

UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea triennale in

Statistical Analysis on Generative AI Adoption in the U.S. Banking Sector

A Comparative Benchmark with the EU and Italy (Eurostat, Istat)

Relatore

Prof. : MARIA CRISTINA RECCHIONI

Rapporto Finale di:

Amare, Firezer Tilahun

Student ID: S1120025

Anno Accademico 2025/2026

Abstract

This thesis investigates whether the "Productivity Paradox" documented by Kikuchi (2026) for generative AI (GenAI) adoption in U.S. banking — in which AI-adopting banks appear more productive in cross-section, yet the act of adoption itself imposes a short-run "Implementation Tax" on profitability — can be detected in a smaller, independently constructed sample. Using a panel of six publicly traded U.S. bank holding companies (2021–2025/2026), built from SEC EDGAR filings, Federal Reserve FR Y-9C regulatory data, and SEC XBRL data where the standard regulatory source was unavailable, this thesis tests four hypotheses adapted from Kikuchi's selection-versus-treatment framework: a cross-sectional selection effect (H1), a causal treatment effect around the November 2022 ChatGPT release (H2), moderation by institutional size (H3), and moderation by pre-existing digital maturity, measured from the text of each bank's pre-2022 annual filings (H4). A descriptive comparison to EU and Italian enterprise AI-adoption trends, drawn from Eurostat, situates the U.S. findings within a broader international pattern.

The results do not reproduce Kikuchi's resolution of the paradox, and the reasons are a central finding in their own right. The selection effect (H1) was directionally consistent with Kikuchi's but not statistically distinguishable from zero. The treatment effect (H2) could not be separately identified from H1 at all, since every adopting bank in the sample first disclosed GenAI activity in 2024 or later, leaving no pre-shock adopter for the difference-in-differences design to compare against. The originally specified size-moderation test (H3) could not be estimated because the sample's sole small-tier bank never adopted GenAI; a substituted large-tier comparison showed a significant effect in the opposite direction from Kikuchi's population-level finding. A newly constructed digital-maturity proxy (H4) showed a significant negative interaction with adoption, suggesting a catch-up rather than complementarity pattern, though this result rests on only four adopting banks. Throughout, the thesis documents specific, traceable causes for each limitation — including a discontinued filer, a Federal Reserve reporting-threshold distinction, and a structural gap in Eurostat's sector coverage — rather than treating them as unexplained noise. The thesis concludes that a six-bank convenience sample lacks the identifying variation needed to

adjudicate between selection and treatment explanations that a much larger panel can separate, and outlines the specific data-collection steps a follow-up study would need to take to do so.

Acknowledgement

I would like to thank my thesis advisor, Professor Maria Cristina Recchioni, for her guidance and support throughout this project. Her willingness to support a transparent analysis of the data's structural limitations – and to push for honest documentation rather than a cleaner but less truthful story – shaped this thesis into a more careful piece of work.

Table of Contents

Abstract ii
Acknowledgement iii
List of tables v
Chapter 1: Introduction and Literature Review 1
1.1 Background and Motivation 1
1.2 Research Problem, Questions, and Hypotheses 2
1.3 Literature Review 3
1.4 Significance of the Study 8
Chapter 2: Data and Methodology 9
2.1 Conceptual Framework: Selection versus Treatment 9
2.2 U.S. Banking Sector Data 10
2.3 Comparative EU/Italy Benchmark Data 13
2.4 Variable Construction and Summary Statistics 14
2.5 Panel Regression, Difference-in-Differences, and Moderation Specifications 17
Chapter 3: Results and Discussion 20
3.1 Descriptive Trends in GenAI Adoption 20
3.2 Panel Regression Results 21
3.3 Difference-in-Differences Results (Implementation Tax) 22
3.4 Moderation Results 23
3.5 EU/Italy Comparative Benchmark 25
3.6 Robustness Checks 26
3.7 Interpretation: Reconciling Selection and Treatment Effects 26
Chapter 4: Conclusion 28
4.1 Summary of Findings 28
4.2 Implications for Bank Management and Policy 29
4.3 Limitations and Future Research 30
Reference 33

List of tables

Table 2.1: Sample bank holding companies 12
Table 2.2: Variable construction 14
Table 2.3: Summary statistics 15
Table 2.4: Digital-maturity proxy (DigMat), pre-2022 10-K filings 16
Table 3.1: GenAI mentions and adoption timing by bank 20
Table 3.2: Return on assets and return on equity by group 21
Table 3.3: Panel fixed-effects results (H1) 21
Table 3.4: Difference-in-differences results (H2) 22
Table 3.5: Size-moderation results (H3, Large-tier interaction substituted for Small-tier) 23
Table 3.6: Digital-maturity moderation results (H4, GenAI × DigMat interaction) 24
Table 3.7: AI adoption rate, EU-27 average vs. Italy 25

Chapter 1: Introduction and Literature Review

1.1 Background and Motivation

Large language models have moved from research labs to knowledge-intensive sectors since the public release of ChatGPT in November 2022. This change has also reached banking and finance. Banks depend on information processing and interpretation, so they become an important setting for the adoption of generative AI for tasks such as credit underwriting, customer service and risk modelling.

Yet the historical relationship between new general-purpose technologies and measured productivity has rarely been immediate. Robert Solow's (1987) observation that computers were visible 'everywhere but in the productivity statistics' described a problem that later appeared with other technologies, including the internet, enterprise software and cloud computing: firms invest heavily in a new technology well before that investment shows up as higher output per worker or higher profitability. Brynjolfsson, Rock and Syverson (2021) formalize this as a "productivity J-curve," in which the intangible investments a firm must make to use a general-purpose technology effectively – retraining staff, redesigning processes, building complementary data infrastructure – depress measured output in the short run, even though they raise the technology's long-run productive potential.

Kikuchi (2026) is, to date, the most direct empirical test of this pattern for generative AI in banking. Using SEC 10-Q filings linked to Federal Reserve regulatory data for 809 U.S. bank holding companies, the study finds that banks mentioning generative AI in their disclosures have higher average performance than non-adopters. However, after estimating the effect of adoption itself, the paper reports a 428-basis-point decline in ROE, which it describes as an 'Implementation Tax'. (the paper's title phrase "Innovation Tax" refers to the overall study; the specific ROE decline is named the "Implementation Tax" throughout its empirical sections). The tax falls more heavily on smaller institutions and coexists with positive productivity spillovers across network-connected banks, raising a further concern about "algorithmic coupling" and systemic risk.

This thesis takes that single-country, single-study finding as its point of departure and asks two things: first, whether an independently constructed panel of U.S. banking data reproduces the same

basic pattern using a more tractable econometric design suited to the scope of a thesis; and second, whether comparable sector-level evidence from the EU and Italy – drawn from Eurostat and Istat, the two databases specified for this research – is broadly consistent with, or diverges from, the U.S. pattern.

The question is relevant beyond banking. GenAI is often discussed as a general-purpose technology with the potential to affect productivity across the economy. Banking provides a useful setting because it is information-intensive and has relatively detailed public regulatory disclosures. If a measurable short-run implementation cost can be identified using public banking data, similar methods could be applied to other information-intensive service sectors such as insurance, professional services and telecommunications.

1.2 Research Problem, Questions, and Hypotheses

1.2.1 Research Problem

Existing evidence does not show a clear immediate productivity gain for banks adopting generative AI. The reasons for this delay, and for who within the sector bears its cost, are not yet well understood beyond a single recent U.S. study. It is also unclear whether the pattern documented for U.S. banks is a feature specific to that market – shaped by U.S. disclosure practices, competitive structure, and the scale of the largest U.S. banks – or a more general feature of how banking sectors absorb a new general-purpose technology, which would be expected to leave a visible trace in European sector statistics as well.

