Akholi

INTRODUCTION / 2026 RESEARCH STUDY

The Study

Agentic AI Outsourcing at Tier 1 Financial Institutions

Phil Hatch

150 engagements · 850 delivery-team respondents

Observational, cross-sectional study of a nonrepresentative sample. Financial institutions did not independently validate the reported experience. The findings do not establish causation or legal breach.

Version 1.8.2 | November 1, 2026

REPORT INFORMATION

Agentic AI Outsourcing at Tier 1 Financial Institutions

Full Study Report

Author	Phil Hatch
Publication date	November 1, 2026
Version	1.8.2
ISBN	979-8-234-26882-2
SSRN DOI	https://doi.org/10.2139/ssrn.7164559
Figshare DOI	https://doi.org/10.6084/m9.figshare.34036890
Contact	info@akholi.com
Website	akholi.com

AUTHOR ONLINE

Akholi author page | ORCID | LinkedIn | ResearchGate

SUGGESTED CITATION

Hatch, Phil. 2026. *Agentic AI Outsourcing at Tier 1 Financial Institutions: Full Study Report*. Version 1.8.2. Akholi. https://doi.org/10.2139/ssrn.7164559.

COPYRIGHT

© 2026 Phil Hatch and Akholi. All rights reserved.

No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the copyright holders, except for brief quotations with full attribution in reviews, research, and other noncommercial uses permitted by copyright law. Requests for permission should be sent to info@akholi.com.

USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence tools were used to prepare this report. Appendix G discloses how they were used.

Contents

Executive Summary **5**
Introduction **7**
Study Design, Evidentiary Basis, and Definitions **9**
1 Composition of the Study Set 11
2 Evidentiary Basis and Its Limits 17
3 The Ten-Point Rating Scale 19
4 Cancellation and Cancellation Risk 21
5 The Four Core Study Measures and Supplementary Measures 23
6 Manual Alteration as Defined and Counted 25
7 Statistical Methods and Reporting Conventions 29
8 Reading Conventions 33
About the Author **35**
About Akholi **37**

“

What we did, in the end, was stop waiting on their change board and put a standing Friday call with the credit desk instead, which is not governance in any formal sense, but it does at least get a decision on the week's flagged names before the book closes. We also keep our own mapping of counterparty identifiers, maintained by two of our analysts, because the definitions their source systems produce disagree between themselves and, in fairness to them, they never claimed they would. It works, more or less. I would not hold it up as a model, and it does depend rather heavily on two people staying.”

PORTFOLIO DIRECTOR

KPO, Credit, underwriting and suitability

Executive Summary

The surveyed engagements had a high cancellation rate and low reported customer satisfaction. By the study date, 56 of 150 engagements had been canceled (37.33%). Delivery teams rated cancellation as at least moderately likely for another 58 of the 94 active engagements. Together, 114 of 150 engagements (76.00%) had either been canceled or met the study's threshold for perceived cancellation risk. Vendor-reported customer satisfaction averaged 43.85 points on a 100-point scale across the 150 engagements.

Positive measured returns were reported for 17 of the 30 engagements that had reached Production, with a mean recorded return of 24.71% among those 17. The mean is the arithmetic average of the recorded values. These are reported return figures, not independently verified financial results. Only 84 of 782 respondents (10.74%) rated their engagement as achieving its stated key performance indicators, with a rating of 3 or below on the statement that it was not doing so.

Four findings warrant particular attention:

1. Manual alteration by the delivery team was reported on all 150 engagements when the five manual alteration methods were combined: manual alteration of financial-institution-provided data, manual alteration of financial institution processes, manual alteration of AI outputs, reporting alteration, and third-party work done by hand. Respondents described these interventions as necessary to sustain delivery. In the structured disclosure questions, discussion of data, process, or AI-output manual alteration with the financial institution was reported in 13 of 1,004 valid answers (1.29%).

2. Respondents reported that financial institution governance, oversight, and control arrangements were not being used as designed. The 19 selected governance indicators from financial institution readiness and co-management had a mean rating of 3.63 on the 10-point scale (n = 15,201 ratings). Here n denotes the number of ratings. Every engagement had at least one respondent reporting that financial institution governance, oversight, and control arrangements were not implemented as designed. This selected subset is distinct from the 47-indicator governance, oversight, and control score examined in Part 3.

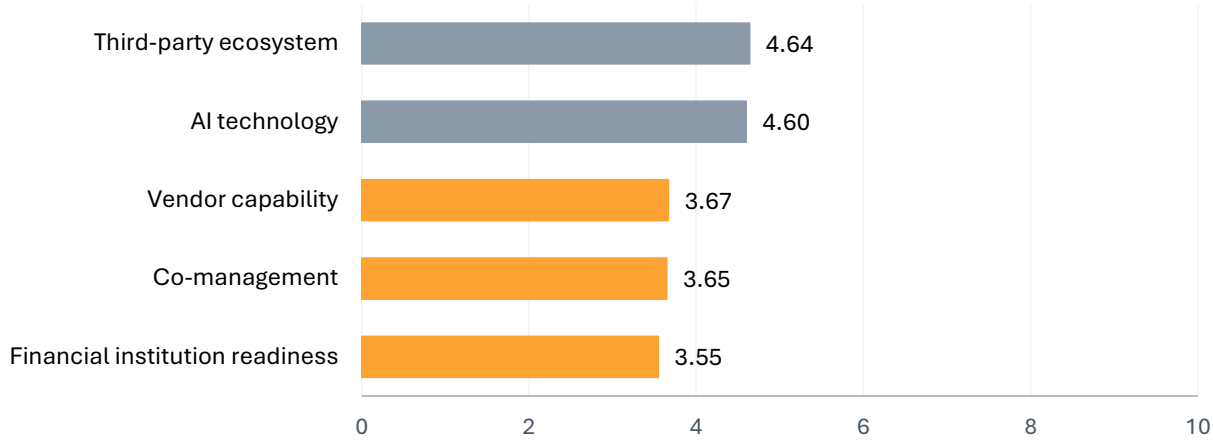
3. Experience with agentic AI was limited. No respondent reported more than one year of personal agentic AI experience, and no respondent reported that the vendor had more than two years of agentic AI delivery experience. Vendor delivery-playbook maturity was distressed in 392 of 807 ratings (48.57%). Respondents also reported limited agentic AI knowledge among financial institution staff, with 418 of 793 ratings classified as distressed (52.71%).

