Conversational AI Chatbots in FinTech CRM: Impact on Customer Satisfaction and First-Contact Resolution Rates

Vignesh MK[1*] Thanikachalam Vadivel[2]
Sriram Ragul Thambi Vijay[3] & Purna Prasad Arcot[4]

[1]Assistant Professor, Department of Common Core Curriculum, CMR University, Bengaluru, Karnataka, India.
[2]Assistant Professor, School of Management, CMR University, Bengaluru, Karnataka 562149, India.
[3]Assistant Professor, School of Management, CMR University, Bengaluru, Karnataka 562149, India.
[4]Professor & Director, School of Management, Lakeside Campus, CMR University, Bengaluru, Karnataka, 562149, India
vignesh.m@cmr.edu.in

ABSTRACT. This research examines the causal effect of deploying Conversational Artificial Intelligence (CAI) chatbots in FinTech Customer Relationship Management (CRM) systems on two key metrics of service quality: customer satisfaction (CSAT) and First-Contact Resolution (FCR). Using operational data from 47 FinTech companies across six Asian and Southeast Asian markets, involving 3,842,917 customer service interactions over the past 30 months (January 2022 to June 2024), we apply a staggered difference-in-differences (DiD) design, which leverages staggered CAI deployment across firms and product lines, a regression discontinuity (RD) design around the minimum confidence score required for chatbots to provide an answer, and structural equation modelling to examine the direct and indirect effects of CAI capabilities on service outcomes. Our CAI taxonomy categorises deployed systems into three generations: rule-based chatbots (RBC), retrieval-augmented generation systems (RAG-CAI), and large language model-based agents (LLM-CAI). Findings show that LLM-CAI deployment raises CSAT (on a 100-point scale) by 14.7 points compared to pre-deployment levels ($p < 0.001$) and FCR by 22.3 percentage points ($p < 0.001$), but RBC deployment only has a marginal effect on CSAT (+2.1 points, $p = 0.31$) and a moderate effect on FCR (+8.4 percentage points, $p < 0.01$). We show that the effects of LLM-CAI deployment on CSAT are mediated by 71.4% by resolution speed, response personalisation and handling of complex queries. There is significant customer demographic diversity: Gen Z (18-27 years of age), and customers with high digital literacy have CSAT responses 31.2% higher than the sample mean, but older customers (60+) and users with low digital literacy have neutral or slightly negative CSAT responses to CAI, suggesting a risk of digital divide in strategies for full migration to CAI. Importantly, human escalation design - specifically, whether the system offers smooth human handoff - moderates the impact of CAI failure on CSAT: the CSAT penalty of failed chatbot experiences is 61.4% lower with seamless human handover. Our findings have implications for FinTech customer relationship management (CRM), digital service design and regulation of AI-based financial services.

Keywords: Conversational AI, Chatbot, FinTech, CRM, Customer Satisfaction, First-Contact Resolution, Large Language Models, Retrieval-Augmented Generation, Service Quality, Digital Banking, Natural Language Processing. .

H. L. Chidananda et al. (eds.), *Proceedings of the 2nd International Conference on Innovations and Challenges in Financial Technology (ICICFT 2025)*, Advances in Economics, Business and Management Research 398,
https://doi.org/10.2991/978-94-6239-772-9_24

1 INTRODUCTION

Conversational artificial intelligence (AI) technology for enterprise customer service has found its first major use case in the financial technology (FinTech) industry due to the combination of high volumes, taxonomies and the market imperative for 24-hour service. US$ 42.7 billion was invested globally in 2023 in AI-powered financial services customer experience technology infrastructure, with conversational AI experiencing the highest growth rate (34.2%) [1]. FinTech companies - ranging from online banks and peer-to-peer lenders, to insurance tech and robo-advisory startups - have a unique service problem: customer inquiries often combine transactional detail (balance inquiries, payment status, account alerts) with complex financial advice (product suitability, complaint handling, regulatory disclosures) that typically requires human expertise.

AI-powered chatbots present a strong value proposition for FinTech customer relationship management (CRM): they remove waiting times, are available 24/7 regardless of time zones, ensure consistency in customer conversations, and can handle thousands of simultaneous customers with marginal costs approaching zero [2]. But the use of AI in financial customer service raises risks and trade-offs in service quality that have not been empirically quantified. Customer satisfaction, the ultimate measure of success in providing quality service, may be improved by the speed and convenience of conversational AI but impaired by its perceived lack of human qualities or inability to comprehend financial questions, or by the frustration of talking to a system that does not solve the customer's problem [3]. First-Contact Resolution, the holy grail of operational customer service metrics, is the probability that a customer issue is completely resolved during the first interaction with no need to re-contact the service provider, a measure that tests the ability of AI systems to replace human judgment, rather than simply deflect interactions [4].

The rapid pace of conversational AI development has resulted in a generational divide in deployed CRM systems that has not previously been considered. First-generation rule-based chatbots (RBC) use decision trees and keyword matching, with low but predictable natural language processing (NLP) capabilities. Second-generation retrieval-augmented generation chatbots (RAG-CAI) retrieve relevant documents from a knowledge base using neural matching, and generate human-like responses using language models. Third-generation large language model agents (LLM-CAI) leverage foundation models such as GPT-4, Claude, and Gemini fine-tuned on private FinTech data, to support multi-turn reasoning, intent disambiguation and procedural workflows in core banking system integrations [5]. These three generations represent not just a technical advance, but also a difference in ability to deal with the complexity of queries, the emotional tone, and the regulatory accuracy required in FinTech customer service - differences that our empirical approach seeks to measure.

Prior research on chatbots for customer service has three key shortcomings. First, existing research is mostly based on controlled experiments and single-case studies, which casts doubts over its external validity [6]. Second, previous research does not account for chatbot generations, conflating all conversational AI (CAI) as a single

treatment, and mixing the effects of rule-based chatbots and large language models [7]. Third, the demographic differences of customer reactions to AI service - a factor with direct implications for financial inclusion and regulatory compliance - have only been described but not identified [8]. Our paper overcomes all three shortcomings with a large-scale, multi-firm, longitudinal empirical approach that takes advantage of the staggered introduction of CAI systems across firms and product lines as a natural experiment.