1.2.2 Research Questions

- RQ1: Is generative AI adoption by U.S. bank holding companies associated with a short-run change in productivity and profitability (ROA, ROE)?
- RQ2: Does bank size or digital maturity moderate the relationship between GenAI adoption and bank performance?
- RQ3: How does the scale and pattern of AI adoption in the U.S. banking sector compare with AI adoption and productivity trends in the EU and Italian financial and insurance sector, as measured by Eurostat and Istat?

1.2.3 Hypotheses

“Building on Kikuchi (2026), this thesis tests four related hypotheses using a smaller panel and a simplified empirical design. H1 and H2 follow Kikuchi's distinction between selection and treatment effects. H3 examines whether the Implementation Tax differs by bank size. However, as discussed in Chapter 2, Kikuchi's empirical results support a larger effect for smaller banks, whereas the original H3 hypothesis in that paper predicted a larger effect for large banks because of legacy-system complexity. This thesis therefore follows the empirical finding rather than the original hypothesis wording. H4 is an extension developed for this thesis and examines whether digital maturity moderates the effect of GenAI adoption. It is not a replication of Kikuchi's H4, which concerns network spillovers.

- H1: AI-adopting banks exhibit higher baseline productivity than non-adopters in the cross-section (a selection, or “frontier firm,” effect).
- H2: The causal effect of AI adoption on short-term profitability is negative (the “Implementation Tax”).
- H3: The magnitude of the “Implementation Tax” is moderated by bank size, with smaller institutions experiencing a larger effect relative to their capital base.
- H4: The magnitude of the effect is moderated by digital maturity, with more digitally mature institutions experiencing a smaller and shorter-lived productivity dip.

1.3 Literature Review

1.3.1 The Productivity Paradox and the Innovation J-Curve

The gap between the potential of a new general-purpose technology and its measured effect on productivity was first documented by Solow (1987), whose observation that computers were visible “everywhere but in the productivity statistics” gave the phenomenon its name. As reviewed by Kikuchi (2026), subsequent work grouped candidate explanations for this gap into two broad families. The first family treats the gap as a measurement or adjustment problem: Brynjolfsson (1993) proposed that mismeasured outputs, learning lags, and the simple redistribution of profits

(rather than their creation) could each produce an apparent productivity shortfall without implying the technology itself is unproductive. The second family instead treats the gap as evidence that complementary organizational investment, not measurement error, is the main constraint: David (1990) argued from the historical case of manufacturing electrification that general-purpose technologies require decades of complementary reorganization before their productive potential is realized, and Brynjolfsson and Hitt (2000) later showed that firms pairing IT investment with decentralized decision-making saw substantially larger productivity gains than firms investing in technology alone — evidence for organizational capital as a complement rather than a substitute for technology, a theme carried forward in Bresnahan, Brynjolfsson and Hitt (2002).

The productivity J-curve proposed by Brynjolfsson, Rock and Syverson (2021) provides a formal explanation for this second view: the intangible investments a firm must make to use a general-purpose technology effectively — retraining staff, redesigning processes, building complementary data infrastructure — depress measured output in the short run precisely while they are raising the technology's long-run productive potential, so a temporary decline in measured performance is not by itself evidence against the technology's value.

Firm-level evidence specific to generative AI, as reviewed by Kikuchi (2026), has so far tended to find sizeable near-term productivity gains at the task level: reported improvements include customer-support agent productivity following the introduction of a generative AI assistant (Brynjolfsson, Li and Raymond, 2023) and reduced completion time for professional writing tasks (Noy and Zhang, 2023). Whether such task-level gains aggregate cleanly into firm- or sector-level productivity is a separate and open question: general equilibrium effects, task reallocation, and organizational adjustment costs could each cause firm- or sector-level measures to lag behind task-level evidence (Dell'Acqua, McFowland, Mollick and colleagues, 2023), which is exactly the possibility a J-curve framework is designed to capture. This thesis examines this question at the firm level using a small sample of U.S. bank holding companies.

1.3.2 Banking, Digital Transformation, and AI in Financial Services

Three themes from the banking-technology literature, as reviewed by Kikuchi (2026), are relevant here. The first concerns how information technology has changed the economics of lending and processing: Berger (2003) traces how earlier waves of IT investment reshaped relationship and transaction-based lending, and Petersen and Rajan (2002) show that IT investment expanded the geographic reach of small-business credit, weakening the traditional link between physical proximity and lending relationships. The second theme concerns how the FinTech era intensified competitive pressure on incumbent banks: Buchak, Matvos, Piskorski and Seru (2018) find that FinTech lenders gained market share specifically during periods of regulatory constraint on incumbents, Fuster, Plosser, Schnabl and Vickery (2019) document meaningfully faster mortgage processing at FinTech lenders without a corresponding increase in default rates, and Berg, Burg, Gombović and Puri (2022) review the broader role of alternative data and machine learning in credit allocation.

The third theme concerns why the capacity to adopt new technology is not evenly distributed across institutions of different sizes. As reviewed by Kikuchi (2026), this scale-economies or "digital divide" literature documents that larger banks can spread the fixed costs of new technology across a larger asset base (Berger, Miller, Petersen, Rajan and Stein, 2005), that community banks in particular face structural disadvantages in adopting new technology given limited IT budgets and specialist expertise (Frame, Wall and White, 2014), and that FinTech partnerships may in some cases let smaller banks leapfrog these barriers rather than simply falling further behind (Jagtiani and Lemieux, 2018). Kikuchi (2026) applies this scale-economies literature to generative AI in banking: its central empirical finding that the short-run "Implementation Tax" is roughly four times larger, in return-on-equity terms, for smaller banks than for larger ones is consistent with the general pattern this literature documents, while adding a causal identification strategy (synthetic difference-in-differences around the ChatGPT release) that the earlier digital-divide literature lacks.

1.3.3 Research Gap

Three gaps motivate this thesis. First, Kikuchi (2026) is, at the time of writing, a single study; its findings on the direction and magnitude of the "Implementation Tax" have not yet been independently reproduced with an alternative panel and a more tractable econometric design. Second, essentially all empirical evidence on this question comes from the United States; whether the same pattern is visible in European banking, where disclosure practices, market structure, and the scale of the largest institutions differ substantially from the U.S., is untested. Third The thesis also brings together two related literatures: the productivity-paradox/J-curve literature and the banking digital-divide literature. which can be examined descriptively using the broader sector-level indicators available from Eurostat and Istat.

1.3.4 AI as an Economic Input

A distinct and faster-growing literature treats artificial intelligence directly as an economic input rather than as one more wave of IT investment. As reviewed by Kikuchi (2026), Agrawal, Gans and Goldfarb (2018) conceptualize AI as a "prediction machine" that sharply lowers the cost of prediction, which changes the value of complementary human judgment rather than simply substituting for labor outright, and Acemoglu and Restrepo (2018) develop a task-based framework in which automation's aggregate effect depends on the relative strength of a displacement effect, which reduces the labor share, against a productivity effect, which can raise wages and employment elsewhere in the economy. This distinction matters for how this thesis's own "Implementation Tax" evidence should be read: a short-run profitability decline following GenAI adoption could reflect either channel, or a mix of both, and a bank-level ROA/ROE panel of the kind constructed in Chapter 2 cannot on its own separate them.

Within financial services specifically, Gu, Kelly and Xiu (2020) show that machine-learning methods meaningfully outperform traditional asset-pricing models in predicting returns, evidence that AI as prediction input has material value even in a well-studied, quantitatively sophisticated corner of finance; Bartlett, Morse, Stanton and Wallace (2022) document algorithmic discrimination in mortgage lending, a reminder that efficiency gains from AI adoption are not automatically distributionally neutral. Neither finding is about network structure or systemic risk

— both concern how AI functions as an input within a single institution's decisions — which is why they are grouped here rather than with the network literature that follows.