4. Respondents reported limits in the financial institution's willingness to redesign processes for agentic AI. Distressed ratings were given in 429 of 804 answers (53.36%).

The study organized 100 engagement indicators into five study-defined failure categories. Financial institution readiness had the lowest category mean (3.55), followed by co-management (3.65) and vendor capability (3.67). AI technology and the third-party ecosystem were rated more favorably, at 4.60 and 4.64, respectively, and both category means were below 5.0.

Figure 1. Mean indicator rating of the five failure categories on the 10-point scale, 150 engagements

Financial institution readiness, co-management, and vendor capability received lower mean ratings



Note: Each category mean averages the engagement-level means of its 20 indicators across 150 engagements. The amber bars identify organizational categories proposed as priorities for institutional review. The colors do not represent a measured degree of institutional control.

The study does not establish a regulatory or legal breach, and it contains no account from a financial institution. Its findings support the recommendation that financial institutions audit their agentic AI outsourcing engagements. Part 6 sets out the proposed response.

Introduction

Reported delivery difficulties, recurring manual alteration, and limited visibility into delivery work connect the engagement outcomes, governance concerns, and workforce experience examined in this study. Delivery teams described work sustained through human intervention alongside weaknesses in preparation and control. Their accounts require the reported performance of agentic AI outsourcing to be considered together with how that performance was achieved and what the two firms could observe. The report presents the structured results and written accounts from engagements at Tier 1 financial institutions, including null and conflicting findings.

The evidence is the reported experience of 850 members of vendor delivery teams on 150 agentic AI outsourcing engagements serving Tier 1 financial institutions, collected once, between January and June 2026.

The report is organized into six Parts.

01 Engagement Outcomes and Their Correlates

Engagement outcomes on the four core study measures, with mixed evidence on their relationships to structural attributes and careful distinction among supported, equivalent, and inconclusive results.

02 The Five Failure Categories

Financial institution readiness, co-management, vendor capability, AI technology, and the third-party ecosystem.

03 Governance, Oversight, and Control

Reported weaknesses in governance, oversight, and control, their relationship to engagement performance, and the operational significance of authority, review, and visibility.

04 The Delivery Workforce

The delivery workforce's experience, working conditions, intent to leave, and the continuity of knowledge used during delivery.

05 Manual Alteration by the Delivery Team

Manual alteration by the delivery team, the reasons given for it and its disclosure to the financial institution.

06 Executive Action Plan

Seven proposed controls and a phased implementation schedule for assessing engagements, correcting identified gaps, and verifying that controls operate effectively.

A Appendices

The definitions, the indicator inventory, the method, the list of companion analyses, and an index.

> *We know how to run the agents; fourteen months on this account and the exception queue has never been the thing that broke. What we do not have is anybody in the delivery line who has sat on a retail bank's servicing floor. Practically: I would take two people off the account for six weeks and put them through the client's own dormancy and deceased-account training, and I would stop hiring for the platform and start hiring for the domain. That is a staffing decision, not a training budget one, and it needs somebody above me to make it."*
>
> **PORTFOLIO DIRECTOR**
> BPO, Payments, cards and disputes

> *Problem the questionnaire cannot reach is why this is working, which is not process but people. Both sides put someone in the room who can actually say yes on a Tuesday afternoon without going back for approval, and in dispute work where Reg E clock is running that single fact is worth more than any governance document we have written. I have run engagements where the joint forum met fortnightly and everything of consequence happened between meetings. Here it is the reverse. My caution, and I would put it no stronger than caution, is that we are three months in with modest volumes, and arrangements of this kind tend to be tested at the first genuine backlog rather than in the quiet period."*
>
> **PORTFOLIO DIRECTOR**
> BPO, Payments, cards and disputes

Study Design, Evidentiary Basis, and Definitions

The study is a single cross-section of vendor delivery teams' reported experience on agentic AI outsourcing engagements serving Tier 1 financial institutions. Each respondent answered once, between January and June 2026, about one engagement. The design establishes what the delivery teams reported and how their reports related to their engagements' outcomes. It does not establish the order in which events occurred, what any financial institution knew, or how far the findings extend beyond the 150 engagements studied. This study did not approach any financial institution or outsourcing vendor to corroborate findings or obtain commentary.

The analysis distinguishes engagement-level outcomes from respondent-level ratings. It uses four core measures: cancellation rate, customer satisfaction, engagement indicator mean, and intent to leave. Manual alteration and its disclosure are measured separately. Statistical interpretation distinguishes supported associations, results sensitive to analytical choices, estimates within specified equivalence bounds, and inconclusive results.

The full method appears in Appendix D. The companion analysis for each indicator includes the full form of every test statistic summarized in the report. (Available online at Akholi.com)

“

The problem for my team is not that client is difficult, they are not. It is that same merchant category is set up one way in the North region and another way in the South, so my agents cannot trust one rule. i have two people who do nothing but check which version applies before a case is closed. That is two people i would rather put on the dispute queue. Access we get quickly, credit to them, and the core servicing process itself is well written. i just wish it was written once.”