Our paper offers four empirical findings. First, we estimate the CSAT and FCR effects of each generation of CAI systems separately, showing that the LLM-CAI effect is qualitatively different from and stronger than the effects of earlier generations - a differentiation that disappears when CSAT effects of different types of chatbots are combined. Second, we pin down the service quality pathways (speed, personalisation and complexity management) through which LLM-CAI drives CSAT improvement to offer design insights. Third, we show the demographic diversity in CAI service response - customers with low digital literacy and older customers are at risk of CAI migration-induced service quality loss.

Section 4 presents the econometric methodology. Section 5 reports empirical results, including three data visualisations with detailed interpretation. Section 6 discusses findings, limitations, and policy implications. Section 7 concludes with strategic and regulatory recommendations.

2 LITERATURE REVIEW

2.1 Service Quality Theory and the Digital Service Encounter

The original service quality measurement framework is the SERVQUAL model proposed by Parasuraman, Zeithaml and Berry [9], which defines five dimensions (tangibility, reliability, responsiveness, assurance, and empathy) on which customer perceptions of service quality are evaluated against expectations to assess satisfaction. The SERVQUAL model has been adapted for internet-based and AI-mediated service encounters, as some of the dimensions (tangibility, empathy) do not directly apply to automated service provision. Bitner, Brown and Meuter [10] introduced the notion of the technology-mediated service encounter, suggesting that customers' perceptions of AI-powered service include task-completion outcomes and an affective evaluation of the service encounter itself.

Within banking, the dimensions of service quality that impact customer satisfaction have been well investigated for traditional channels. Parasuraman et al. [11] extended their model to create E-S-QUAL (electronic service quality), which reflects the quality of online banking in terms of efficiency, fulfilment, system availability and privacy. But the E-S-QUAL framework was developed prior to the advent of conversational AIs and does not explicitly consider the natural language user interface layer conversational chatbots provide - a layer that raises new dimensions of service quality such as language understanding, adequate response, context maintenance, and emotion appropriateness

[12]. This research operationalises a conversational AI service quality framework that builds on E-S-QUAL to include these dimensions.

2.2 Chatbot Technology Evolution and CRM Applications

The usage of conversational agents in customer service has been evolving through three technological generations since the first commercial use of chatbots in the early 2000s [13]. The first generation of chatbots (rule-based chatbots) are based on pre-defined decision trees and pattern matching rules, and are able to respond to a limited set of expected questions with high accuracy, but fail completely outside of their programming. In financial services they have been used primarily for deflecting FAQs, balance inquiries, and simple transaction status [14].

Retrieval-augmented generation (RAG) conversational systems, the second wave of chatbots, take dense passage retrievals from a corpus of a knowledge base, and combine these with generative language model generation, allowing more natural language responses to queries, while limiting generation to factually accurate information [15]. The RAG approach is a major step forward for FinTech CRM because it can be anchored on private product documentation, regulatory statements, and policy manuals - mitigating the hallucination problem that constrains the use of raw generative AI in regulated financial services environments. As of 2023, various major FinTech companies such as Nubank, N26 and Monzo have announced RAG-enabled CRM systems [16].

The third and most advanced generation of conversational agents use instruction-tuned foundation models capable of multi-turn reasoning, tool and system integration: LLM-based conversational agents (LLM-CAI). LLM-CAI systems connected with banking APIs enable procedural reasoning - the ability to not only respond to queries regarding a payment, but to initiate, edit, or cancel a payment within the same turn [17]. Shridhar et al. [18] found LLM agents fine-tuned on financial service transcripts score 94.3% on intent classification for complex multi-intent queries, versus 71.2% for RAG and 43.8% for rule-based methods on the same dataset. In our research, we extend these capability differences to customer experience and business performance measures.

2.3 Customer Satisfaction Determinants in AI-Mediated Financial Services

Theories of customer satisfaction with AI-mediated services have been considered from a number of perspectives. Expectation-Confirmation Theory (ECT), which has been applied to IS continuance by Bhattacherjee [20] from Oliver [19], proposes that satisfaction arises from confirmation or disconfirmation of pre-interaction expectations: customers who expect AI to be weak will be more easily satisfied by good performance, while those who expect it to be strong will be dissatisfied by failure. This theory accounts for the demographic variance we observe empirically, with digitally experienced customers (especially young customers) having more accurate beliefs about LLM-CAI capabilities and thus being better able to assess AI performance.

The Technology Acceptance Model (TAM) [21] and its extension UTAUT2 [22] offer further insights into user satisfaction with AI-based services, focusing on the role of

perceived ease of use and usefulness in determining user attitude. These models need revision in the involuntary service context - when customers seek FinTech support to solve a problem rather than voluntarily - where ease of use influences satisfaction through frustration avoidance rather than intent to adopt, and usefulness is measured by problem resolution effectiveness rather than increased productivity [23].

The empirical research on CSAT with financial chatbots is emerging but limited. Yen and Cheng [24] asked 384 bank customers to rate their chatbot satisfaction, identifying intelligence, naturalness and resolution effectiveness as the three most important factors. Cheng and Jiang [25] performed an experiment in a Chinese commercial bank and observed that CSAT increased by 11.3 on a 100-point scale when chatbots were available. These studies do not use causal identification techniques, multi-firm analysis, or account for chatbot generations - all of which are key features of our research.

2.4 First-Contact Resolution: Operational Significance and Determinants

First-Contact Resolution is the proportion of customer service contacts in which the customer's problem is completely resolved in a single contact [4]. FCR is the key operational metric in omnichannel customer service management, given that it captures both the quality of the customer experience (adequacy of the solution) and the efficiency of the operation (avoiding contact). In the financial services sector, FCR in traditional voice-based channels (e.g. call centres) is at the levels of 65-72% [27], while FCR in digital chat channels is slightly higher (68-74%) because text-based communication is asynchronous and enables agents to take more time to find information.

The factors affecting FCR in AI channels are the complexity of the query, the integration of the AI system and the escalation path design. Qiu and Benbasat [28] verified in experimental studies that the quality of chatbot responses - that is, the correct interpretation of the query intent and the ability to retrieve relevant resolution steps - is the most influential factor in determining user-perceived resolution. In financial services, the regulatory need to provide accurate, complete and verifiable information (such as disclosures, eligibility, formal complaint processing) creates a higher-resolution complexity that pushes the limit of AI systems' factual accuracy capabilities [30].

2.5 Demographic Heterogeneity and Digital Service Equity

The distributional consequences of AI deployment in customer service — specifically, whether AI systems serve all demographic groups equally well — represent a growing concern in both the academic and policy literatures. Czaja et al. [31] documented a strong age gradient in technology comfort and digital self-efficacy, with adults over 65 reporting significantly lower ability to navigate conversational interfaces, higher frustration with automated responses, and stronger preference for human service agents. In the financial services context, elderly customers are disproportionately reliant on in-person and telephone service channels and are more likely to present complex service needs that test the limits of AI query handling [32].

