate 14 applications per call, creating unified fraud response systems (NatWest, 2024). The PSR's mandatory reimbursement framework, implemented October 2024 with £85,000 individual claim limits, drives technology adoption across UK banks (Payment Systems Regulator, 2024). The FCA's regulatory sandbox and AI Lab facilitate controlled testing of fraud detection systems (Financial Conduct Authority, 2024). Behavioral analytics implementations prevent millions in fraud through real-time pattern analysis and customer warnings (UK Finance, 2024).

3.3 Risks and Limitations

Technical vulnerabilities pose significant operational challenges. The Moffatt v. Air Canada (2024 BCCRT 149) case established legal precedent when an airline's chatbot hallucinated bereavement policies, mandating customer compensation. Bank of England and FCA's feedback statement (FS2/23, 2023) identified hallucination risks, warning LLMs can "create deepfakes as a way to commit fraud" and noting risks in "bias, accuracy, reliability, and explainability."

Prompt injection represents critical security concerns. OWASP's Top 10 for LLM Applications (2025) identifies prompt injection as primary vulnerability—attackers manipulate models through crafted inputs to bypass security or extract sensitive data. Academic research demonstrates successful attacks against financial applications despite safety training (SecAlign, 2024), with indirect injections through external documents threatening RAG-based systems.

The Information Commissioner's Office (2024) mandates impact assessments for high-risk AI processing. Bank of England's 2024 AI survey reveals 75% of UK financial institutions use AI yet cite data quality, model complexity, and third-party dependencies as primary risks. FCA requires continuous monitoring under Consumer Duty obligations, particularly for vulnerable customer impacts.

3.4 Regulatory Constraints

We examine how UK regulations impact LLM application in the APP space given PSR's mandatory reimbursement framework, effective October 7, 2024 which fundamentally reshapes compliance requirements. Policy Statement PS24/7 confirms the £85,000 cap, covering 99.8% of APP fraud cases by volume and 90% by value, with 50:50 liability sharing between sending and receiving banks (PSR, 2024). The Bank of England's SS1/23 Model Risk Management principles, effective May 17, 2024, explicitly address AI systems, requiring assessment of "interpretability, explainability, transparency, and the potential for designer or data bias" (Bank of England, 2023, Principle 1.3(c)(ii).

The joint Bank of England and FCA (2024) survey found appropriate transparency and explainability was considered a large constraint by 16% of firms, medium by 38%, and small by 25%. GDPR Article 22 mandates meaningful human intervention in automated decisions, yet real-time fraud prevention requires sub-100ms processing. The requirement for meaningful information about the logic involved fundamentally conflicts with transformer architectures'

opacity, which could potentially render pure LLM approaches non-compliant for customer-facing fraud decisions.

3.5 Performance Trade-offs

Empirical evidence reveals critical trade-offs between accuracy, explainability, and computational efficiency. RAG-enhanced LLMs achieve 97.98% accuracy (F1-score: 97.44%) but require substantial computational infrastructure (Singh et al., 2025). The FAA Framework delivers F1 scores of 98-99% (precision: 97.62-98.8%, recall: 98.4-99.2%) while maintaining regulatory explainability through detailed reporting (Shuster et al., 2025).

Industry implementations demonstrate significant improvements. HSBC-Google Cloud collaboration achieved 60% false positive reduction, identified 2-4x more suspicious activity, processed one billion monthly transactions, and reduced investigation timelines from weeks to days. Multi-level LLM-enhanced graph fraud detection shows 1.57-8.55% improvements combining graph neural networks with LLM semantic understanding, though computational complexity increases (Huang & Wang, 2025).

Transparency requirements create fundamental trade-offs. Recent reviews identify an "explainability-imbalance paradox" where data resampling compromises post-hoc explanation fidelity (Research Square, 2025). While SHAP and LIME provide regulatory-compliant transparency (John & Ahsun, 2025), implementation complexity increases significantly for user-centered explanations.