ENGAGEMENT MANAGER

CXM, Payments, cards and disputes

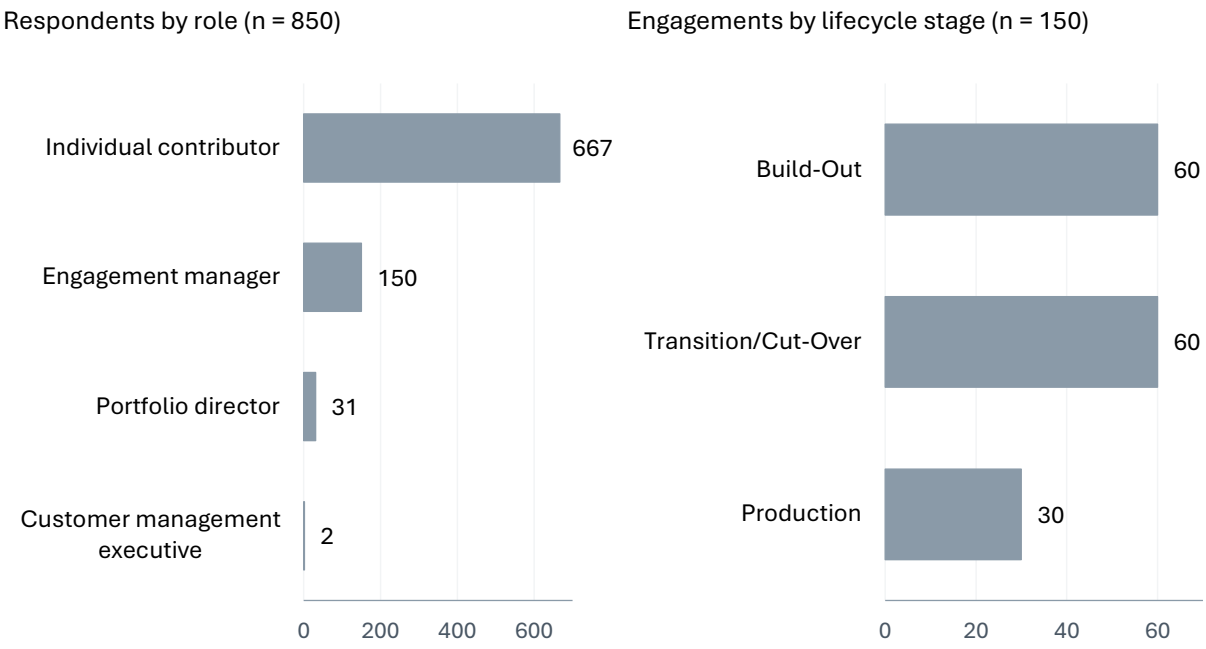
1 Composition of the Study Set

The study includes 150 engagements and 850 respondents. An engagement is one contracted outsourcing arrangement between a vendor and a financial institution. A respondent is a member of the vendor's delivery team assigned to that engagement. Engagement-level summaries give each engagement one observation. Respondent-level summaries include multiple observations from some engagements. The relevant inferential procedures and their treatment of clustering are specified in Appendix D.

The study set spans the engagement lifecycle stages, four outsourcing models, two general work classifications, eight use case families, and 10 types of financial institutions. Each composition is reported at its own base in sections 1.1 to 1.6.

Figure 2. Respondents by role and engagements by lifecycle stage

The sample includes 850 respondents across 150 engagements



Note: Respondents are counted at respondent level and engagements at engagement level. The panels use separate bases.

The panels use different units and bases. Role counts describe respondents. Lifecycle-stage counts describe engagements.

1.1 Respondents, Roles, and Engagements

The study covers 150 engagements and 850 respondents. Individual contributors, who perform the engagement's tasks, numbered 667 of the 850 respondents (78.47%). Engagement managers numbered 150 of the 850 respondents (17.65%), with exactly one per engagement. Portfolio directors, who oversee several engagements, numbered 31 of the 850 respondents (3.65%). Two respondents were customer management executives.

Sample size is reported for each analysis. Estimates based on fewer than 30 observations require particular caution. Descriptive subgroups with fewer than 10 observations are reported as counts rather than rates. These thresholds refer to the denominator, not the number of affirmative answers. The two customer management executives are reported only as counts. Small subsets of voluntary written answers are also identified explicitly and are not treated as prevalence estimates.

1.2 Structured Fields and Written Answers

The survey collected two forms of evidence. The structured answers total 149,619, given to 172 numeric questions and 23 categorical questions (195 total). The numeric questions include ratings on a 10-point scale and typed numbers such as months, team size, or contract term. The 100 rated engagement indicators, 20 in each of the five failure categories, form the core of the numeric questions. The categorical questions record attributes such as the type of financial institution, the outsourcing model, and the respondent's role, and the yes-or-no questions on manual alteration and its disclosure.

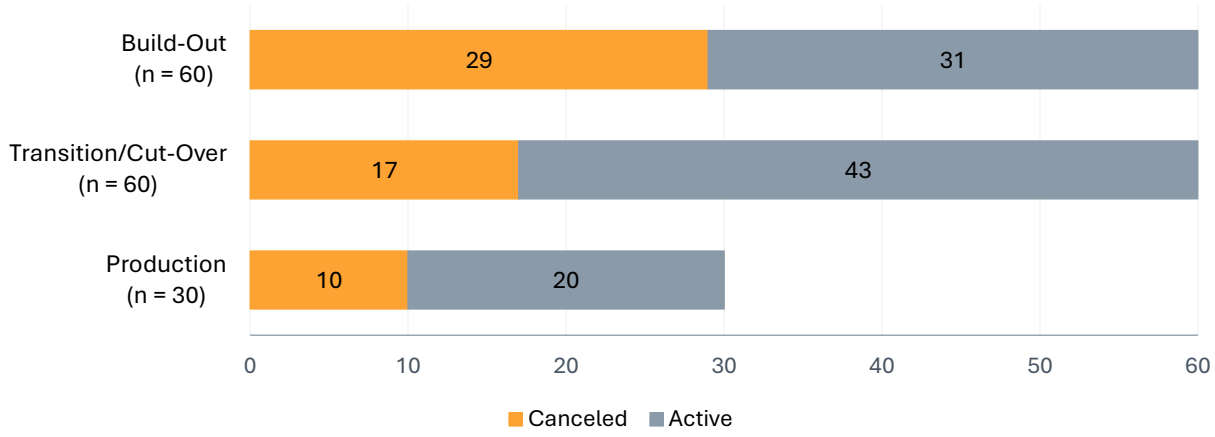
Written answers totaled 8,918 responses to the study's 16 open questions. Each substantive written response was assigned a sentiment score on a scale from −10 to +10.

Written answers were voluntary and describe the experiences respondents chose to report. Their frequency is a count of supplied and coded accounts, not an estimate of prevalence among all engagements or the wider industry. Structured questions provide separate estimates for the conditions or practices they explicitly measure. A broad rating does not measure the prevalence of every mechanism mentioned in comments.