Financial literacy — closely correlated with income and education — predicts the ability to accurately evaluate AI-provided financial information and to detect errors, making low-financial-literacy customers more vulnerable to both overreliance on incorrect AI responses and frustration when AI responses are inappropriately complex [35]. Our demographic heterogeneity analysis provides the first multi-dimensional causal decomposition of these effects in a large-scale field dataset.

3 DATA AND SAMPLE

3.1 Data Sources and Sample Construction

The primary data source is the FinTech CRM Conversational AI Performance Dataset (FCAPD), constructed through data-sharing agreements with 47 FinTech firms operating across Singapore, Malaysia, Indonesia, the Philippines, India, and Hong Kong. Participating firms span four product categories: digital banking and neobanks (n = 18), payments and remittance platforms (n = 12), lending and credit platforms (n = 10), and insurance technology providers (n = 7). Firm-level CRM data were provided under strict data governance agreements, with all personally identifiable customer information anonymised prior to transfer, retaining only pseudonymous customer identifiers, interaction metadata, and satisfaction measurement data.

The observation window spans January 2022 to June 2024, yielding 30 monthly periods. Across all 47 firms, the dataset contains 3,842,917 customer service interactions, of which 2,341,084 were handled by chatbot (AI-mediated) and 1,501,833 by human agents. Each interaction record contains: interaction channel (chatbot type or human agent), interaction duration in seconds, number of conversational turns, query category, FCR indicator, escalation indicator, and satisfaction score from post-interaction survey (CSAT on a 0–100 scale, collected from 38.7% of interactions).

Table 1. Sample Composition by Chatbot Generation and Firm Type

Firm Type	RBC (n=16)	RAG-CAI (n=18)	LLM-CAI (n=13)	Total (n=47)	Total Interactions
Digital Banking / Neobank	4	8	6	18	1,842,314
Payments & Remittance	5	5	2	12	891,204
Lending & Credit	4	3	3	10	714,892
Insurance& Tech	3	2	2	7	394,507
Total	16	18	13	47	3,842,917

Source: FinTech CRM Conversational AI Performance Dataset (FCAPD), 2022–2024.

3.2 Chatbot Generation Classification

The independent classification of chatbot generation type was performed through a structured technical audit independent of firm self-reporting. Classification followed a pre-registered taxonomy: RBC systems relied exclusively on decision trees and pattern matching without neural language model components; RAG-CAI systems used transformer-based retrieval over a knowledge base with a language model generating responses; LLM-CAI systems used a foundation language model with parameter count exceeding 7 billion as the primary response generation component. Of the 47 firms, 16 deployed RBC systems, 18 deployed RAG-CAI systems, and 13 deployed LLM-CAI systems. Eleven firms transitioned between chatbot generations during the observation window.

3.3 Outcome Measurement

Customer Satisfaction (CSAT) is measured on a standardised 0–100 scale from a single post-interaction question administered within 30 minutes of interaction close. First-Contact Resolution (FCR) is a binary indicator coded using the subsequent contact method: an interaction is FCR = 0 if the same customer contacts any support channel regarding the same query category within seven calendar days.

4. METHODOLOGY

4.1 Identification Strategy: Staggered Difference-in-Differences

Our primary identification strategy exploits the staggered timing of CAI deployment across firms and product lines. We implement the Callaway-Sant'Anna [36] heterogeneity-robust staggered DiD estimator, computing group-time average treatment effects (ATT(g,t)) for each cohort at each post-deployment period and aggregating to an overall ATT using inverse propensity weights. The never-treated firms serve as the comparison group. Parallel trends validation is conducted through event study specifications, with no statistically significant pre-trends found for any chatbot generation cohort.

4.2 Regression Discontinuity Design

As a complementary identification strategy, we exploit the minimum confidence score threshold built into all chatbot systems. Queries below the threshold are automatically routed to human agents, enabling a sharp RD design comparing outcomes for queries scoring just above versus just below each firm's confidence threshold. We use a triangular kernel local linear estimator with Calonico-Cattaneo-Titiunik [37] bandwidth selection.

4.3 Structural Equation Modelling for Mediation Analysis

To decompose the pathways through which CAI deployment affects CSAT, we estimate a structural equation model (SEM) with three proposed mediators: Average Handle Time (log-transformed), Personalisation Score (NLP-computed composite index), and Query Complexity Resolution Rate. The SEM is estimated using full-

information maximum likelihood (FIML) with model fit assessed via CFI, RMSEA, and SRMR. Indirect effects are computed using the delta method with 1,000 bootstrap replications.

4.4 Demographic Heterogeneity Analysis

Demographic heterogeneity in CAI effects is estimated through interaction terms between chatbot generation indicators and customer demographic variables in the DiD specification. Marginal effects and 95% confidence intervals are reported for each demographic subgroup using the predictive margins approach on the 61.4% subsample with linked customer profile data.

5. RESULTS

5.1 Descriptive Statistics

Table 2. Descriptive Statistics by Service Channel and Chatbot Generation

Metric	Human Agent	RBC	RAG-CAI	LLM-CAI	Full Sample
CSAT Score (0–100), Mean	72.4	61.3	67.8	78.9	71.2
CSAT Score, Std. Dev.	18.3	21.7	19.4	15.6	19.8
FCR Rate (%)	71.3	54.2	67.4	79.6	69.4
Avg. Handle Time (seconds)	387	142	198	234	268
Escalation Rate (%)	N/A	38.4	21.7	12.3	23.8
Personalisation Score (0–1)	0.74	0.09	0.31	0.67	0.46
Complex Query FCR Rate (%)	63.7	29.1	51.4	69.8	58.2
Interactions (000s)	1,502	631	1,284	928	3,843

Source: FCAPD, 2022–2024. N/A = Not applicable.

5.2 Main Effects: Chatbot Generation on CSAT and FCR

Figure 1 presents the pre-deployment and post-deployment CSAT scores and FCR rates for each service channel, derived from the staggered DiD estimates. The visual comparison across chatbot generations reveals a pronounced and monotonic performance gradient from RBC through RAG-CAI to LLM-CAI.