1.3.5 Financial Networks, Systemic Risk, and Algorithmic Coupling

Another strand of the literature examines how network structure affects the way financial systems absorb or amplify shocks. In the context of traditional interbank credit and liquidity linkages, Acemoglu, Ozdaglar and Tahbaz-Salehi (2015) and Elliott, Golub and Jackson (2014) show that network topology — how integrated and how diversified a system of counterparty relationships is — determines whether a financial system absorbs or amplifies shocks. Neither paper is about technology adoption. Kikuchi (2026) connects this literature to technology adoption through the concept of ‘algorithmic coupling as a new, technology-mediated channel for the same kind of correlated fragility: banks that rely on similar AI model architectures or vendor solutions can become correlated in their AI-driven decisions even absent any direct balance-sheet exposure to one another. This thesis's own U.S. panel, built from a small sample of bank holding companies, is not designed to re-estimate that network channel directly — doing so would require the kind of spatial weight matrix Kikuchi constructs across 809 institutions, which is outside this thesis's scope — but the concept motivates the systemic-risk framing carried through to the policy discussion in Chapter 4.

1.3.6 Managerial Characteristics and Technology Adoption

A smaller but relevant strand of the literature links firm-level technology adoption to observable managerial characteristics. Bertrand and Schoar (2003) show that individual manager “fixed effects” explain a meaningful share of variation in corporate investment and financing policy, including technology-related decisions. This thesis includes CEO age as a control variable in the empirical specification (Section 2.4), following the same choice made by Kikuchi (2026); it does not independently test or claim a specific relationship between executive age and technology-adoption risk-taking, since no source verified for this thesis establishes that relationship. Including the control is a precaution against omitted-variable bias — if manager-specific characteristics

correlated with age also correlate with GenAI adoption timing and with other drivers of bank profitability, omitting the control could bias the estimated relationship between adoption and performance — rather than a test of a specific age-related hypothesis.

1.4 Significance of the Study

This thesis provides an independent statistical test of a recently documented pattern in banking productivity and examines whether the pattern can be observed using a smaller panel and a simplified empirical design. It also places the U.S. evidence alongside broader European and Italian sector-level indicators from Eurostat and Istat. The comparison is descriptive because the European data are not directly matched to the U.S. bank-level observations.

The remainder of the thesis is organized as follows. Chapter 2 develops the conceptual framework distinguishing selection from treatment effects, describes the data sources and variable construction, and specifies the econometric models. Chapter 3 presents the empirical results and discusses the findings in relation to the hypotheses and previous literature. Chapter 4 summarizes the main findings, discusses the study's limitations and implications, and outlines directions for future research.

Chapter 2: Data and Methodology

2.1 Conceptual Framework: Selection versus Treatment

The main issue in this analysis is distinguishing between selection and treatment. These two mechanisms can produce similar observed relationships between GenAI adoption and bank performance. Selection would occur if banks that adopt GenAI differ systematically from banks that do not. For example, adopting banks may be larger, better capitalized, or more technologically mature. institutions that would have outperformed non-adopters regardless of whether GenAI itself contributed anything to performance. A positive cross-sectional correlation between adoption and profitability under this account reflects the characteristics that led a bank to adopt, not a causal effect of adoption.

The treatment effect is different. In this case, GenAI adoption itself may affect a bank's costs and productivity in the short run, and that effect may be either positive or negative. Kikuchi (2026) documents an "Implementation Tax" using a Synthetic Difference-in-Differences design around the November 2022 ChatGPT release: across 809 U.S. bank holding companies observed 2018–2025, adopting banks experience a 428-basis-point average decline in ROE (46 basis points for ROA) as they absorb GenAI integration costs, a pattern the paper interprets through the "Innovation J-Curve" (Brynjolfsson, Rock, and Syverson, 2021) and situates within Solow's (1987) earlier observation that computer-related productivity gains were, for a time, difficult to detect in aggregate output statistics. Kikuchi additionally documents that this tax is not evenly distributed: bottom-three-quartile banks by assets suffer a 517-basis-point ROE decline, compared to 129 basis points for top-quartile banks the opposite direction from what a pure legacy-system-complexity account would predict, and a pattern this thesis's own H3 test is designed to probe in the smaller six-bank sample described in Sec. 2.2.

The analysis does not assume in advance that either mechanism is more important. Section 2.5 uses separate specifications to examine the two possibilities. A static panel specification (H1) captures the selection-consistent cross-sectional association between adoption and performance, while a difference-in-differences specification centered on the ChatGPT shock (H2) isolates within-bank changes in performance that occur specifically after adoption is observed, net of any

time-invariant bank characteristics that a fixed-effects design absorbs. The comparison between these two estimates is discussed further in Sect. 3.7 and is treated as part of the main findings rather than as a robustness check.

2.2 U.S. Banking Sector Data

The U.S. analysis draws on two public regulatory data sources, following the approach in Kikuchi (2026).

GenAI adoption indicator (SEC EDGAR).

Adoption is coded from the full-text search index of SEC EDGAR, which covers the complete text of 10-K annual and 10-Q quarterly filings. For each sample bank, filings are searched for five core GenAI terms: "generative AI," "generative artificial intelligence," "ChatGPT," "large language model," and "LLM." The dictionary is narrower than the one used by Kikuchi (2026), which also includes named models such as GPT-4, Claude, Gemini, and Copilot. I use the narrower set because the analysis is based on a smaller thesis sample. This means that the adoption measure may be more conservative than Kikuchi's. Once a bank is classified as an adopter, the indicator remains equal to one in subsequent quarters.

An initial 2022–2024 filing window, matching the immediate aftermath of ChatGPT's public release, produced almost no matches restricted to 10-K/10-Q forms: exploratory searches show the earliest substantive 10-K disclosures of GenAI activity among the sample banks appear only in filings for fiscal year 2024, submitted in early 2025. The observation window was therefore extended to 2021–2025. This provides one to two years of pre-shock observations and allows more time for GenAI disclosures to appear in later filings. Under this window, four of the six sample banks (JPMorgan Chase, Bank of America, U.S. Bancorp, and PNC) show at least one qualifying GenAI mention in a 10-K or 10-Q filing; 28 of 121 bank-quarter observations with complete financial data (23.1 percent) are coded as GenAI-adopting, concentrated in the post-2023 period (Table 2.3).

Financial performance (Federal Reserve FR Y-9C).

Quarterly financial data are drawn from the FR Y-9C, the Consolidated Financial Statements for Bank Holding Companies, obtained from the Federal Reserve's National Information Center (NIC) bulk data files for 2021 Q1 through 2025 Q4. Each quarterly file is a caret-delimited text extract keyed by the reporting institution's RSSD ID, containing every FR Y-9C data item filed that quarter across all reporting bank holding companies industry-wide; sample banks are identified by filtering on their RSSD identifiers (Table 2.1).

Return on assets and return on equity are constructed from three reported items: net income (MDRM code BHCK4340), average total consolidated assets (BHCK3368, Schedule HC-K), and total equity capital (BHCKG105 for the 2009 Q1–present period, with BHCK3210 retained as a fallback for any earlier filings). Because net income is reported on a year-to-date cumulative basis rather than for the single quarter, the raw figure is first divided by the number of elapsed quarters in the filing year before being annualized:

$$ROA_it = [(NetIncome_YTD,it \div q_t) \div AvgAssets_it] \times 400$$

Where q_t denotes the number of quarters elapsed in the fiscal year as of quarter t. The corresponding ROE formula substitutes average total equity for average total assets in the denominator. This adjustment was necessary because the unadjusted calculation produced increasingly high ROA and ROE values across quarters within the same year. For example, one initial calculation showed a large bank's ROA increasing from 1.12% in Q1 to 4.05% in Q4. This resulted from the year-to-date reporting convention rather than an actual change in profitability. the corrected formula above resolves this, yielding a stable quarterly ROA in the 0.8–1.2 percent range typical of large U.S. bank holding companies over the sample period.

Small-tier financial data: resolution.