1.3 Distribution of Engagements by Lifecycle Stage

The sample included 60 engagements in Build-Out, 60 in Transition/Cut-Over, and 30 in Production. By the study date, 56 had been canceled, and 94 remained active. Cancellation occurred at every stage. The chi-square test compares cancellation frequencies across the stages. The difference in cancellation rates across stages was not statistically significant (chi-square 5.39, p = 0.068, n = 150 engagements). The p value is the probability, under the null model, of a result at least as extreme as the observed result.

Figure 3. Engagements by status and lifecycle stage, 150 engagements



Note: canceled engagements: 56 (56 of 150 engagements, across all three stages).

Canceled engagements are counted across all three stages, 29 in Build-Out, 17 in Transition/Cut-Over and 10 in Production. The active engagements total 94.

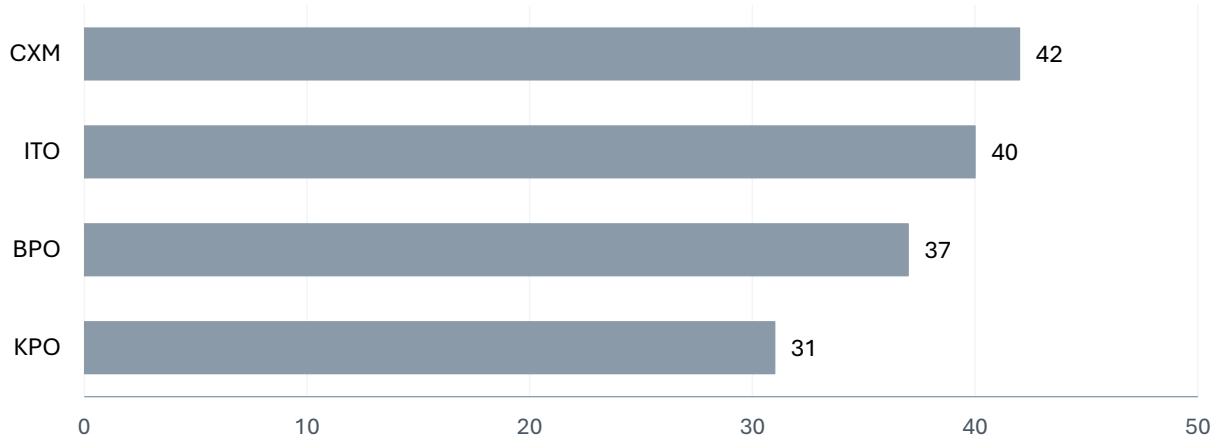
Counts are engagements. Cancellation rates were 48.33% in Build-Out, 28.33% in Transition/Cut-Over, and 33.33% in Production. Overall, 56 of 150 engagements were canceled (37.33%). The lifecycle stages are categories, not elapsed-time observations.

1.4 Distribution of Engagements by Outsourcing Model

The four outsourcing models were represented in similar numbers. Customer experience management (CXM) accounted for 42 engagements, information technology outsourcing (ITO) for 40, business process outsourcing (BPO) for 37, and knowledge process outsourcing (KPO) for 31. The two classes of work were nearly even, with 76 engagements in a line-of-business operation and 74 in a corporate function.

Figure 4. Engagements by outsourcing model, 150 engagements

The four outsourcing models were represented in similar numbers.



1.5 Distribution of Engagements by Institution Type

The engagements served 10 types of financial institutions, shown in Figure 5. Insurers were served by 14 engagements, and private equity and private capital firms by nine. The 23 engagements serving those two types fall outside the banking types. The report uses the term financial institution throughout.

Figure 5. Engagements by type of financial institution, 150 engagements

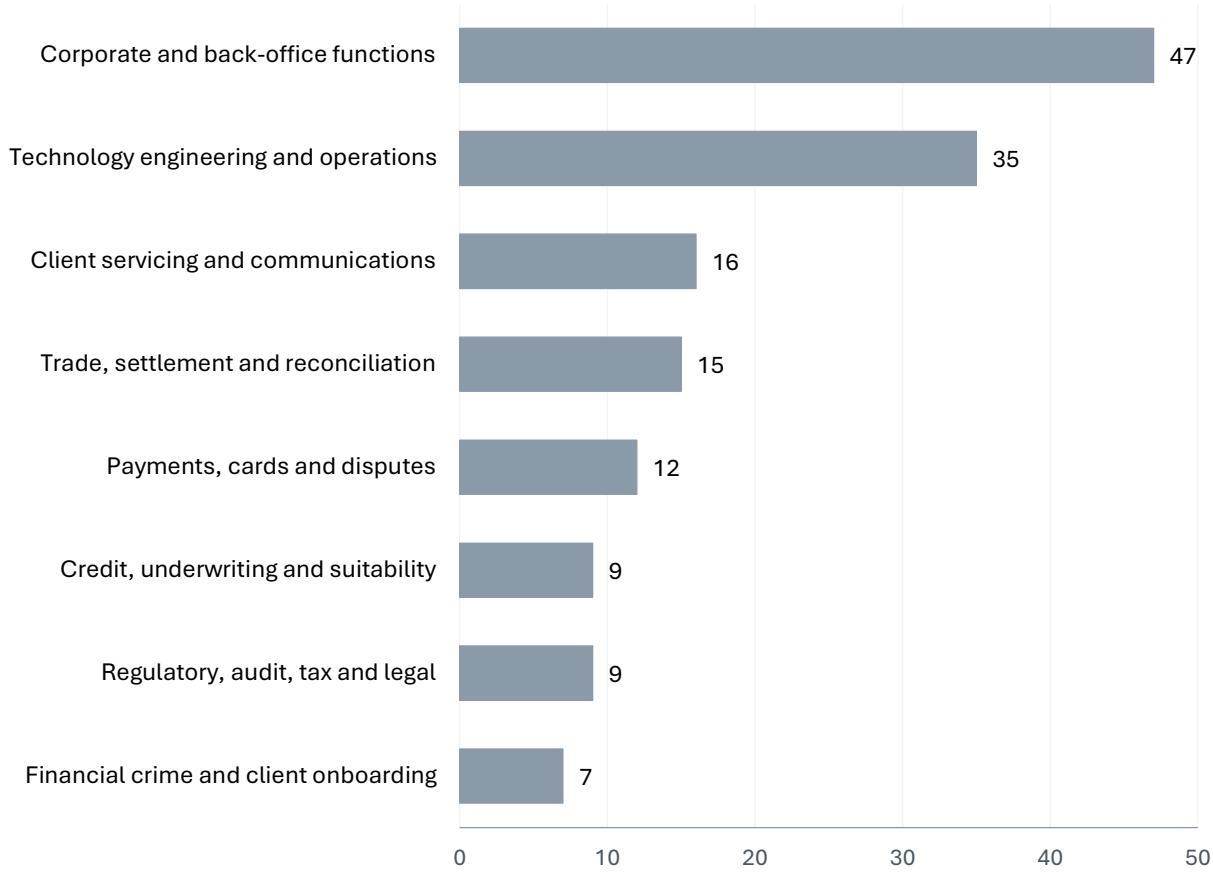
The engagements served ten types of financial institutions