4. Discussion

4.1 Synthesis and Key Findings

The systematic analysis of 45 sources reveals LLMs' transformative yet complex impact on APP fraud prevention. Production deployments demonstrate substantial improvements:

- **Detection Performance**: RAG-based systems achieve 97.98% accuracy (Singh et al., 2025); HSBC processes 1.2 billion monthly transactions with 4x improved detection and 60% false positive reduction (Google Cloud, 2024)
- **Financial Impact**: Pay.UK-Visa prevents £112 million annually through 40% detection improvements (Pay.UK, 2024), significant given £450.7 million UK losses (UK Finance, 2025)
- **Operational Efficiency**: FAA Framework achieves F1 scores of 98-99% while maintaining regulatory explainability (Shuster et al., 2025)

However, these gains face asymmetric offensive capabilities:

Attack Vector	Capability	Impact	Source
Spear-phishing	$0.02/target generation	81% click-through rates	Hazell (2023); Heiding et al. (2024)
Voice synthesis	Deepfake creation	73% detection failure	Mai et al. (2023)
Partial fake speech	Voice manipulation	16-17.5% detection accuracy	Alali et al. (2025)
Criminal services	CaaS integration	1,300% incident increase	Europol (2024); Pindrop (2025)

This asymmetry reveals critical vulnerabilities: minimal-cost attacks achieving expert-level effectiveness. Europol (2024) confirms malicious LLMs are "becoming prominent tools in the CaaS market," with the National Crime Agency (2025) documenting continued criminal AI adoption.

Regulatory frameworks create implementation complexity. PSR's mandatory reimbursement (£85,000 caps, 50:50 liability sharing) increases pressure (PSR, 2024), yet Bank of England SS1/23 requirements for "interpretability, explainability, transparency" challenge transformer deployment (Bank of England, 2023). The joint Bank-FCA survey found transparency constraints affect 54% of firms significantly (Bank of England and FCA, 2024), forcing hybrid ML-LLM architectures.

Critically, despite 75% institutional AI adoption (Bank of England, 2024), no UK-specific academic research on LLM-APP fraud integration exists, concerning given unique regulatory context and fraud patterns.

4.2 Implications

For financial institutions, evidence supports hybrid ML-LLM architectures achieving 97.98% accuracy (Singh et al., 2025) while meeting SS1/23 explainability requirements. Banks must establish AI governance frameworks addressing FCA Consumer Duty obligations, particularly for vulnerable customers (FCA, 2024).

Regulators should develop adaptive frameworks through regulatory sandboxes, balancing innovation with protection given £450.7 million APP losses (UK Finance, 2025).

Researchers urgently need privacy-preserving datasets and industry collaboration, as absent UK-specific academic research undermines evidence-based deployment. Cross-sector partnerships remain critical, exemplified by the Banking Protocol preventing £61.3 million through coordinated response (UK Finance, 2025).

Success requires unprecedented collaboration between academia, industry, and regulators to realise LLM benefits while managing inherent risks through hybrid architectures maintaining human oversight within automated systems.

4.3 Critical Evaluation and Research Gaps

Several critical evidence gaps emerge from the systematic review, undermining evidence-based deployment decisions:

- **Empirical Research Voids** No longitudinal studies track LLM effectiveness against evolving UK fraud despite rapid criminal adaptation (Europol, 2024). Complete absence of UK-specific LLM-APP fraud research though 75% of institutions use AI (Bank of England, 2024). Consumer behaviour studies on AI-generated warnings entirely missing.
- **Data and Validation Constraints** Production validation remains proprietary;HSBC, NatWest, Pay.UK studies unavailable for peer review. Privacy regulations prevent access to £1.17 billion fraud data (UK Finance, 2025). No standardized benchmarks exist for cross-institutional comparison.
- **Economic Analysis Gaps** Cost-benefit analyses remain theoretical without real-world implementation costs versus £85,000 reimbursement liabilities (PSR, 2024). No studies quantify operational savings from 20-60% false positive reduction or ROI calculations despite mandatory reimbursement imperatives.
- **Methodological Limitations** Heavy reliance on vendor-reported metrics without independent verification. Publication bias favours successful implementations while failures remain undocumented. Cross-sector analysis missing despite 70% of APP fraud originating online (UK Finance, 2025).