The sample's two originally identified small-bank-tier institutions — FFD Financial Corp. and Consumers Bancorp, Inc. both fall below the $3 billion consolidated-asset threshold that determines FR Y-9C filing status, and instead file the abbreviated, semi-annual FR Y-9SP with the Federal Reserve. Because the FR Y-9SP does not populate the standard FR Y-9C MDRM item

codes used for the other five sample banks, financial data for these two institutions could not be recovered from the FFIEC bulk files by that route.

As a fallback, financial data for Consumers Bancorp were instead constructed from the company's own SEC XBRL filings (data.sec.gov), using the standardized us-gaap:Assets, us-gaap:StockholdersEquity, and us-gaap:NetIncomeLoss concepts that all SEC-reporting companies tag regardless of their Federal Reserve filing status. Because Consumers Bancorp's fiscal year does not align with the calendar year for reporting purposes in the way its quarterly XBRL facts are tagged, one quarter per fiscal year (the quarter ending in June) is not separately reported as a standalone ~90-day fact; that quarter's net income was derived by subtracting the three known quarters from the corresponding annual (10-K) total. Five of Consumers Bancorp's twenty-one quarterly observations were constructed this way and are labeled accordingly in the underlying data.

FFD Financial Corp. could not be included using either data source: inspection of its complete SEC XBRL filing history (all 93 distinct us-gaap concepts the company has ever tagged) shows no data of any kind under any concept filed after March 31, 2012. Cross-referencing against the company's SEC filing history confirms its last 10-K or 10-Q filing of any type was submitted May 15, 2012. FFD Financial Corp. therefore appears to have ceased SEC reporting obligations more than eight years before the start of the sample period and has been excluded from the sample. The final sample therefore contains six bank holding companies instead of the seven originally planned. the Small tier is represented by Consumers Bancorp alone (Table 2.1).

Sample composition.

The final sample contains six publicly traded U.S. bank holding companies across three size tiers. This provides variation in both GenAI adoption and bank size for the H1 and H3 analyses while keeping the panel manageable for the thesis. The Small tier, originally planned with two institutions, is represented by one.

Table 0.1: Sample bank holding companies

Bank Holding Company	Size Tier	SEC CIK	FRB RSSD ID	Financial Data Source

JPMorgan Chase & Co.	Large	0000019617	1039502	FR Y-9C
Bank of America Corp.	Large	0000070858	1073757	FR Y-9C
Wells Fargo & Co.	Large	0000072971	1120754	FR Y-9C
U.S. Bancorp	Regional	0000036104	1119794	FR Y-9C
PNC Financial Services Group	Regional	0000713676	1069778	FR Y-9C
Consumers Bancorp, Inc. (OH)	Small	0001006830	2286228	SEC XBRL

2.3 Comparative EU/Italy Benchmark Data

The EU/Italy analysis is descriptive rather than causal because comparable firm-level financial data are not available from Eurostat and Istat. Two aggregate sources are used, there is an important limitation concerning the sector covered by the first source. Eurostat's isoc_eb_ai dataset reports the share of enterprises using artificial intelligence technologies, disaggregated by country, enterprise size class, and a single NACE Rev. 2 activity aggregate, published under the code C10-S951_X_K. On verification, this code covers the broad non-financial business economy and explicitly excludes NACE Section K (financial and insurance activities); Eurostat does not publish enterprise-level AI-adoption data for the financial sector in this public series at all, for any country, at any level of disaggregation currently available. This limitation comes from the design of Eurostat's ICT usage in enterprises survey. Financial and insurance companies are not included in the relevant public series. Therefore, changing the dataset filter would not provide a financial-sector-specific measure. The EU/Italy figures reported in this thesis (Sec. 3.5) are therefore economy-wide, non-financial-sector figures, and The EU and Italy figures should be treated as general indicators of AI-adoption timing. They are not directly comparable with the U.S. banking sample. Second, Istat's enterprise ICT-usage surveys provide the corresponding Italian national figures, subject to the same sector-scope limitation, allowing Italy to be benchmarked against the broader EU-27 average drawn from the same Eurostat series.

Because neither Eurostat nor Istat publishes firm-level financial statements alongside adoption statistics, the comparative section cannot replicate the bank-level panel design used for the U.S.

sample. Instead, Sec. 3.5 presents adoption-rate trends for the EU aggregate and Italy over the sample period, situating the U.S. findings within a broader international pattern.

2.4 Variable Construction and Summary Statistics

Table 2.2 summarizes the construction of each variable used in the empirical specifications of Sec. 2.5. The resulting panel is bank-quarter structured, comprising six banks: five observed quarterly over twenty quarters (2021 Q1–2025 Q4) via FR Y-9C data, and Consumers Bancorp observed over twenty-one quarters via SEC XBRL data (2021 Q1–2026 Q1, five of which are derived from annual totals as described above). This yields 121 bank-quarter observations with complete ROA/ROE data.

Table 0.2: Variable construction

Variable	Definition	Source	Construction
GenAI_it	Binary indicator: bank i has disclosed GenAI-related activity by quarter t	SEC EDGAR full-text search (10-K/10-Q)	=1 from the first quarter in which a qualifying term appears in a filing, persisting thereafter
ROA_it	Return on average assets, annualized (%)	FR Y-9C, Schedule HI / HC-K	(Net income YTD ÷ elapsed quarters) ÷ average total assets × 400
ROE_it	Return on average equity, annualized (%)	FR Y-9C, Schedule HI / HC-K	(Net income YTD ÷ elapsed quarters) ÷ average total equity × 400
Size_i	Categorical bank-size tier (Large / Regional / Small)	SEC EDGAR (10-K disclosed total assets)	Assigned once per bank at sample construction, time-invariant
Post_t	Binary indicator for the post-ChatGPT period	Calendar quarter	=1 for quarters beginning Q1 2023 onward (see §2.5)

DigMat_i	Digital-maturity proxy, time invariant per bank	SEC EDGAR filing text (plus EX-13 exhibits where incorporated by reference)	Keyword count (39 multi-word phrases + 8 single words, word-boundary matched) ÷ total filing word count, ×10,000; see Table 2.4

Table 0.3: Summary statistics

Group	n	Mean ROA	SD ROA	Mean ROE	SD ROE	Adoption Rate
Overall	121	1.052	0.210	12.652	3.112	0.231
Large tier	60	1.063	0.222	12.237	3.181	0.200
Regional tier	40	1.088	0.197	11.869	1.739	0.400
Small tier	21	0.950	0.169	15.330	3.616	0.000
Pre-shock (before 2023 Q1)	48	1.093	0.231	13.064	3.584	0.000
Post-shock (2023 Q1 onward)	73	1.024	0.191	12.382	2.751	0.384
Non-adopters	93	1.038	0.210	12.741	3.183	0.000
GenAI adopters	28	1.097	0.206	12.357	2.898	1.000

Mean ROA is broadly stable across the Large and Regional tiers (1.06–1.09 percent) and somewhat lower for the single Small-tier bank (0.95 percent), though with only one institution in that tier this figure should not be read as representative of small community banks generally. Across adopter status, GenAI adopters show a modestly higher mean ROA than non-adopters (1.10 percent versus 1.04 percent), a small unconditional gap consistent with — though not yet a test of — the positive selection effect hypothesized in H1; the panel and difference-in-differences estimates in sec. 3.2–3.3 are required to assess whether this gap is statistically distinguishable from zero once bank and quarter fixed effects are included.