1.6 Distribution of Engagements by Process Family

Each engagement's use case was assigned to one of eight process families, shown in Figure 6. Corporate and back-office functions accounted for 47 engagements and financial crime and client onboarding for seven.

Figure 6. Engagements by process family, 150 engagements

Corporate and back-office functions accounted for the largest process family



2 Evidentiary Basis and Its Limits

The evidence consists of vendor delivery personnel's reported experience. The study did not directly observe operations, inspect engagement records, or interview the financial institutions. Ratings, counts, and written responses must be interpreted within that evidentiary boundary.

The findings do not establish what financial institutions knew or received through channels outside the respondents' accounts. Each engagement was surveyed once. Selection sought variation in engagement types rather than an industry-representative sample.

2.1 Vendor Delivery Team Members as the Sole Source of Evidence

Every respondent was employed by the vendor delivering the engagement. The study did not approach or obtain commentary from any financial institution, outsourcing vendor, AI provider, or third-party ecosystem vendor. The study views each engagement from one side: the people who performed its tasks and managed its delivery.

The sample was selected to include a range of financial institution types, outsourcing models, and use cases. The resulting composition does not estimate their relative prevalence in the industry.

2.2 Anonymization of the Collected Data

The study was designed to limit the collection of data that identifies the financial institution, its customer, and the outsourcing vendor. A third-party audit of all data collected was performed to ensure no identifiable data was provided to Akholi or this report’s author. The reviewer conducting the audit made one alteration to the collected data to ensure anonymity: Financial institution names were replaced by “the customer” or “the desk”, and the outsourcing vendor’s name was replaced by “the vendor”. An additional third party audited the work to ensure no other alterations to any collected data occurred.

2.3 Financial Institution Conditions Reported at Second Hand

Financial institution readiness is described by 20 of the 100 rated indicators, covering its data, processes, staff, governance, and response speed. Each rating reflects the delivery team's account of the circumstances it worked within.

A reported absence of disclosure identifies the communication described by the respondent. It does not establish institutional ignorance: other personnel, records, or reporting channels may have conveyed the information.

2.4 Observational Reach of Each Respondent Role

Each role sees a different part of the engagement. Individual contributors see the data, the agents, and the daily tasks firsthand. Engagement managers see the engagement as a whole and the relationship with the financial institution. Portfolio directors see several engagements and the commercial terms behind them.

Respondents may have limited direct knowledge of indicators outside their operational remit. Individual contributors may lack visibility into management arrangements, while senior respondents may have limited or no direct knowledge of the delivery team's activities. Unless otherwise noted, this study did not filter responses to an indicator according to the respondent's role.

2.5 Standing of a Finding: Association Without Causal Direction

The primary customer-satisfaction analysis uses the 94 active engagements. This restriction excludes ratings recorded after cancellation but does not remove selection bias, confounding, or uncertainty about causal direction. Analyses of realized cancellation use all 150 engagements and are descriptive contrasts. Ratings from the 56 canceled engagements may reflect the cancellation itself. These contrasts are not forecasts of future cancellation.

Results are interpreted as associations. Statistical significance, sensitivity to alternative analyses, and equivalence within specified bounds are distinct properties of an estimate. None establishes causal direction in this cross-sectional design.

3 The Ten-Point Rating Scale

The questionnaire used 10-point ratings with several orientations. Some items assigned higher values to more favorable conditions. Others assigned higher values to adverse conditions. A further group had no favorable direction. The analysis preserves these distinctions and identifies transformations before interpreting scores.

3.1 Favorable-Direction and Unfavorable-Direction Indicators

In most cases, favorability increases across the ten-point scale, where 1 is the most unfavorable score, and 10 is the most favorable. Such indicators are labeled as “favorable-direction” indicators. For example, respondents rated the quality of the financial institution’s data across the systems the engagement depended on, where a higher rating denotes higher data quality.

Some indicators are inverse, with 1 most favorable and 10 least favorable. These indicators are referred to as “unfavorable-direction” throughout this study. For example, respondents rated the statement that the engagement was not achieving its stated key performance indicators, where a higher rating denotes the less favorable outcome. Where the report converts such ratings to the favorable direction, each rating is subtracted from 11, so that a higher figure denotes the more favorable outcome, and the conversion is stated beside the figure.

Eight rated questions have neither a favorable nor an unfavorable end. For example, they ask whether the financial institution is scaling its agentic AI outsourcing or shifting work to its own global capability centers. For these questions, the report states whether each association was positive or negative, and does not define a favorability score for the indicator itself.

3.2 The Distressed Threshold

The study defines a distressed rating as 3 or below on a favorable-direction indicator, and 8 or above on an unfavorable-direction indicator. The threshold is an analytical classification of adverse ratings. It does not establish that the respondent diagnosed operational failure.

The threshold also applies to engagements. An engagement is distressed on a question when the mean of its respondents' ratings falls in the distressed range. For example, in answers to the statement that the engagement was not achieving its stated key performance indicators, 241 of 782 answers were distressed (30.82%). The engagement mean was distressed on 32 of the 150 engagements.

Indicators distressed is the share of an engagement's distressed indicator ratings. It is computed for each respondent across the 100 rated indicators, averaged within the engagement and then across the 150 engagements. Across the study set, 42.49% of indicator ratings were distressed (n = 150 engagements).