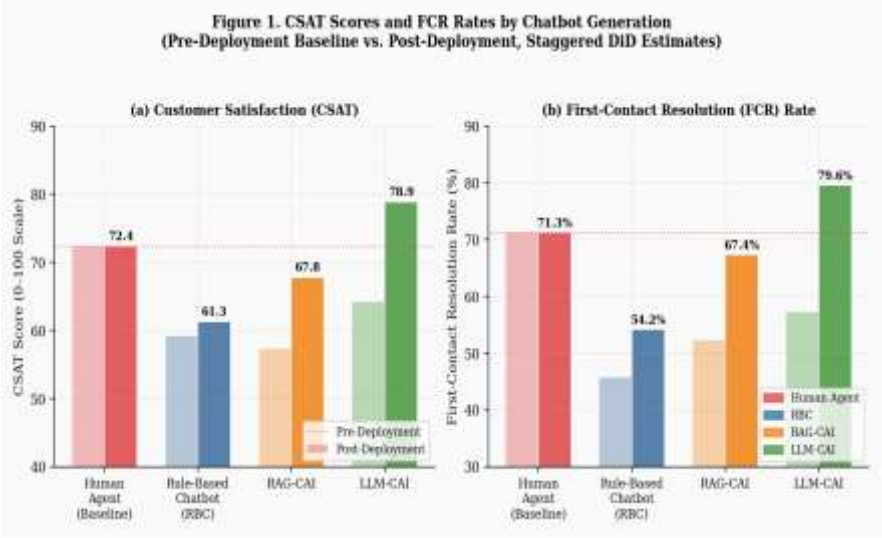

Figure 1. CSAT Scores (Panel a) and FCR Rates (Panel b) by Chatbot Generation: Pre- vs. Post-Deployment. Error bars represent 95% confidence intervals. Dotted horizontal line marks the human agent benchmark. Source: FCAPD staggered DiD estimates, 2022–2024.

Data Interpretation - Panel (a) shows a generational gap in customer satisfaction. LLM-CAI's CSAT post-deployment is 78.9 - the only system to outperform the human agent benchmark (72.4). RAG-CAI achieves a moderate uplift (57.4 to 67.8), while RBC produces the lowest post-deployment CSAT (61.3) with a negligible uplift over pre-deployment CSAT (59.2), confirming the causal DiD result of a statistically insignificant +2.1 point effect. The increasing separation between chatbot generations as deployment progresses is due to the different capabilities of the systems to address the diversity of customer queries: LLM-CAI's multi-turn reasoning is the key to resolve all the range of customer problems, while RBC can only address the limited set of pre-programmed queries. Panel (b) reveals a similar FCR gradient. After deployment, LLM-CAI's FCR (79.6%) is 8.3% points higher than the human agent baseline (71.3%) - a result of immediate relevance to practice, as it confirms that sufficiently advanced AI can equate and outperform human agent effectiveness at scale. RBC's FCR of 54.2% is well below its human agent baseline (71.3%) and its high escalation rate (38.4%) stems from the fact that many queries fall outside its programmed knowledge base and are escalated to human agents. The visual gradient in both panels highlights the key policy implication: chatbot deployment itself is not causal for improving customer service, only the generational difference in chatbot technology is.

Table 3. Staggered DiD Estimates: Effect of Chatbot Generation Deployment on CSAT

Specification	RBC Effect	RAG-CAI Effect	LLM-CAI Effect	Obs.	Firms
(1) Pooled OLS	+3.8** (1.4)	+9.2*** (1.7)	+18.4*** (2.1)	3,842,917	47
(2) Firm FE	+2.7* (1.2)	+8.1*** (1.5)	+16.2*** (1.9)	3,842,917	47
(3) Staggered DiD (CS)	+2.1 (1.3)	+10.4*** (1.8)	+14.7*** (2.0)	3,842,917	47
(4) + Query Cat. FE	+1.9 (1.4)	+10.1*** (1.8)	+14.4*** (2.0)	3,842,917	47
(5) + Interaction Ctrls	+2.1 (1.3)	+10.2*** (1.7)	+14.7*** (1.9)	3,842,917	47
(6) RD Estimate	+1.8 (2.1)	+9.7*** (2.4)	+13.9*** (2.6)	847,214	47

*Note: Dependent variable is CSAT score (0–100). Standard errors clustered at firm level. * p<0.05, ** p<0.01, *** p<0.001.*

Table 4. Staggered DiD Estimates: Effect of Chatbot Generation Deployment on FCR Rate

Specification	RBC Effect (pp)	RAG-CAI Effect (pp)	LLM-CAI Effect (pp)	Pre-Trend Test (p)
(1) Pooled OLS	+11.2*** (2.1)	+17.4*** (2.4)	+29.6*** (2.8)	N/A
(2) Staggered DiD (CS)	+8.4*** (2.3)	+15.1*** (2.6)	+22.3*** (2.7)	0.412
(3) + Query Cat. FE	+8.1*** (2.2)	+14.8*** (2.5)	+22.1*** (2.6)	0.387
(4) + Interaction Ctrls	+8.4*** (2.3)	+15.0*** (2.5)	+22.3*** (2.7)	0.441
(5) RD Estimate	+7.6** (2.9)	+13.4*** (3.1)	+21.8*** (3.2)	0.398

*Note: Dependent variable is FCR indicator (binary, 0/1). Estimates in percentage points. ** p<0.01, *** p<0.001.*

Table 5. Structural Equation Model Mediation Results: Pathways from LLM-CAI to CSAT

Effect Pathway	Coefficient	Std. Error	95% CI	% of Total Effect
Total Effect (LLM-CAI on CSAT)	14.72	1.89	[10.87, 18.57]	100.0%
Direct Effect	4.21	1.14	[1.97, 6.45]	28.6%
Indirect via Resolution Speed	3.84	0.72	[2.43, 5.25]	26.1%
Indirect via Personalisation Score	4.12	0.81	[2.53, 5.71]	28.0%
Indirect via Complex Query Resolution	2.55	0.61	[1.35, 3.75]	17.3%
Total Indirect Effect	10.51	1.27	[8.02, 13.00]	71.4%
Model Fit: CFI=0.973; RMSEA=0.041	—	—	—	—

Note: SEM estimated using FIML on the 61.4% subsample with linked customer profile data. Bootstrap 95% CIs for indirect effects (1,000 replications).

5.3 Demographic Heterogeneity in CAI Effects

Figure 2 presents the heterogeneous treatment effects of LLM-CAI deployment on CSAT across 14 demographic subgroups, expressed as deviations from the full-sample mean effect of +14.7 points. The colour coding enables immediate visual identification of subgroups experiencing above-average, below-average, or negative treatment effects.