The digital-maturity proxy for H4 was constructed using the approach described above: each bank's most recent pre-2022 10-K filing (the fiscal-year-2020 annual report, filed in early-to-mid 2021 for all six sample banks) was retrieved in full from SEC EDGAR, including EX-13 annual-report exhibits where the primary 10-K document incorporates detailed business description and financial content by reference, Some larger banks incorporate detailed business information into exhibits by reference. Including these exhibits helps avoid understating their digital-maturity scores. Each filing's full text was searched for a dictionary of 39 multi-word phrases (e.g., “cloud computing,” “digital banking,” “application programming interface”) plus 8 single words matched with word-boundary regular expressions to avoid partial-word false positives (e.g., “cloud,” “fintech,” “blockchain”). The word “mobile” was deliberately excluded as a standalone term, since bank 10-Ks frequently discuss “mobile home loans” as a lending-portfolio category unrelated to digital maturity; “mobile banking,” “mobile app,” and “mobile device” remain covered as multi-word phrases. The resulting DigMat_i score is the total keyword count divided by the filing's total word count, expressed per 10,000 words (Table 2.4)

Table 0.4: Digital-maturity proxy (DigMat), pre-2022 10-K filings

Bank	Size Tier	Word Count	Keyword Hits	DigMat Score (per 10,000 words)
JPMorgan Chase & Co.	Large	184,524	13	0.7045
Bank of America Corp.	Large	138,234	24	1.7362
Wells Fargo & Co.	Large	159,971	26	1.6253
U.S. Bancorp	Regional	109,926	18	1.6375
PNC Financial Services Group	Regional	134,807	30	2.2254
Consumers Bancorp, Inc. (OH)	Small	31,973	1	0.3128

Consumers Bancorp has the lowest DigMat score (0.31) in the sample. This is substantially below the scores of the other five banks. Among the remaining five banks, scores do not scale

monotonically with size: PNC Financial Services Group records the highest score (2.23) and JPMorgan Chase the lowest among the non-Small-tier banks (0.70), despite JPMorgan's larger asset base, This suggests that the keyword-based measure does not simply increase with bank size in this sample.

2.5 Panel Regression, Difference-in-Differences, and Moderation Specifications

Kikuchi (2026) tests the same four-hypothesis structure this thesis adopts (frontier-firm selection, implementation tax, size heterogeneity, and a fourth moderation hypothesis) using two complementary estimators: a Dynamic Spatial Durbin Model (DSDM) with Bayesian MCMC estimation to capture cross-sectional "frontier firm" effects and network spillovers across 809 institutions, and Synthetic Difference-in-Differences (SDID) with bootstrapped inference (200 replications) to isolate the causal within-bank effect of adoption around the 2023 Q1 ChatGPT shock. This thesis does not replicate those methods. The spatial weight matrices, MCMC estimation, and synthetic-control weighting are beyond the scope of the six-bank sample. Instead, the analysis uses a simplified two-way fixed-effects panel and difference-in-differences approach. These models address the same four hypotheses but are designed for the smaller sample. The results should therefore not be interpreted as a replication of Kikuchi's causal estimates, which were identified from a sample roughly two orders of magnitude larger.

Panel regression (H1).

The baseline specification tests whether GenAI-adopting banks exhibit systematically higher profitability, consistent with a positive selection effect — the pattern Kikuchi's DSDM estimates as $\beta > 0$ — using a two-way fixed-effects panel model:

$$Performance_it = \beta_1 \, GenAI_it + \alpha_i + \gamma_t + \varepsilon_it$$

Where Performance_it is ROA or ROE for bank i in quarter t, α_i is a bank fixed effect absorbing time-invariant institutional characteristics, and γ_t is a quarter fixed effect absorbing macroeconomic and industry-wide shocks common to all banks. Standard errors are clustered at the bank level. A positive and significant β_1 is consistent with H1.

Difference-in-differences (H2).

To isolate a within-bank treatment effect around the ChatGPT shock rather than a cross-sectional association the role Kikuchi's SDID estimator plays the panel is re-estimated with an interaction between adoption status and a post-shock period indicator:

$$Performance_it = \beta_2 (GenAI_i \times Post_t) + \alpha_i + \gamma_t + \varepsilon_it$$

Post_t is defined as equal to one for quarters from 2023 Q1 onward, matching Kikuchi's treatment-period definition. A negative and significant β_2 would be directionally consistent with the Implementation Tax Kikuchi estimates (−46 basis points ROA, −428 basis points ROE, aggregate sample), though this thesis's simplified two-way fixed-effects design is not equivalent to Kikuchi's synthetic-control-weighted estimator and should not be expected to recover comparable magnitudes from a six-bank sample.

Size moderation (H3).

H3 tests whether the adoption effect varies with institutional scale by interacting the adoption indicator with the size-tier variable:

$$Performance_it = \beta_3 GenAI_it + \beta_4 (GenAI_it \times Size_i) + \alpha_i + \gamma_t + \varepsilon_it$$

There is an important difference between Kikuchi's H3 hypothesis and the empirical result. The hypothesis suggests that large banks should experience a more pronounced decline because of greater legacy-system complexity. However, the paper's own SDID results find the opposite: smaller banks (bottom three quartiles) show a substantially larger ROE decline of 517 basis points, compared with 129 basis points for large banks (top quartile). The paper attributes this result to economies of scale in implementation rather than to legacy-system complexity. Therefore, the H3 test in this thesis should be interpreted in relation to Kikuchi's empirical finding that smaller banks experience a larger tax, rather than the original hypothesis that large banks would be more affected.

Digital-maturity moderation (H4)

Kikuchi's H4 differs from the H4 used in this thesis. In Kikuchi (2026), H4 concerns positive network spillovers and is tested using the DSDM spatial spillover parameter θ. It is not a digital-maturity moderation hypothesis. Kikuchi does use a "Digital Index" (a weighted count of technology keywords in 10-K filings), but only as a control variable in the DSDM specification, not as an interaction term testing a distinct hypothesis. This thesis's proposal frames H4 as "the effect of AI adoption depends on the digital maturity of the organization," The H4 is used as an extension of Kikuchi's Digital Index rather than a replication of the paper's H4. The specification below tests this thesis-specific hypothesis, not Kikuchi's:

$$Performance_it = \beta_5\ GenAI_it + \beta_6\ (GenAI_it \times DigMat_i) + \alpha_i + \gamma_t + \varepsilon_it$$

The thesis retains the digital-maturity moderation specification because estimating spatial spillovers with only six banks would require a substantially more complex model and would provide limited additional information.

Chapter 3: Results and Discussion

3.1 Descriptive Trends in GenAI Adoption

SEC full-text search identified 10 qualifying filings mentioning GenAI-related terms ("generative AI" or "generative artificial intelligence") across the six sample banks, 2021–2026. Table 3.1 breaks these down by bank. Four of the six banks — JPMorgan Chase, PNC, U.S. Bancorp, and Bank of America show at least one qualifying mention; Wells Fargo and Consumers Bancorp show none in this window. First mentions cluster tightly around February 2024 (JPMorgan, PNC, and U.S. Bancorp all first mention GenAI within a five-day window, coinciding with fiscal-year-2023 10-K filings), with Bank of America's first mention arriving roughly a year later, in its fiscal-year-2024 10-K (February 2025). Once flagged, the adoption indicator is held at 1 for all subsequent quarters (Sec. 2.2), yielding 28 adopter-quarters out of 121 total bank-quarter observations (23.1 percent).

Table 0.1: GenAI mentions and adoption timing by bank

Bank	Mentions (raw filings)	First mention date	Adopter-quarters (persistent)
JPMorgan Chase & Co.	3	2024-02-16	8 / 20
PNC Financial Services Group	3	2024-02-21	8 / 20
U.S. Bancorp	3	2024-02-20	8 / 20
Bank of America Corp.	1	2025-02-25	4 / 20
Wells Fargo & Co.	0	—	0 / 20
Consumers Bancorp, Inc. (OH)	0	—	0 / 21

An important finding is that none of the banks in the sample reported using or mentioning GenAI before 2024. Therefore, the 48 observations covering the period before Q1 2023 contain no

GenAI adopters. This timing pattern has important implications for the difference-in-differences analysis presented in Sec. 3.3.