3.3 Response Counts and the Base Convention

Percentages identify the relevant denominator: respondents, answers, or engagements. Percentages are generally reported to two decimal places, with small sample sizes stated explicitly.

The yes-or-no questions allowed “I do not know” and “I prefer not to answer.” Rates use valid yes-or-no answers only. For data alteration, 421 of 845 respondents selected yes and 418 selected no. The resulting rate is 421 of 839 valid answers (50.18%). The other six answers are excluded from that denominator.

For perceived factual attributes such as institutional asset size or contract term, “I do not know” is reported separately. It records the answer provided without explaining why the respondent could not or did not provide the information.

4 Cancellation and Cancellation Risk

Realized cancellation is the status recorded for each engagement in the study's Engagements sheet. It differs from a respondent's rating of future cancellation likelihood. The study recorded 56 canceled engagements among 150.

Perceived cancellation likelihood is a respondent rating, not an observed outcome or calibrated probability. Although the question was worded for active engagements, numeric answers were recorded on all 150 engagements. Only ratings from the 94 active engagements were used to define prospective risk bands. Full-sample correlations are identified as descriptive comparisons.

4.1 Cancellation Likelihood on Active Engagements

The cancellation-likelihood question asked about engagements not yet canceled. Nevertheless, its 328 numeric responses represented all 150 engagements, including canceled engagements. Another 269 respondents selected “I do not know.”

The rating records perceived likelihood and was not validated as a forecast. Responses from canceled engagements require retrospective interpretation and cannot be treated as observations made before cancellation.

4.2 High Risk, Moderate Risk, and the Canceled-or-at-Risk Population

All 94 active engagements were placed in a risk band based on the mean of their respondents' ratings of cancellation likelihood on the original scale. An engagement whose mean was 7 or above is high risk, and one whose mean was 5 or above and below 7 is moderate risk. By that rule, 19 active engagements were high risk, 39 moderate risk, and 36 below moderate risk.

The combined category comprises 56 canceled engagements and 58 active engagements classified as having moderate or high perceived cancellation risk. It contains 114 of 150 engagements (76.00%). The total combines realized status with a subjective risk classification and should not be interpreted as a realized or predicted cancellation rate.

“

I found this section difficult to answer genuinely, because it asks about joint decisions as though both firms are standing in the same room when something goes wrong. In a cardholder servicing operation the thing that matters is who holds the switch, and at two months in that is one named person on our side and nobody on theirs after six in the evening. A number between one and ten will not tell you that. Ask instead how long it takes to pull an agent off live cardholder traffic, and who is permitted to do it — that answer would have been thirty minutes on a weekday, and I would rather have given you it.”

PORTFOLIO DIRECTOR

CXM, Payments, cards and disputes

5 The Four Core Study Measures and Supplementary Measures

Four core study measures are applied to every indicator in the report, always in the same order. They are cancellation rate, customer satisfaction, engagement indicator mean, and intent to leave. Together they describe whether an engagement survived, the satisfaction recorded against it, the circumstances it was delivered under, and whether its delivery team intended to stay.

The four measures were strongly correlated across the 150 engagements, with every pair having an absolute correlation of at least 0.72. One principal component, a statistical combination that captures shared variation among the measures, accounted for 88.23% of their variance. This shared statistical structure does not make the constructs interchangeable. Two supplementary measures, recorded return and the share of distressed indicator ratings, are reported separately.

5.1 Engagement Indicator Mean

The engagement indicator mean summarizes the circumstances under which an engagement was delivered. It is computed in three steps.

1. Each respondent's ratings on the 100 rated indicators are averaged.

2. The respondent means within each engagement are then averaged.

3. The figure for the study set is the mean across the 150 engagements. Across the study set, it was 4.02 on the 10-point scale (n = 150 engagements).

Each indicator is also one of the 100 components of the mean. When an indicator is tested against the engagement indicator mean, the mean is recomputed without that indicator. The leave-one-out rule prevents an indicator from self-correlation.

5.2 Cancellation Rate

Cancellation rate is the proportion of a specified group of engagements recorded as canceled by the study date. Across the study sample, it was 56 of 150 engagements (37.33%). A respondent-level calculation would weight engagements by their number of sampled respondents, which is not necessarily proportional to delivery-team size. Such respondent-level comparisons are identified separately from engagement cancellation rates.

5.3 Customer Satisfaction

Customer satisfaction is the vendor-reported score recorded for each engagement on a 100-point scale. The source field is labeled “Customer Satisfaction %”. The values are treated here

as scale points, not as the percentage of customers who were satisfied. The mean was 43.85 points across 150 engagements and 64.10 points among the 94 active engagements. No financial institution record was available to verify these scores.

The primary estimate of an association with customer satisfaction uses active engagements. Scores for canceled engagements were recorded after cancellation and are used only in explicitly identified descriptive comparisons.

5.4 Intent to Leave

Intent to leave is the respondent's rating, from 1 to 10, of their likelihood of leaving the current job within 12 months. Higher original-scale scores indicate greater stated likelihood. The mean was 6.77 among 798 respondents. 430 of 798 (53.88%) rated their likelihood at 7 or above. This high-likelihood threshold differs from the study's distressed threshold of 8 or above for unfavorable-direction indicators.

Intent to leave is measured at the respondent level. Respondents on the same engagement share its circumstances, and tests against respondent-level questions use standard errors clustered by engagement. Where an engagement-level form is needed, the measure is the mean of its respondents' ratings.

5.5 Measured Return and Indicators Distressed

Measured return is the percentage recorded in the study's return field for the 30 engagements that had reached Production, as reported by the vendor’s delivery team. Positive values were reported for 17 engagements, averaging 24.71%. The other 13, including 10 canceled engagements, had no positive return recorded. The field does not establish a common investment denominator, measurement period, or gross-versus-net convention. The values should be interpreted as reported return percentages rather than standardized, independently verified returns on investment. No return is imputed for preproduction engagements.

6 Manual Alteration as Defined and Counted

Manual alteration denotes delivery personnel's reported changes to institutional data or processes, agent output, or performance reporting. Examples include correcting input records, supplying an unresolved process step, and revising output before submission. The study also records manual work performed to compensate for third-party system limitations.