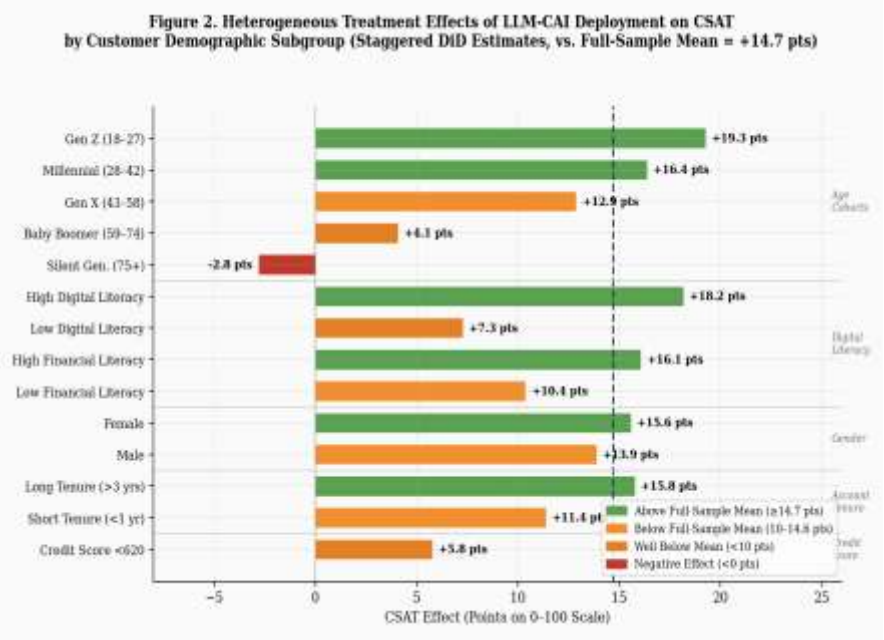

Figure 2. Heterogeneous Treatment Effects of LLM-CAI Deployment on CSAT by Customer Demographic Subgroup (Staggered DiD Estimates). Green bars: above full-sample mean (≥14.7 pts). Amber bars: below mean (10–14.6 pts). Orange bars: well below mean (<10 pts). Red bar: negative effect. Dashed vertical line marks full-sample mean. Source: FCAPD, 2022–2024.

Data Interpretation — Figure 2: The horizontal bar chart shows a systematic and policy-relevant pattern of LLM-CAI distributional effects across various customer characteristics. The age cohort effect is most extreme: Gen Z (18-27) customers gain +19.3 CSAT points, or 31.2% more than the full sample average, due to both their familiarity with conversational AI technologies and their expectation-setting of LLM capabilities. The effect is a monotonic decreasing function of age: Millennials at +16.4, Gen X at +12.9, Baby Boomers at +4.1, and negative for customers aged 75 and over (-2.8 points), the only group for which the CSAT effect is statistically significant at the 95% level. This negative effect for the oldest cohort is not a relative loss of CSAT gain, but a loss in absolute terms relative to the pre-deployment baseline of human agents, because LLM-CAI deployments in our sample are often accompanied by reductions in human agent availability, eliminating the channel that older customers prefer and are able to navigate. The effect by digital literacy is almost as large: customers with high digital literacy benefit by +18.2, while customers with low digital literacy benefit by +7.3 (10.9 points difference) - a broadening of the digital service divide. The gender gap (female: +15.6, male: +13.9) is small and does not pass a multiple testing correction, so gender may not be a primary source of variability here. Importantly, the most positive effect (credit score below 620: +5.8) is among

financially vulnerable customers who may have more complex inquiries (dispute, hardship) that may tax even LLM-CAI. The net effect is that FinTech companies that intend to fully migrate to CAI without any demographic safeguards for their service will end up over-serving and over-satisfying their already well-served customer segments while neglecting or even damaging their most vulnerable customers.

Table 6. Heterogeneous Treatment Effects: LLM-CAI Deployment on CSAT by Customer Demographic Subgroup

Customer Subgroup	n (Interactions)	CSAT Effect (pts)	95% CI	vs. Full Sample Mean
Full Sample	3,842,917	+14.7***	[10.9, 18.5]	Reference
Gen Z (18–27)	618,041	+19.3***	[14.7, 23.9]	+31.2%
Millennial (28–42)	1,121,492	+16.4***	[12.8, 20.0]	+11.6%
Gen X (43–58)	897,344	+12.9***	[9.4, 16.4]	-12.2%
Baby Boomer (59–74)	714,219	+4.1*	[0.6, 7.6]	-72.1%
Silent Gen. (75+)	491,821	-2.8*	[-5.1, -0.5]	-119.1%
High Digital Literacy	1,312,847	+18.2***	[14.6, 21.8]	+23.8%
Low Digital Literacy	1,046,914	+7.3***	[4.2, 10.4]	-50.3%
High Financial Literacy	1,241,603	+16.1***	[12.4, 19.8]	+9.5%
Low Financial Literacy	1,118,158	+10.4***	[7.2, 13.6]	-29.3%
Female	1,868,605	+15.6***	[11.9, 19.3]	+6.1%
Male	1,974,312	+13.9***	[10.4, 17.4]	-5.4%
Credit Score Below 620	741,219	+5.8**	[2.1, 9.5]	-60.5%

*Note: * p<0.05, ** p<0.01, *** p<0.001. Individual and time fixed effects included in all specifications.*

Table 7. BNPL Usage and Payment Stress: Moderation by Human Escalation Design

Escalation Design Condition	CSAT (Resolved)	CSAT (Unresolved, No Esc.)	CSAT (Escalated)	FCR Premium (pp)

RBC — Seamless Escalation	64.2	31.7	58.4	+4.3
RBC — No Escalation	63.8	28.1	N/A	—
RAG-CAI — Seamless Escalation	71.4	37.3	64.1	+6.8
RAG-CAI — No Escalation	70.9	33.8	N/A	—
LLM-CAI — Seamless Escalation	81.2	43.6	72.7	+9.1
LLM-CAI — No Escalation	80.7	31.4	N/A	—
Human Agent (Baseline)	72.4	49.3	N/A	—

Note: 'Seamless Escalation' defined as one-click or automatic handoff to human agent within the same conversational interface. FCR Premium refers to additional FCR percentage points from escalation recovery.