Table 0.2: Return on assets and return on equity by group

Group	n	Mean ROA (%)	Mean ROE (%)
Large tier	60	1.063	12.237
Regional tier	40	1.088	11.869
Small tier	21	0.950	15.330
Non-adopters	93	1.038	12.741
GenAI adopters	28	1.097	12.357
Pre-shock (before 2023 Q1)	48	1.093	13.064
Post-shock (2023 Q1 onward)	73	1.024	12.382

Unconditionally, GenAI adopters show a modestly higher mean ROA than non-adopters (1.097 percent versus 1.038 percent) and a modestly lower mean ROE (12.357 percent versus 12.741 percent) a small, mixed-direction gap that does not by itself indicate either a selection or a treatment effect; sec. 3.2 and sec. 3.3 test whether either pattern survives controlling for bank and quarter fixed effects.

3.2 Panel Regression Results

Table 3.3 reports the two-way fixed-effects panel regression specified in Sec. 2.5 (H1): Performance_it = β1·GenAI_it + α_i + γ_t + ε_it, estimated separately for ROA and ROE with standard errors clustered at the bank level.

Table 0.3: Panel fixed-effects results (H1)

Dependent variable	Coefficient (β1)	Std. Err.	t-stat	p-value
ROA (%)	0.047	0.130	0.36	0.720
ROE (%)	0.833	1.981	0.42	0.675

Neither coefficient is statistically distinguishable from zero ($p = 0.720$ for ROA, $p = 0.675$ for ROE). Therefore, the results do not provide statistically significant support for H1 in this sample. Two characteristics of the data may contribute to the lack of statistical significance. First, the analysis uses only six bank clusters, which creates a small-cluster problem for cluster-robust standard errors. Second, the point estimates are positive, although they are imprecisely estimated: with only 6 bank clusters, cluster-robust standard errors are known to be unreliable and generally too large (a well-documented small-cluster problem in panel econometrics), and the point estimates themselves are directionally positive (consistent with Kikuchi's $\beta > 0$ finding) even though imprecisely estimated. Kikuchi (2026) estimates the equivalent selection coefficient from 24,270 bank-quarter observations across 809 institutions and finds it positive and significant at the 1–5 percent level for both ROA and ROE; this thesis's 121-observation, 6-bank sample is roughly two orders of magnitude smaller and The much smaller sample provides less statistical power to detect an effect of a similar size.

3.3 Difference-in-Differences Results (Implementation Tax)

Table 3.4 reports the difference-in-differences specification from §2.5 (H2): $Performance_{it} = \beta 2 \cdot (GenAI_i \times Post_t) + \alpha_i + \gamma_t + \varepsilon_{it}$, where $Post_t = 1$ for quarters from 2023 Q1 onward.

Table 0.4: Difference-in-differences results (H2)

Dependent variable	Coefficient (β2)	Std. Err.	t-stat	p-value
ROA (%)	0.047	0.130	0.36	0.720
ROE (%)	0.833	1.981	0.42	0.675

The coefficients are identical to the H1 estimates in Table 3.3 because of the timing of GenAI adoption in the sample. As shown in sec. 3.1, every adopting bank in this panel first mentions GenAI after the first quarter of 2023. Because there are no pre-shock adopters, the interaction term ($GenAI_i \times Post_t$) is mathematically identical to the baseline indicator ($GenAI_{it}$). Therefore, the difference-in-differences model cannot separately identify a within-bank treatment effect from the cross-sectional association estimated in H1; β2 is simply not identified separately from β1.

Therefore, the six-bank sample does not allow H2 to be tested independently. Kikuchi (2026) bypasses this issue because the paper's broader 2018–2025 dataset includes institutions that adopted AI well before ChatGPT's release, supplying the pre-treatment variation that the estimator requires. Replicating that variation here would require either extending the observation window earlier than 2021 or broadening the sample size to capture pre-2023 adopters.

3.4 Moderation Results

H3, as specified in Sec. 2.5, interacts GenAI adoption with the Small size tier. That specification could not be estimated: Table 3.1 shows Consumers Bancorp, the sample's only Small-tier institution, has zero GenAI mentions throughout the panel, so the interaction term GenAI_it × Small_i is a column of zeros and is perfectly collinear with (uninformative relative to) the main effect. Because the original Small-tier interaction could not be estimated, a Large-tier interaction was used as a substitute. This is not a replacement for the original H3 test, but it provides the closest estimable comparison in this sample. Table 3.5 reports GenAI_it × Large_i instead, comparing the three Large-tier banks against the pooled Regional and Small tiers.

Table 0.5: Size-moderation results (H3, Large-tier interaction substituted for Small-tier)

Dependent variable	GenAI (β3)	GenAI × Large (β4)	p-value (β4)
ROA (%)	−0.036	0.178	0.132
ROE (%)	−0.158	2.133	0.023**

For ROE, the interaction is positive and significant at the 5 percent level (β4 = 2.133, p = 0.023): among adopters, Large-tier banks show a return-on-equity roughly 2.1 percentage points higher than non-Large adopters, relative to their respective non-adopter baselines. This result differs from Kikuchi (2026), whose SDID estimates show a larger Implementation Tax for smaller banks. the paper's own SDID results find the opposite pattern, smaller banks bearing a larger “Implementation Tax” (a 517-basis-point ROE decline versus 129 basis points for large banks), meaning adoption is relatively more costly, not more beneficial, for large institutions in Kikuchi's much larger sample. The difference from Kikuchi's finding should be interpreted cautiously because the present

analysis is based on only three Large-tier and three non-Large-tier banks, The result may therefore reflect differences between the specific firms in the sample rather than a general size effect.

H4 was tested using the digital-maturity proxy described in Sec. 2.4, which is available for all six sample banks (Table 2.4). The specification follows Sec. 2.5: Performance_it = β5·GenAI_it + β6·(GenAI_it × DigMat_i) + α_i + γ_t + ε_it, estimated with the same two-way fixed-effects, bank-clustered design used throughout this chapter. Because DigMat_i is time-invariant per bank, its own main effect is absorbed by the bank fixed effect α_i and cannot be separately estimated, only the interaction with GenAI_it, which varies within-bank as adoption switches on, is identified. Table 3.6 reports the results.

Table 0.6: Digital-maturity moderation results (H4, GenAI × DigMat interaction)

Dependent variable	GenAI (β5)	GenAI × DigMat (β6)	p-value (β6)
ROA (%)	0.308	−0.167	0.076*
ROE (%)	3.626	−1.786	0.033**

The GenAI × DigMat interaction (β6) is negative for both outcomes, and statistically significant for ROE (β6 = −1.786, p = 0.033) with a similarly signed but only marginally significant estimate for ROA (β6 = −0.167, p = 0.076). The interaction has the same negative sign for both outcomes. Among adopting banks, higher pre-existing digital maturity is associated with a smaller increase in performance after adoption. This is contrary to the expectation that digitally mature banks would obtain greater benefits from GenAI and more consistent with a diminishing-returns or substitution pattern in which GenAI adds comparatively less incremental value where digital infrastructure was already extensive.

Two limitations should also be considered when interpreting H4. First, Because the interaction is estimated from only four adopting banks, the result may reflect bank-specific differences rather than a general relationship between digital maturity and GenAI adoption. Second, a computational note for transparency: these estimates were computed directly in Python (pandas/linearmodels) rather than the R/fixest environment Sec. 2.5 specifies as the thesis's intended implementation. The H1 and H3 point estimates were reproduced to within rounding in the Python cross-check,

providing an internal consistency check, but cluster-robust standard errors can differ slightly across statistical packages when the cluster count is this small (6 clusters), due to differing small-sample corrections. The significance pattern reported here ROE significant, ROA marginal should be re-confirmed by running the same specification in fixest before this result is considered final.

3.5 EU/Italy Comparative Benchmark

The AI-adoption figures reported in Table 3.7 are economy-wide figures that exclude the financial and insurance sector (NACE Section K) they are not a financial-sector-specific benchmark**,** Hence, these figures are not directly comparable with the U.S. banking-sector results in Sec. 3.1 through Sec. 3.4. Table 3.7 reports the trend extracted from the uploaded Eurostat dataset (isoc_eb_ai), using the "enterprises using any AI technology" indicator (E_AI_TANY) for enterprises with 10 or more employees, the dataset's standard reporting threshold.