The five indicator categories describe assessed engagement conditions. Manual-alteration measures describe reported actions taken during delivery. Their association does not establish whether conditions caused the actions, and alteration is not presumed improper merely because it occurred.

6.1 The Five Manual Alteration Methods

The study counts manual alteration by five methods, shown in Table 1. Three methods were counted by a direct yes-or-no question, each followed by a written question asking why. Respondents reported altering financial-institution-provided data in 421 of 839 answers (50.18%), the financial institution's processes in 490 of 841 (58.26%), and AI outputs in 278 of 828 (33.57%).

Table 1. The five manual alteration methods

Method	How it was counted	Respondents	Engagements
Manual alteration of financial-institution-provided data	Yes-or-no question, then a written question asking why	421 of 839	127
Manual alteration of financial institution processes	Yes-or-no question, then a written question asking why	490 of 841	132
Manual alteration of AI outputs	Yes-or-no question, then a written question asking why	278 of 828	108
Reporting alteration	Written answers on how the vendor altered the reports the financial institution received, other than “I do not know”	484 of 775	127
Third-party work done by hand	Two rated questions and the written answers describing the method	87	63

Reporting alteration concerns changes to performance reports and dashboards. Of 775 respondents answering how the vendor altered such reports, 484 gave an answer other than “I do not know” (62.45%), representing 127 engagements. This broad response count is distinguished from narrower substantive and indicator-coded subsets in Part 5.

Third-party work done by hand, the last of the five methods, is the delivery team performing by hand a task that a third-party system did not perform. Respondents described pulling records from a third-party portal that offered no interface to an agent and keying in data that a broken feed did not deliver. Others described reopening records that a change by the third party had stopped. The method was rated on two questions and described in 91 written answers from 87 respondents on 63 engagements.

6.2 Counting Conventions and Coverage

The coverage measures use overlapping definitions. Direct affirmative answers on data, process, or output alteration came from 619 respondents on 146 engagements. Adding reporting alteration and third-party manual work identified at least one reported manual alteration on all 150 engagements. The latter coverage remained after excluding answers that expressed no knowledge, denied alteration, supplied no substantive content, or criticized the survey.

The difference between 146 and 150 engagements reflects the additional measurement channels. It is not a discrepancy between estimates of the same practice.

6.3 Disclosure as a Separate Object

Disclosure is measured separately from intervention. Respondents reporting data, process, or output alteration were asked whether they had discussed it with vendor management and whether they or their management had discussed it with the financial institution.

Internal disclosure to the vendor's management was reported in 783 of 1,113 answers (70.35%). Financial institution disclosure was reported in 13 of 1,004 answers (1.29%). Both counts sum the answers to three questions, one for each method, and are counts of answers rather than of respondents.

“

The problem, nine months in, is that change control on funding-rule updates still routes through the client, and by the time it clears, the settlement window has already closed. I do not touch the reconciliation myself now, but the team overrides settled versus funded positions by hand most weeks, and none of that reaches the dashboard I report from, which sits uneasily against the change-control rules on paper.”

ENGAGEMENT MANAGER

BPO, Trade, settlement and reconciliation

7 Statistical Methods and Reporting Conventions

The companion analyses document the tests applied to each indicator, field, and derived measure. This report summarizes the estimates, relevant uncertainty, analysis units, and material qualifications. Appendix D describes the statistical procedures and the distinction between the original analysis and subsequent analytical or coding revisions. All companion analyses are available online at Akholi.com.

At n = 150, the reported minimum detectable correlation was approximately 0.23 for a two-sided 5% significance level and 80% power. Equivalence testing used a separate bound of ±0.30. Meeting that criterion supports an estimate within the study's chosen bounds. It does not establish literal zero or prove that every smaller association is operationally unimportant.

7.1 Primary Estimates, Descriptive Contrasts, and Partial Tests

For numeric indicators, the principal analyses report Pearson correlations with the four core measures. Pearson correlation describes the strength and direction of a linear association. Cancellation contrasts use 150 engagements. Primary customer-satisfaction estimates use the 94 active engagements. The engagement indicator mean is recomputed without the indicator under examination. Intent-to-leave analyses use respondent observations with engagement-clustered inference where specified. The companion record identifies each estimate's sample and interval method.

A partial test asks whether an indicator was associated with a measure beyond what the engagement indicator mean already accounts for. Each of the 100 rated indicators was tested in this way against cancellation, customer satisfaction, and intent to leave (300 tests in total) with the false discovery rate controlled across the full set. This procedure limits the expected proportion of false discoveries among results declared statistically significant under its assumptions. Comparisons across segments of the study set were corrected for the number of comparisons by the Holm method, which limits the probability of at least one false-positive conclusion within the specified family of comparisons under its assumptions.

7.2 The Four States of a Reported Result

An association is described as supported when it meets the relevant significance criterion and remains consistent under the applicable multiplicity adjustment and sensitivity checks. The rated-indicator protocol includes 22 checks distributed across the four outcomes. It does not apply 22 separate checks to every association. Appendix D identifies their scope.

A result is qualified when it fails an applicable adjustment or sensitivity check. Equivalence means that the tested effect meets the specified equivalence criterion, rather than being literally zero. A statistically significant effect can also fall within the equivalence bounds. A

nonsignificant result that does not meet those bounds remains inconclusive. These distinctions prevent failure to detect an association from being interpreted as proof of absence.

7.3 Tests for Unordered Attributes

Attributes without an order, such as the outsourcing model or the type of financial institution, were tested by comparing the levels of the attribute. Cancellation, a yes-or-no outcome, was compared across levels with the chi-square test, or with Fisher's exact test or a Monte Carlo test where an expected count was under five. The other measures were compared with Welch's analysis of variance. That test does not assume equal spread across levels.

Engagement-level attribute comparisons use one assigned value per engagement and exclude response levels with fewer than 10 eligible engagements. Separate indicator analyses also compare respondents by role and vendor type. Those analyses use different units and should not be interpreted as the same tests.