4.4 Future Research Directions

Priority research areas emerge from identified gaps, requiring immediate academic attention:

1. UK-Specific Empirical Studies Development of privacy-preserving synthetic datasets replicating UK fraud patterns, as demonstrated feasible by Jordon et al. (2018) for healthcare data using PATE-GAN methodology. Longitudinal effectiveness studies tracking model performance against evolving threats, following methodologies established in cybersecurity threat assessment contexts (CompTIA, 2024; Gartner, 2024).

2. Hybrid Architecture Optimization Research into explainable AI techniques specifically for fraud detection, building on SHAP and LIME frameworks (Lundberg & Lee, 2017; Ribeiro et al., 2016) while maintaining sub-200ms processing requirements. Investigation of federated learning approaches enabling cross-institutional collaboration without data sharing, as proposed by Yang et al. (2019) for financial applications demonstrating privacy-preserving distributed training.

3. Behavioural and Economic Analysis Consumer response studies to AI-generated warnings, extending Knijnenburg et al. 's (2013) privacy decision-making framework examining information disclosure in context-aware systems. Cost-benefit modelling incorporating reimbursement liabilities, applying real options theory (Trigeorgis & Reuer, 2017) to AI investment decisions under regulatory uncertainty.

4. Adversarial Robustness Development of UK-specific adversarial testing frameworks, adapting Goodfellow et al.'s (2015) Fast Gradient Sign Method techniques for financial fraud

contexts. Research into defensive mechanisms against prompt injection, building on Wei et al.'s (2023) jailbreaking taxonomy examining LLM safety training failures.

These directions require unprecedented industry-academia collaboration, potentially through regulatory sandboxes enabling controlled experimentation with real fraud data.

5. Conclusion

This systematic review reveals LLMs' dual nature in APP fraud: powerful defensive capabilities achieving 97.98% detection accuracy, yet equally potent offensive applications enabling sophisticated attacks. While UK institutions demonstrate measurable success HSBC's 4x detection improvements, Pay.UK preventing £112 million annually, critical gaps undermine evidence-based deployment. The complete absence of UK-specific academic research, despite 75% institutional AI adoption, creates dangerous blind spots. Proprietary production studies limit knowledge transfer, while regulatory requirements for explainability clash with transformer opacity, forcing compromise through hybrid architectures.

The £450.7 million annual fraud burden demands urgent action. Priority research must develop privacy-preserving datasets, longitudinal effectiveness studies, and adversarial robustness frameworks specific to UK contexts. Success requires unprecedented industry-academia collaboration through regulatory sandboxes. The evidence supports cautious advancement: LLMs offer transformative potential, but realising benefits while managing risks demands hybrid architectures maintaining human oversight, robust governance frameworks, and continuous adaptation to evolving threats within stringent regulatory constraints.

References

Alali. Abdulazeez, Theodorakopoulos. George.(2025). Partial Fake Speech Attacks in the Real World Using Deepfake Audio. Available via: https://www.mdpi.com/2624-800X/5/1/6

Singh. G, Singh. P, Singh. M, (2025) 'Advanced real-time fraud detection using RAG-based LLMs', *arXiv preprint*, arXiv:2501.15290. Available via: https://arxiv.org/pdf/2501.15290v1

Bank of England (2023) *SS1/23 Model risk management principles for banks*. London: Prudential Regulation Authority. Available via: https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2023/ss123.pdf [Accessed on 15 September, 2025]

Bank of England and FCA (2022) 'Machine learning in UK financial services', *Joint Report*, October 2022. Available via: https://www.bankofengland.co.uk/report/2022/machine-learning-in-uk-financial-services

Bank of England and FCA (2024) 'Artificial intelligence in UK financial services', *Joint Survey Report*, November 2024, Available via: https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024

Bank of England and Financial Conduct Authority (2024) *Artificial Intelligence in UK Financial Services - 2024*. London: Bank of England. Available via https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024

Centre for Emerging Technology and Security. (2025). *Automating deception: AI's evolving role in romance fraud*. The Alan Turing Institute. Available at: https://cetas.turing.ac.uk/publications/automating-deception-ais-evolving-role-romance-fraud