Table 0.7: AI adoption rate, EU-27 average vs. Italy

Year	EU-27 average (%)	Italy (%)
2021	7.65	6.17
2023	8.06	5.05
2024	13.48	8.20
2025	19.95	16.40

Both series show a substantial increase in AI adoption after 2023. By 2025, the reported adoption rates were roughly three times their 2023 levels. A similar timing pattern appears in the U.S. bank sample, where the first GenAI mentions occur in 2024 or later.

A structural limitation in the Eurostat survey design affects this comparison. The Eurostat dataset (isoc_eb_ai) aggregates the surveyed enterprise population under the code C10-S951_X_K, which explicitly excludes NACE Section K (financial and insurance activities); Eurostat's underlying ICT usage in enterprises survey does not sample financial and insurance companies, so no filter within this dataset can isolate the financial sector (Sec. 2.3). Consequently, the trends reported in Table 3.7 represent an economy-wide, non-financial-sector baseline rather than a direct financial-sector

benchmark. The lack of financial-sector observations prevents a direct sector-to-sector comparison. However, both datasets show increased AI adoption after 2023.

3.6 Robustness Checks

Because the sample contains only six banks, the robustness checks are limited. They are included to assess whether the main results change under alternative specifications.

Dropping quarter fixed effects (bank fixed effects only) leaves the GenAI coefficient small and statistically insignificant for both outcomes (ROA: −0.039, p = 0.569; ROE: −0.790, p = 0.258) notably, the point estimate flips sign relative to the two-way fixed-effects specification in Table 3.3, though neither is distinguishable from zero. A pooled OLS specification with no fixed effects at all (the naive cross-sectional comparison) shows a small positive, insignificant GenAI coefficient for ROA (0.059, p = 0.492) and a small negative, insignificant coefficient for ROE (−0.384, p = 0.777). Across the three specifications, none of the H1/H2 coefficients is statistically significant. However, the small number of bank clusters limits the statistical power of these tests. Because the coefficient changes sign across some specifications, individual point estimates should be interpreted cautiously. This also applies to the H3 and H4 results, which are based on the same six-bank sample.

3.7 Interpretation: Reconciling Selection and Treatment Effects

Kikuchi (2026) reconciles a positive cross-sectional association (DSDM, $\beta > 0$) with a negative causal estimate (SDID, ATT < 0) by distinguishing between selection and treatment: frontier banks self-select into early adoption, while the adoption process itself imposes short-run implementation costs. The results of this thesis do not reproduce that distinction because the sample does not provide the variation needed to estimate the two effects separately.

The panel estimate for H1 (Sec. 3.2) is directionally positive but statistically indistinguishable from zero, offering inconclusive support for a selection effect. Furthermore, the difference-in-differences estimate for H2 (Sec. 3.3) is not separately identified from H1 in this sample. Because no bank in the six-bank panel adopted GenAI prior to the 2023 Q1 shock date, there is no genuine "treatment" variation for the DiD design to isolate. Therefore, the data cannot determine whether

a separate negative ‘Implementation Tax’ exists independently of the selection effect. The H4 result does not resolve this identification problem. Instead, it shows that the association between GenAI adoption and performance differs according to pre-existing digital maturity: since H1 and H2 are not separately identified to begin with, H4's interaction is itself a modifier of an underlying effect that this sample cannot cleanly characterize as selection, treatment, or some mixture of the two.

These results therefore reflect the limitations of the six-bank sample rather than providing evidence against Kikuchi's findings. Kikuchi's reconciliation relies on identifying variation across 809 institutions over an eight-year panel, including adopters both before and after the ChatGPT shock variation this smaller sample lacks. Future research could address this limitation by expanding the sample to include banks whose first GenAI mentions occurred before Q1 2023. Extending the observation period back to 2018, as in Kikuchi (2026), could also capture earlier adoption activity. This would help capture earlier, less publicized adoption activity that the narrow five-term keyword search used here (Sec. 2.2) likely missed. Without these modifications, the data indicates that GenAI adoption in this specific U.S. bank sample has no statistically detectable association with short-run profitability in either direction. With the current sample, it is not possible to determine whether the observed association reflects selection, treatment, or a combination of both

Chapter 4: Conclusion

4.1 Summary of Findings

This thesis set out to test, on a small sample of U.S. bank holding companies, the same selection-versus-treatment structure Kikuchi (2026) documents at scale: that GenAI-adopting banks might appear more productive in cross-section (a selection effect, H1) while the act of adoption itself imposes a short-run "Implementation Tax" on profitability (a treatment effect, H2), with that tax varying by institutional size (H3) and, potentially, by pre-existing digital maturity (H4).

The results presented in Chapter 3 do not reproduce Kikuchi's findings. The differences between the two studies are important for interpreting the results of this thesis.H1's panel estimate was directionally positive but not statistically distinguishable from zero for either ROA or ROE (Sec. 3.2), offering only weak, inconclusive support for a selection effect. H2 could not be tested independently of H1 because every bank in the sample that mentions GenAI does so no earlier than 2024, so there is no pre-2023 adopter for the difference-in-differences design to compare against, and the H2 coefficients turned out numerically identical to H1's as a direct, mechanical consequence (Sec. 3.3). H3, as originally specified, could not be estimated either — the sample's only Small-tier bank never mentions GenAI, leaving that interaction term uninformative, so a Large-tier interaction was substituted instead, and it produced one significant result: Large-tier banks showed roughly 2.1 percentage points higher ROE among adopters than non-Large adopters (Sec. 3.4), the opposite direction from Kikuchi's own population-level finding that the "Implementation Tax" falls more heavily on smaller banks. Because the sample contains only three Large-tier and three non-Large-tier banks, this result should be interpreted as a firm-specific pattern rather than as evidence of a general size effect. H4 has since been tested using a digital-maturity proxy built from the banks' actual pre-2022 filing text (Sec. 2.4): the GenAI × DigMat interaction was negative for both outcomes, statistically significant for ROE and only marginally so for ROA (Sec. 3.4), meaning that among adopters, banks with higher pre-existing digital maturity showed a smaller performance gain from GenAI adoption than banks with lower digital

maturity. Only four banks in the sample adopted GenAI, so this result should be interpreted cautiously. It may reflect differences between the specific banks in the sample rather than a general moderation effect; and because H1 and H2 are not separately identified in this sample to begin with (Sec. 3.3), H4's interaction is itself a modifier of an underlying selection-or-treatment effect that this data cannot cleanly characterize as one, the other, or some mixture of the two (Sec. 3.7).

The clearest pattern in the analysis was found in the descriptive comparisons in Sections 3.1 and 3.5: GenAI mentions across the sample banks cluster tightly in 2024 and later, and the Eurostat AI-adoption trend — though, as Sec. 3.5 makes clear, an economy-wide rather than financial-sector-specific figure — shows the same post-2023 acceleration, roughly tripling between 2023 and 2025 for both the EU-27 average and Italy. The timing pattern is consistent across both data sources even though the causal and moderation tests built on top of it were not cleanly identified in this sample.

4.2 Implications for Bank Management and Policy

Because H1 and H2 could not be separately identified, the results do not establish whether GenAI adoption causes a short-run reduction in profitability. They also do not establish whether banks that adopt GenAI were already higher-performing before adoption. the central question the thesis set out to answer remains open on the evidence gathered here. What can be said with more confidence is about timing: across every data source used in this thesis, the U.S. banks' own SEC filings and the EU/Italy Eurostat series alike, substantive GenAI adoption activity is concentrated from 2024 onward, not immediately following ChatGPT's November 2022 release. For banks considering when to adopt GenAI, the results suggest that adoption in 2024–2025 was still relatively early compared with the banks in this sample. The evidence does not suggest that the early-adoption period had already ended in 2023.