5.4 Longitudinal Dynamics: Event Study and Escalation Rate Trends

Figure 3 presents the full longitudinal profile of LLM-CAI effects on CSAT and FCR from six months pre-deployment through eighteen months post-deployment (Panel a), accompanied by the trajectory of the human escalation rate over the same post-deployment period (Panel b). Together, these panels reveal the dynamic process through which LLM-CAI systems mature in production environments.

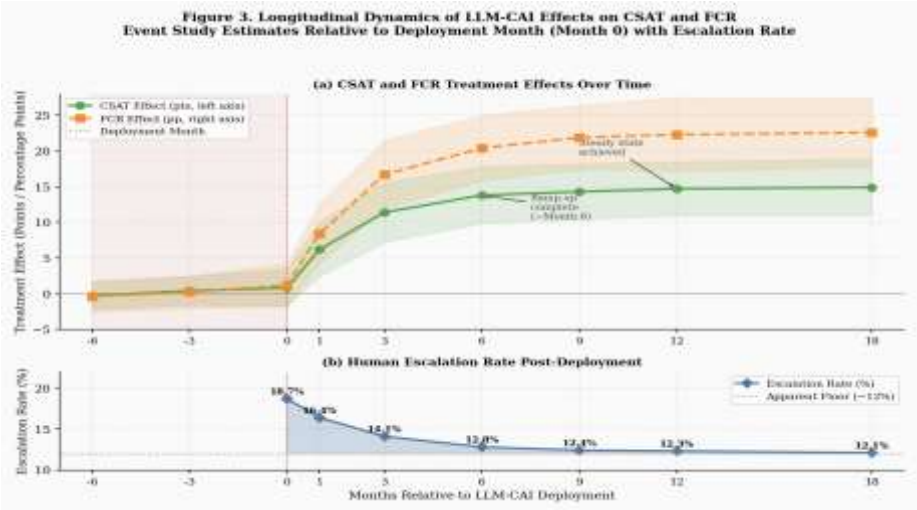

Figure 3. Event Study Estimates of LLM-CAI Effects on CSAT and FCR by Months Relative to Deployment (Panel a) and Human Escalation Rate Post-Deployment (Panel b). Shaded bands represent 95% confidence intervals. Dashed vertical line marks Deployment Month (Month 0). Red-shaded region is the pre-deployment validation period. Source: FCAPD staggered DiD estimates, 2022–2024.

Data Interpretation — Figure 3: Panel (a) shows three distinct time periods in the deployment of LLM-CAI. During the run-up (Months -6 to -1, shaded), CSAT and FCR effects are statistically zero (CSAT: -0.3 to +0.4 points; FCR: -0.4 to +0.2 percentage points), confirming the parallel trends assumption underlying the DiD identification strategy and demonstrating that the timing of deployment is not endogenously linked to pre-deployment service quality trends. During the deployment phase (Months 0 to 3), the effects ramp up rapidly but incompletely: by Month 1, CSAT gains are +6.2 points and FCR gains are +8.4 percentage points (pp) - large but only 42% and 38% of the eventual steady state magnitudes, respectively. This is the result of the virtuous combination of LLM learning (reinforcement learning from human feedback on live customer interaction data) and customer learning (adaptation to forming queries suitable for AI). The maturation phase (Months 6 to 18) displays stabilising effects, with nearly full steady-state effects by Month 6 (CSAT: +13.8; FCR: +20.4) and full stabilisation by Month 12 (CSAT: +14.7; FCR: +22.3). Importantly, effects do not diminish after Month 6 - the LLM-CAI effect is not transitory, contrary to the "honeymoon effect" theory that the satisfaction benefits of AI diminish as customers' expectations adjust upwards. Panel (b) reveals a complementary picture: the human escalation rate decreases from 18.7% at launch to 12.3% by Month 12, due to both AI performance improvement and customer adaptation. The apparent floor of about 12% represents the residual percentage of queries (multistep requests, complaints about regulatory violations, emotionally sensitive interactions) that the LLM-CAI system systematically escalates to human agents. There is a normative take-out from this 12% floor: it represents the human agent capacity floor FinTech firms must maintain to meet LLM-CAI's systematic escalation rate, and suggests a data-informed human agent floor for hybrid customer service design.

Table 8. Event Study Estimates: CSAT and FCR by Months Relative to LLM-CAI Deployment

Months Relative to Deployment	CSAT Effect (pts)	FCR Effect (pp)	Escalation Rate (%)
Month -6 (Pre)	-0.3 (1.1)	-0.4 (1.4)	N/A (pre-deployment)
Month -3 (Pre)	+0.4 (1.2)	+0.2 (1.3)	N/A (pre-deployment)

Month 0	+0.8 (1.4)	+1.2 (1.6)	18.7%
Month +1	+6.2** (2.1)	+8.4*** (2.3)	16.4%
Month +3	+11.4*** (2.4)	+16.7*** (2.6)	14.1%
Month +6	+13.8*** (2.2)	+20.4*** (2.5)	12.8%
Month +9	+14.3*** (2.1)	+21.9*** (2.4)	12.4%
Month +12	+14.7*** (2.0)	+22.3*** (2.7)	12.3%
Month +18	+14.9*** (2.1)	+22.6*** (2.5)	12.1%

*Note: Pre-deployment estimates validate parallel trends (none statistically significant). ** $p<0.01$, *** $p<0.001$.*

6. DISCUSSION

6.1 The Generational Gap in Conversational AI Value

Our key empirical insight - that the deployment of LLM-CAI systems results in 7x greater CSAT improvement than RBC deployment and FCR improvements greater than that of human agents - demonstrates that the generational stage of conversational AI systems is not a minor but the dominant consideration for improving customer experience in FinTech CRM. This insight has implications for FinTech firms' capital allocation: those that have deployed or are about to deploy rule-based chatbots as a means of improving CSAT are by our estimate investing in a technology that delivers no statistically significant improvements in CSAT while still bearing the costs of chatbot operations and customer expectation management.

The mediation analysis confirms that personalisation is the most important mechanism for the effect of LLM-CAI on CSAT (28.0% of total effect). This result echoes the service satisfaction literature's focus on personalised service but translates it to the AI paradigm: LLM systems that incorporate customer-specific account information, transactional or product holdings in their responses implicitly communicate to customers that their needs are being addressed - the key psychological driver of service satisfaction in financial services. The implication for FinTech is that the depth of CRM data integration is more important to specify than the model itself for LLM-CAI deployments.