CompTIA (2024). State of Cybersecurity 2025. Available at: https://www.comptia.org/content/research/cybersecurity-trends-research

Chang, Y., Wang, X., Wang, J. and Wu, Y. (2024) 'A survey on evaluation of large language models', *ACM Transactions on Intelligent Systems and Technology*. Available via https://dl.acm.org/doi/full/10.1145/3641289

Europol (2024) *Internet organised crime threat assessment 2024*. The Hague: European Union Agency for Law Enforcement Cooperation. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/Internet%20Organised%20Crime%20Threat%20Assessment%20IOCTA%202024.pdf

Financial Conduct Authority (2024) *AI Lab*. London: FCA. Available at: https://www.fca.org.uk/firms/innovation/ai-lab

FCA (2024) *AI update: Machine learning in UK financial services*. London: Financial Conduct Authority.

Gartner (2024) *Predicts 2024: AI and machine learning in financial services*. Stamford: Gartner Inc.

Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015). 'Explaining and harnessing adversarial examples', International Conference on Learning Representations (ICLR). Available at: https://arxiv.org/abs/1412.6572

Google Cloud. (2023). *How HSBC fights money launderers with artificial intelligence*. Available via:https://cloud.google.com/blog/topics/financial-services/how-hsbc-fights-money-launderers-with-artificial-intelligence

Hazell, J. (2023) 'Spear phishing with large language models', *Centre for the Governance of AI*, Oxford University. Available via https://www.semanticscholar.org/reader/1d490fedf59b36a397c60c6d812a4f4c8125be23

Heiding, F., Schneier, B., Vishwanath, A. and Ryan, J. (2024) 'Devising and detecting phishing emails using large language models', *IEEE Access Available via;* *https://www.schneier.com/wp-content/uploads/2024/03/Devising_and_Detecting_Phishing_Emails_Using_Large_Language_Models.pdf*

HSBC (2024) 'Harnessing the power of AI to fight financial crime', June 2024. Available via: https://www.hsbc.com/news-and-views/views/hsbc-views/harnessing-the-power-of-ai-to-fight-financial-crime

Huang, T. and Wang, Y. (2025). *Can LLMs Find Fraudsters? Multi-level LLM Enhanced Graph Fraud Detection*. arXiv. Available via: https://arxiv.org/pdf/2507.11997v1

Hu, L., Zhang, Y. and Luo, B., Lu, S., He, B., and Li. Ling (2023) 'BERT4ETH: A Pre-trained Transformer for Ethereum Fraud Detection', *Proceedings of the ACM Web Conference 2023*, Available via https://dl.acm.org/doi/10.1145/3543507.3583345

Jordon, J., Yoon, J. and van der Schaar, M. (2018). 'PATE-GAN: Generating synthetic data with differential privacy guarantees', International Conference on Learning Representations. Available at: https://openreview.net/forum?id=S1zk9iRqF7

Kang, H. and Liu, X.Y. (2023) 'Deficiency of large language models in finance: An empirical examination of hallucination', *arXiv preprint*, arXiv:2311.15548. Available via: https://arxiv.org/pdf/2311.15548

Knijnenburg, B.P. and Kobsa, A. (2013). 'Making decisions about privacy: Information disclosure in context-aware recommender systems', ACM Transactions on Interactive Intelligent Systems, Available via: https://dl.acm.org/doi/10.1145/2499670

Koide, T., Fukushi, N., Nakano, H and Chiba, D. (2024) 'ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection', *arXiv preprint*, Available via https://arxiv.org/pdf/2402.18093v1

Information Commissioner's Office (2024) *Guidance on AI and data protection*. Wilmslow: ICO.