7.4 Engagement Indicator Quintiles

The indicator quintiles divide the 150 engagements into five bands of 30, ranked by engagement indicator mean, with ties broken by engagement identifier. The lowest indicator quintile contains the 30 engagements with the lowest engagement indicator means, and the highest contains the 30 with the highest means. The indicator quintiles allow every study measure to be compared across the full range of these reported assessments in five steps. Because the bands are built on the engagement indicator mean, the rise of that mean from band to band is true by construction and is not reported as a finding.

7.5 Reporting of Null Results

Null and inconclusive results are reported alongside supported associations. For several structural attributes, estimated relationships met the study's equivalence criteria. Those results constrain the size of the tested associations under the specified models. They do not eliminate the attributes as possible contributors through other mechanisms or in other populations.

Restricted variation limits what an analysis can detect. For example, all 571 respondents answering the vendor-experience item reported zero, one, or two years. That range restricts conclusions about more experienced vendors, but it does not mathematically require a zero association within the observed sample.

7.6 The Written Answers as Evidence

Each substantive written response is assigned a sentiment score from −10 to +10, in five fixed steps. A score of −10 marks a part of the engagement described at or near its worst, and −5 a

clear but partial problem. A score of 0 marks no position, or good and bad in balance. A score of +5 marks a clear but partial strength, and +10 marks at or near its best. Each response is also identified by the indicators it addresses and by the recurring themes it contains.

A response is primary when it answers the indicator's own open question and secondary when it addresses the indicator in answer to another question. A respondent asked for comment on the financial institution may describe its data, its processes, and its speed of response in one answer, and that answer is evidence for each of the indicators it describes. When written responses provide direct contribution to the affiliated indicator, the response is marked as a primary theme contribution. Additional indicators that a written response references are marked as secondary theme contributions.

The report quotes written responses as the respondents wrote them, spelling and grammar unchanged, with words omitted only where an ellipsis marks the omission and the omission is immaterial. Names of financial institutions and vendors were replaced by a third party before the answers reached the author, as set out in Appendix D, section 10. Each quotation is attributed by role, outsourcing model, and process family.

“

I feel we are being judged on something nobody in this market has finished learning yet. Our people can read a medical file and see what matters in it, that part is solid, and the staff they put on this desk inside three weeks is more than I expect. But the agent side we are building while we fly. There is nobody in the firm I can ring at ten at night when the summariser produces a different risk note on the same chart, and the QA method we use is still the one written for human summarisers. So I score those lower. It is not that leadership is hiding from it. But wanting the capability and having it are two different things, and right now we are somewhere in between.”

ENGAGEMENT MANAGER

KPO, Credit, underwriting and suitability

8 Reading Conventions

Readers may consult individual indicators independently. Local sections retain the sample sizes and qualifications necessary to interpret their findings, while shared definitions and methods are consolidated in this Introduction and Appendix D.

8.1 Provenance of Every Figure

The numerical findings derive from the structured responses, coded written answers, and engagement records. The online study-data index links to the corresponding companion analyses, identified by subject and survey-item code. Appendix E provides the access location and explains the supporting materials. Repeated tables of diagnostics are retained in those analyses rather than reproduced throughout the narrative.

8.2 Relationship to the Companion Indicator Analyses

The 216 published companion analyses provide distributions, segment comparisons, sensitivity checks, equivalence tests, and the relevant coded written evidence. Each report section identifies the indicator by its subject heading. The companion record supplies the detailed calculations. The main text retains qualifications that materially affect interpretation.

8.3 Use of a Single Block in Isolation

The Introduction defines distressed ratings, favorable-direction scores, the four core measures, and the distinction between active-engagement estimates and cancellation contrasts. Appendix D supplies the analytical methods, and Appendix E identifies the online companion index. These references support detailed consultation without repeating the full method under every indicator.

“

It is disappointing for us, because the records themselves are not the problem. But the client wants the agent to produce the summary in exactly the same shape their underwriters have always done it, box for box, including the sections nobody reads. We ask twice about redesigning the template for a machine and each time we are told to keep it as is. And that one decision costs us more rework than anything else in month one.”

ENGAGEMENT MANAGER

KPO, Credit, underwriting and suitability

About the Author

Phil Hatch

Outsourcing and workforce development

Akholi author page

LinkedIn

ORCID 0000-0003-1709-8860

SSRN research paper

Phil Hatch is an outsourcing and workforce-development executive with more than 30 years of experience. He previously led the global outsourcing business of ManpowerGroup/Experis. His work includes service delivery, workforce strategy, and organizational development across industries and countries.

Hatch has also advised governments on labor and economic development, including workforce strategies intended to support investment and employment.

Following Russia's 2022 invasion of Ukraine, Hatch was asked by the U.S. Department of State and USAID to advise Ukraine on workforce and economic development. His work has addressed technology, outsourcing, skills development, and international investment in support of economic recovery.

Through Akholi, Hatch advises organizations and governments on outsourcing and the implications of AI for work and service delivery.

“

The vendor's change log recorded release counts and cycle time but not which releases were later rolled back for validation failures, so the reported release velocity looked strong right up until the engagement was terminated.”

INDIVIDUAL CONTRIBUTOR

ITO, Trade, settlement and reconciliation

About Akholi

Akholi, established in 2003, is a research and advisory firm focused on outsourcing, service delivery, and global capability centers. Its work includes the 2026 study of 150 agentic AI outsourcing engagements described in this report.

The study examines relationships among financial institution readiness, co-management, vendor capability, AI technology, the third-party ecosystem, and reported engagement outcomes. Its findings are observational and do not establish that any one category is the sole or primary cause of those outcomes.

Akholi applies this research through advisory and educational work with institutional leaders.

Akholi

INTRODUCTION / 2026 RESEARCH STUDY

The Study

The Introduction establishes the study design, evidentiary limits, definitions, four core study measures, rating conventions, statistical methods, and organization of the report.

Phil Hatch | Akholi

Study ISBN 979-8-234-26882-2
SSRN DOI: https://doi.org/10.2139/ssrn.7164559
Figshare DOI: https://doi.org/10.6084/m9.figshare.34036890

Full study and supporting data

akholi.com/outsourcing-2026-agentic-ai-tier-1-banks

Version 1.8.2 | November 1, 2026

The evidence reflects respondents' reported experience. It is not an independent finding of regulatory breach.