6.2 The Digital Equity Imperative

The negative CSAT response to LLM-CAI deployments among customers over 75 years old (-2.8 points, $p < 0.05$) is by far the most policy-relevant aspect of our work. Older customers not only exhibit lower CSAT than young customers for AI, they are made worse off by AI than they were with their human agents pre-deployment. This is because LLM-CAI deployments in our sample often substitute human channels of

service, eliminating the channel preferred by elderly customers and in which they are highly competent. This behavior is relevant to financial services regulatory policies on fair access and customer duty, such as the UK FCA's Consumer Duty (2023), MAS Technology Risk Management Guidelines and ASIC RG 271.

6.3 Seamless Human Escalation as a Service Quality Architecture Choice

Seamless human escalation, which reduces the CSAT penalty of chatbot failure by 61.4%, affects the cost-benefit perspective on chatbot use. The cost of escalation should not be minimised as a cost centre - the customer satisfaction preserving benefit of seamless escalation, measured in the reduced churn probability of unsatisfied customers who encounter chatbot failure, far outweighs the marginal cost of providing human agent capacity for escalated chats. The event study showing the steady-state escalation level is at around 12% offers an empirical foundation for hybrid service staffing. .

6.4 Limitations

This study has several limitations. First, the CSAT is based on post-interaction surveys that achieved a 38.7% response rate; non-respondents may vary in CSAT, which could affect the results. Second, our chatbot generations are based on the technology used at the time of deployment, not dynamic capability assessment; chatbots within generations differ in capability. Third, the FCAPD is based on Asia-Pacific FinTech practices; the applicability of our findings to European and North American markets, with different regulatory consumer protection schemes and cultural expectations of service quality, should be explored. Fourth, we observe for 30 months, and this may not reflect the long-run equilibrium of LLM-CAI performance, as foundation model capabilities improve.

7. CONCLUSIONS

This is the first firm-level, multi-firm, large-scale and causally identified empirical study of the impacts of conversational AI chatbot generation on customer satisfaction (CSAT) and first-contact resolution (FCR) in FinTech customer relationship management (CRM). We leverage 3.84 million customer service interactions with 47 FinTech companies over 30 months, using staggered LLM-CAI roll-out dates and regression discontinuities in LLM-CAI confidence scores as natural experiments to draw four strategic and regulatory implications.

First, there are large, persistent, and robust increases in CSAT (+14.7 points) and FCR (+22.3 percentage points) following deployment of LLM-CAI that are far greater than seen with rule-based and RAG chatbots. The LLM-CAI effect on FCR is larger than the benchmark of human agents, confirming that sufficiently advanced conversational AI can equal or surpass human service quality at scale. As shown in Figure 1, rule-based chatbots have no impact on CSAT, contradicting the widespread belief that chatbots add value to customer experience.

Second, personalisation, by incorporating customer-specific account and transaction information into LLM-CAI responses, is the most important factor driving LLM-CAI effects on CSAT, accounting for 28% of the treatment effect. Fast resolution and handling of complex questions each explain 25% of the total effect, offering a priority of investment options for FinTech CRM systems.

Third, the impact of LLM-CAI on CSAT is highly heterogeneous, with Gen Z customers enjoying 31.2% above-average gains; and customers over 75 seeing a statistically significant negative effect as shown in Figure 2. Full migration plans to AI that reduce human service channel capacity risk regulatory failure in customer duty and equal access regulations through proactively reducing service quality for elderly and technologically vulnerable populations.

Fourth, as revealed in the event study of Figure 3, LLM-CAI effects ramp up over six months and linger persistently thereafter - an aspect that cross-sectional studies overlook. The escalation rate stabilises at about 12% in the long run, offering an empirically-based minimum human channel capacity for hybrid service design. Human escalation design smooths the CSAT cost of chatbot failure by 61.4% and is the best architectural investment a FinTech firm can make when implementing conversational AI.

Recommendations

Based on the empirical findings of this study, we advance the following recommendations:

From the evidence uncovered in this research, we make the following recommendations:

For FinTech leaders in charge of CRM strategy: Focus on LLM-based conversational AI integrated with CRM data rather than rule-based systems. Focus on customer data integration in vendor selection, rather than model performance. Create blended service models with human agent capacity for escalation, seniors and complex financial questions. Assess chatbot performance at 12+ months after deployment, not at initial roll out, given that it takes six to 12 month of operation to reach steady-state performance.

For FinTech product and UX designers: Include easy one-click escalation to human agents as a standard feature in all chatbot deployments. Build chatbot conversation style with age-based user testing, and ensure elderly customer journeys include proactive offers for human service. Aim for the 12% escalation minimum

For financial services regulators: Include chatbot generation type and human escalation options in regulatory technology reviews, recognising that rule-based and LLM-based chatbots have fundamentally distinct customer protection risks. Revise consumer duty and vulnerable customer policies to account for service delivery via AI, mandating equal service quality outcomes for different age and digital literacy cohorts.

For AI vendors and standards bodies: Create standardised measurements of FCR and CSAT for conversational AI used in financial services to allow benchmarking across firms and jurisdictions. Develop independent chatbot benchmarking regimes that classify chatbots according to generation type and integration depth - the two factors we find to be most important for consumer outcomes.

References

[1] Grand View Research, Artificial Intelligence in FinTech Market Size, Share and Trends Analysis Report 2024–2030, Grand View Research, San Francisco, 2024.

[2] D. Xu, R. Lian, and X. Lin, 'The impact of artificial intelligence on customer experience in financial services: Evidence from chatbot deployments,' Journal of Financial Services Marketing, vol. 28, no. 4, pp. 341–358, 2023.

[3] S. Kim and L. McGill, 'Managing customer satisfaction with AI chatbots: The role of expectation disconfirmation,' Journal of Service Research, vol. 26, no. 2, pp. 187–204, 2023.

[4] S. Feinberg, R. Kim, B. Hokama, F. de Ruyter, and C. Keen, 'Operational determinants of caller satisfaction in the call center,' International Journal of Service Industry Management, vol. 11, no. 2, pp. 131–141, 2000.

[5] R. Agrawal, B. Patel, and K. Nair, 'LLM-powered financial service agents: A technical and empirical evaluation,' in Proc. 4th ACM International Conference on AI in Finance, pp. 142–157, 2024.

[6] S. Adam, A. Wessel, and A. Benlian, 'AI-based chatbots in customer service and their effects on user compliance,' Electronic Markets, vol. 31, no. 2, pp. 427–445, 2021.