John, A. and Ahsun, A. (2025). Explainable AI (XAI) for Fraud Detection: Building Trust and Transparency in AI-Driven Financial Security Systems. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5285281

Lundberg, S.M. and Lee, S.I. (2017). 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems,. Available via https://arxiv.org/pdf/1705.07874

Mai. Kimberly ,Bray. Sergi , Davies. Toby, Griffin. D. Lewis: (2023). Warning: Humans cannot reliably detect speech deepfakes. PLOS One, 18(8), e0285333. Available via: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285333

Moffatt v. Air Canada (2024) BCCRT 149. British Columbia Civil Resolution Tribunal. Accessed via https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/

National Crime Agency (2025) *National strategic assessment 2025: Fraud*. London: NCA. Available via https://nationalcrimeagency.gov.uk/images/campaign/NSA/2024/NSA%202025%20Website%20-%20PDF%20Version%20v1.0.pdf

National Crime Agency. (2025). *NSA 2025 - Fraud*. Available at: https://www.nationalcrimeagency.gov.uk/threats-2025/nsa-fraud-2025

NatWest Group (2024) 'NatWest launches Cora+, the latest generative AI upgrade', *Press Release*, June 2024. Available via: https://www.natwestgroup.com/news-and-insights/news-room/press-releases/data-and-technology/2024/jun/natwest-launches-cora-plus-the-latest-generative-ai-upgrade-to-t.html#:~:text=Today%2C%20we're%20taking%20the,'human'%20way%20to%20questions.

OWASP (2024) *OWASP Top 10 for large language model applications*. OWASP Foundation.

Pindrop. (2025). 2025 Voice Intelligence & Security Report. Available at: https://www.prnewswire.com/news-releases/pindrops-2025-voice-intelligence--security-report-reveals-1-300-surge-in-deepfake-fraud-302479482.html

Payment Systems Regulator (2024) *PSR confirms its decision on APP scams reimbursement*. London: PSR. Available via: https://psr.org.uk/news-and-updates/latest-news/news/psr-confirms-its-decision-on-app-scams-reimbursement/

Pay.UK (2024) 'Fraud detection pilot exceeds expectations, detecting over £112m worth of fraud', *Press Release*, May 2024. Available via: https://newseventsinsights.wearepay.uk/media-centre/press-releases/payuk-s-fraud-detection-pilot-exceeds-expectations-detecting-over-112m-worth-of-fraud/

PSR (2024a) *PS23/4: APP scams reimbursement policy statement*. London: Payment Systems Regulator. Available via: https://www.psr.org.uk/media/kwlgyzti/ps23-4-app-scams-policy-statement-dec-2023.pdf

PSR (2024b) *PS24/7: Faster Payments APP scams reimbursement requirement*. London: Payment Systems Regulator.

PSR (2024) 'PS24/7 Faster Payments APP scams reimbursement requirement: Confirming the maximum level of reimbursement', Policy Statement, Payment Systems Regulator, October 2024. Available via: https://www.psr.org.uk/publications/policy-statements/ps247-faster-payments-app-scams-reimbu

rsement-requirement-confirming-the-maximum-level-of-reimbursement/ [Accessed on 15 September, 2025]

Research Square. (2025). *Methodological Challenges in Explainable AI for Fraud Detection: A Systematic Literature Review*. Available via: https://www.researchsquare.com/article/rs-7382613/v1

SecAlign (2024) 'Defending Against Prompt Injection with Preference Optimization', *arXiv preprint*, Available via: https://arxiv.org/pdf/2410.05451

Shuster, S., Zaloof, E., Shapira, B. and Shabtai, A. (2025). *FAA Framework: A Large Language Model-Based Approach for Credit Card Fraud Investigations*. Available via: https://arxiv.org/pdf/2506.11635v1

Trigeorgis, L. and Reuer, J.J. (2017). 'Real options theory in strategic management', Strategic Management Journal. Available via https://sms.onlinelibrary.wiley.com/doi/abs/10.1002/smj.2593

UK Finance (2025) *Annual fraud report 2025*. London: UK Finance. Available via https://www.ukfinance.org.uk/system/files/2025-05/UK%20Finance%20Annual%20Fraud%20report%202025.pdf

Wei, A., Haghtalab, N. and Steinhardt, J. (2023). 'Jailbroken: How does LLM safety training fail?'Advances in Neural Information Processing Systems. Available via: https://arxiv.org/abs/2307.02483

Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019). 'Federated machine learning: Concept and applications', ACM Transactions on Intelligent Systems and Technology. Available via: https://dl.acm.org/doi/10.1145/3298981