The one significant result this thesis did produce, a positive association between Large-tier adoption and ROE (Sec. 3.4) runs counter to Kikuchi's (2026) own finding that smaller banks bear a disproportionately larger “Implementation Tax.” This result is based on only three Large-tier banks and is also affected by the lack of pre-shock variation that limited the H2 analysis. Therefore, it should not be used on its own to draw conclusions for bank management or policy; if anything,

the discrepancy between this thesis's result and Kikuchi's larger-sample finding argues for weighting Kikuchi's population-level evidence more heavily than this thesis's own small-sample result when the two disagree. For financial-stability policy specifically, this thesis's own data cannot speak to the “algorithmic coupling” and systemic-spillover concerns Kikuchi (2026) raises, since testing them would require the kind of spatial network analysis across many institutions that this thesis's six-bank sample was never designed to support (Sec. 1.3.5); to the extent those concerns are relevant, they rest on Kikuchi's own larger-sample evidence, not on any finding reported in this thesis.

The negative GenAI × DigMat interaction is consistent with the Large-tier result. In this sample, the positive association between GenAI adoption and performance was larger among banks with lower levels of pre-existing digital maturity. This result should be treated as a hypothesis rather than a firm conclusion. One possible explanation is that GenAI may substitute for some existing digital capabilities rather than complement them. a bank with less legacy digital capability may have more low-hanging fruit for GenAI to capture. Given the same four-adopter identification limitation that qualifies the H3 result, and given this estimate has not yet been reproduced in the R/fixest environment Sec. 2.5 specifies as the thesis's intended implementation, Given these limitations, the results are not sufficient to conclude that existing digital maturity predicts the returns from GenAI adoption.

4.3 Limitations and Future Research

Several limitations affect the interpretation of these findings and should be addressed in future research. Sample size and power. The core limitation is scale: six banks and 121 bank-quarter observations, clustered into only six groups for cluster-robust inference, is roughly two orders of magnitude smaller than Kikuchi's (2026) 809-institution, 24,270-observation panel. The robustness checks in Sec. 3.6 also showed that the sign of the GenAI coefficient changed across some alternative fixed-effects specifications, and that instability extends to the significant H3 and H4 results as well, since they rest on the same six-bank base and the same small-cluster inference concern — informative, in itself, about how little this sample can support strong inference in either direction.

Loss of a planned sample bank. The original seven-bank design (Sec. 2.2) was reduced to six after FFD Financial Corp. was found to have ceased SEC reporting in 2012, well before the sample period begins. The remaining Small-tier bank, Consumers Bancorp, required a different financial-data source (SEC XBRL rather than FR Y-9C) than the other five banks, and five of its twenty-one quarterly observations were derived from annual totals rather than directly reported (Sec. 2.2).

No pre-treatment variation. Because every adopter in this sample first mentions GenAI in 2024 or later, the difference-in-differences design could not separate a causal treatment effect from the cross-sectional association captured by H1 (Sec. 3.3). A follow-up study would need either a substantially larger bank sample — large enough that some adopters' first mentions predate the November 2022 ChatGPT release — or a keyword search extended further back in time to catch earlier, less publicized adoption activity that this thesis's five-term dictionary (narrower than Kikuchi's, Sec. 2.2) may have missed.

H4 requires further confirmation. H4 has been tested using a digital-maturity proxy built from the banks' actual pre-2022 filing text (Sec. 2.4), and the resulting GenAI × DigMat interaction is negative and significant for ROE, with a similarly signed but only marginally significant estimate for ROA (Sec. 3.4). Two qualifications limit how far this result can be pushed. First, only four of the six sample banks ever adopt GenAI, so the interaction is effectively estimated from four banks' fixed digital-maturity scores interacting with their own adoption timing — closer to a firm-specific effect than a generalizable moderation result. Second, these estimates were computed directly in Python (pandas/linearmodels) rather than the R/fixest environment Sec. 2.5 specifies as the thesis's intended implementation; a cross-check reproduced the H1 and H3 point estimates to within rounding, but cluster-robust standard errors can differ slightly across statistical packages at this cluster count (six), so the significance pattern reported here should be confirmed by re-running the specification in fixest before being treated as final.

Limitation of the EU/Italy benchmark data. The Eurostat series used in Sec. 3.5 was found, in the course of writing Chapter 3, to report AI-adoption figures for the non-financial business economy only — Eurostat's isoc_eb_ai survey does not sample financial and insurance-sector enterprises (NACE Section K) at all, for any country, at any level of disaggregation. This limitation comes from the design of the Eurostat survey rather than from the way the data were extracted. Changing

the filter within isoc_eb_ai would therefore not solve the problem. Sec. 2.3 and Sec. 3.5 present the EU/Italy figures as an economy-wide, non-financial-sector benchmark rather than a sector-matched one for this reason. A genuinely financial-sector-specific EU/Italy comparison would require a different data source entirely — a country-level, financial-sector-specific ICT-usage survey, if one exists, or firm-level disclosures from EU banks analogous to the SEC-filing approach used for the U.S. sample.

These limitations suggest several priorities for future research. First, the bank sample should be expanded and the observation period extended so that the difference-in-differences analysis has meaningful pre- and post-adoption variation. Second, the H4 digital-maturity result should be re-estimated using the R/fixest approach specified in Section 2.5 rather than the Python-based estimation used in Section 3.4. Finally, a different data source would be needed if future research aims to construct a financial-sector-specific EU/Italy benchmark.

Reference

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review, 105*(2), 564–608. https://doi.org/10.1257/aer.20130456

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Berger, A. N. (2003). The economic effects of technological progress: Evidence from the banking industry. *Journal of Money, Credit and Banking, 35*(2), 141–176. https://doi.org/10.1353/mcb.2003.0009

Berger, A. N., Miller, N. H., Petersen, M. A., Rajan, R. G., & Stein, J. C. (2005). Does function follow organizational form? Evidence from the lending practices of large and small banks. *Journal of Financial Economics, 76*(2), 237–269. https://doi.org/10.1016/j.jfineco.2004.06.003

Bertrand, M., & Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics, 118*(4), 1169–1208. https://doi.org/10.1162/003355303322552775

Bresnahan, T. F., Brynjolfsson, E., & Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics, 117*(1), 339–376. https://doi.org/10.1162/003355302753399526

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM, 36*(12), 66–77. https://doi.org/10.1145/163298.163309

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives, 14*(4), 23–48. https://doi.org/10.1257/jep.14.4.23

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics, 13*(1), 333–372. https://doi.org/10.1257/mac.20180386

Reference

Buchak, G., Matvos, G., Piskorski, T., & Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics, 130*(3), 453–483. https://doi.org/10.1016/j.jfineco.2018.03.011

David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review, 80*(2), 355–361.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science, 37*(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

DeYoung, R., Hunter, W. C., & Udell, G. F. (2004). The past, present, and probable future for community banks. *Journal of Financial Services Research, 25*(2), 85–133. https://doi.org/10.1023/B:FINA.0000020656.65653.79

Elliott, M., Golub, B., & Jackson, M. O. (2014). Financial networks and contagion. *American Economic Review, 104*(10), 3115–3153. https://doi.org/10.1257/aer.104.10.3115

Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2019). The role of technology in mortgage lending. *Review of Financial Studies, 32*(5), 1854–1899. https://doi.org/10.1093/rfs/hhz018

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies, 33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Jagtiani, J., & Lemieux, C. (2018). Do FinTech lenders penetrate areas that are underserved by traditional banks? *Journal of Economics and Business, 100*, 43–54. https://doi.org/10.1016/j.jeconbus.2018.03.001

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Petersen, M. A., & Rajan, R. G. (2002). Does distance still matter? The information revolution in small business lending. *Journal of Finance, 57*(6), 2533–2570. https://doi.org/10.1111/1540-6261.00505

Solow, R. M. (1987, July 12). We'd better watch out. *New York Times Book Review*.