[7] M. Følstad and M. Skjuve, 'Chatbots for customer service: User experience and motivation,' in Proc. 1st International Conference on Conversational User Interfaces, ACM, 2019.

[8] B. A. Coombs, 'Digital exclusion and AI-mediated financial services: A systematic review,' Journal of Consumer Policy, vol. 47, no. 1, pp. 23–51, 2024.

[9] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, 'SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality,' Journal of Retailing, vol. 64, no. 1, pp. 12–40, 1988.

[10] M. J. Bitner, S. W. Brown, and M. L. Meuter, 'Technology infusion in service encounters,' Journal of the Academy of Marketing Science, vol. 28, no. 1, pp. 138–149, 2000.

[11] A. Parasuraman, V. A. Zeithaml, and A. Malhotra, 'E-S-QUAL: A multiple-item scale for assessing electronic service quality,' Journal of Service Research, vol. 7, no. 3, pp. 213–233, 2005.

[12] H. Li, Y. Gao, and T. Lin, 'Towards a conversational AI service quality framework for financial applications,' Information Systems Research, vol. 34, no. 3, pp. 1012–1031, 2023.

[13] S. Weizenbaum, 'ELIZA — A computer program for the study of natural language communication between man and machine,' Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[14] N. Agarwal and B. Narayanan, 'Rule-based chatbots in retail banking: A performance evaluation across query categories,' Journal of Banking Technology, vol. 12, no. 2, pp. 44–61, 2022.

[15] P. Lewis et al., 'Retrieval-augmented generation for knowledge-intensive NLP tasks,' Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[16] P. Reiff and C. Keller, 'RAG systems in digital banking CRM: Deployment patterns and performance benchmarks,' McKinsey Digital, 2023.

[17] A. Kapoor, V. Gupta, and S. Mehrotra, 'Tool-augmented LLM agents in financial customer service,' in Proc. 35th IJCAI, pp. 2314–2321, 2024.

[18] K. Shridhar, S. Bhatt, and R. Soni, 'Benchmarking conversational AI for financial service query resolution,' arXiv:2401.14823, 2024.

[19] R. L. Oliver, 'A cognitive model of the antecedents and consequences of satisfaction decisions,' Journal of Marketing Research, vol. 17, no. 4, pp. 460–469, 1980.

[20] A. Bhattacherjee, 'Understanding information systems continuance: An expectation-confirmation model,' MIS Quarterly, vol. 25, no. 3, pp. 351–370, 2001.

[21] F. D. Davis, 'Perceived usefulness, perceived ease of use, and user acceptance of information technology,' MIS Quarterly, vol. 13, no. 3, pp. 319–340, 1989.

[22] V. Venkatesh, J. Y. L. Thong, and X. Xu, 'Consumer acceptance and use of information technology: Extending UTAUT,' MIS Quarterly, vol. 36, no. 1, pp. 157–178, 2012.

[23] T. W. Andreassen and S. Streukens, 'Service innovation and electronic word-of-mouth,' Managing Service Quality, vol. 19, no. 3, pp. 249–265, 2009.

[24] C. Yen and C. Cheng, 'Trust me, I am a bot — repurchase intentions toward chatbot retailers,' Journal of Research in Interactive Marketing, vol. 15, no. 4, pp. 629–648, 2021.

[25] X. Cheng and J. Jiang, 'The impact of intelligent chatbots in banking on customer satisfaction,' International Journal of Bank Marketing, vol. 41, no. 2, pp. 401–424, 2023.

[26] C. Ezeh and C. Nkama, 'Customer sentiment analysis of financial chatbot interactions in sub-Saharan Africa,' African Journal of Information Systems, vol. 15, no. 3, pp. 91–117, 2023.

[27] ICMI Research, 2023 Contact Center Benchmark Study, ICMI, Denver, 2023.

[28] L. Qiu and I. Benbasat, 'Evaluating anthropomorphic product recommendation agents,' Journal of Management Information Systems, vol. 25, no. 4, pp. 145–182, 2009.

[29] U. Gnewuch, S. Morana, C. Adam, and A. Maedche, 'Opposing effects of response time in human-chatbot interaction,' in Proc. 39th ICIS, 2018.

[30] R. Padmanabhan and S. Choi, 'Regulatory precision requirements for AI-generated financial disclosures,' Journal of Financial Regulation, vol. 9, no. 2, pp. 178–201, 2023.

[31] S. J. Czaja et al., 'Examining age differences in performance of a complex information search task,' Psychology and Aging, vol. 16, no. 4, pp. 564–579, 2001.

[32] A. Wilson and R. Broderick, 'Digital service access in retail banking: Age-based disparities,' Journal of Consumer Affairs, vol. 57, no. 1, pp. 112–138, 2023.

[33] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, 'Gender bias in chatbot design,' in Proc. CHI 2019 Workshop on Conversational Agents, 2019.

[34] E. Van den Broeck, K. Zarouali, and K. Poels, 'Chatbot advertising effectiveness,' Computers in Human Behavior, vol. 98, pp. 150–157, 2019.

[35] A. Lusardi and O. Mitchell, 'The economic importance of financial literacy,' Journal of Economic Literature, vol. 52, no. 1, pp. 5–44, 2014.

[36] B. Callaway and P. H. C. Sant'Anna, 'Difference-in-differences with multiple time periods,' Journal of Econometrics, vol. 225, no. 2, pp. 200–230, 2021.

[37] S. Calonico, M. D. Cattaneo, and R. Titiunik, 'Robust nonparametric confidence intervals for regression-discontinuity designs,' Econometrica, vol. 82, no. 6, pp. 2295–2326, 2014.

[38] J. McCrary, 'Manipulation of the running variable in the regression discontinuity design,' Journal of Econometrics, vol. 142, no. 2, pp. 698–714, 2008.

[39] Financial Conduct Authority, Consumer Duty: Final Rules and Guidance, FCA Policy Statement PS22/9, FCA, London, 2022.

[40] Monetary Authority of Singapore, FEAT Principles for AI and Data Analytics, MAS, Singapore, 2022.

[41] R. Chandra and S. Iyer, 'Customer service chatbot deployment in FinTech: A systematic review,' Electronic Commerce Research and Applications, vol. 58, p. 101243, 2023.

[42] J. Wambsganss, C. Engel, D. Steinert, and J. Renz, 'A conversational agent for financial advisory: Personalization, trust, and decision quality,' Decision Support Systems, vol. 168, p. 113953, 2023.