Citi Wholesale Banking

The Delivery Mandate — a Target Operating Model, Demonstrated

One sentence of mandate, then the three artifacts its loop produces — populated for the Q2 2026 cycle · a discussion starter

Context. This document defines the machine — and then shows it running. A target operating model argued in the abstract is a philosophy; one demonstrated through its artifacts can be corrected, which is worth more. So after one sentence of mandate, everything here is output: the **delivery radar** populated with eight dated signals, the **initiative portfolio** actually sorted with one charter rendered end-to-end, and the **scorecard** at the Q2 2026 close with its quarterly reconciliation shell. The machine's two sibling instruments — the capacity map (where agentic AI creates capacity) and the operating-leverage review (where it banks) — run alongside; each quarter the loop turns, the instruments refresh, and the reconciliation fills.

Reading convention. Values marked ° are illustrative placeholders shaped like the real artifact, for the inside view to overwrite; unmarked figures are from Citi's public record. No new sizing appears anywhere in this document — gross opportunity sizing lives in the capacity map, economics in the operating-leverage review. The mandate-definition discipline was field-tested at a AAA-rated multilateral development bank across 57 member-country programs; it is re-weighted here for a delivery organization inside a bank, and offered — like everything in the system — as a reference to react against.

1. The mandate — one sentence, then the evidence

> **"As the delivery engine of the transformation, we convert AI capacity into the outcomes the firm has committed to — by running three disciplines that keep delivery FIT for purpose: Futures — sense. Initiatives — shape. Targets — bank."**

"Convert," not "align" — a delivery organization under an operator CFO is judged on what banks, and the verb carries that accountability. A mandate is a promise, and promises are judged by what they produce; the rest of this document is what this one produces. The loop runs at two speeds: the existing weekly review of 100+ processes is the heartbeat — the machine wraps it, it does not replace it — and the quarter closes with a reconciliation. Next turn: after Q3 earnings, mid-October.

2. Artifact one — the delivery radar, Q2 2026

Futures, as a log rather than an essay. Every entry must land on a flow and a date with an action — or it does not make the radar. The lens grid is 3×3 (Impact / Results / Resources rows); sensing takes ~5% of the engine's effort, and this table is its entire output, refreshed quarterly into the weekly review.

Lens	Signal — Q2 2026	So-what — the dated decision	Flow · when
Model capability	Agent-assisted investigation crosses the reliability bar for draft-and-approve triage in internal and vendor evals	Charter cross-border exception triage now, at tier-2 autonomy (human sign-off)	Cross-border exceptions · Q3 — charter now
Regulatory horizon	MRM evidence expectations for tier-2 autonomy stabilizing; consent-order lineage work overlaps the same data foundations	Pre-agree the evidence-pack template with Model Risk before the next three charters need it	All regulated flows · Q3 — charter now
Data readiness	Entity golden-source coverage reaches pilot threshold in two client segments	Sequence the perpetual-KYC pilot behind the exceptions charter — same pod pattern	Client lifecycle · Q4 — prepare

Lens	Signal — Q2 2026	So-what — the dated decision	Flow · when
Process shifts	Citi Payments Express ramping; instant rails shrink tolerance for exception latency	Capacity plan for 24/7 exception queues before volumes force it	Payments ops · Q4 — prepare
Peer moves	Peer firmwide assistants at near-total adoption; the Street begins asking for productivity disclosure, not tooling counts	Have the productivity-led leverage share measured before external questions arrive	Scorecard · Q4 — prepare
Client expectations	Corporate treasurers increasingly price payment certainty; exception latency appears in RFP scoring	Fold exception-aging targets into cross-border client SLAs at the next renewal cycle	TTS cross-border · 2027 — watch
Vendor & compute	Inference cost per case falling quarter on quarter; the build-vs-BPO boundary is moving	Re-price the reconciliations BPO contract against an agentic build before renewal	Reconciliations · 2027 — watch
Talent & skills	Eval-engineering is the scarce skill; each pod needs one and market supply is thin	Grow eval engineers internally through the pattern library rather than hire against the market	Spine & pods · Q3 — prepare

Signals drawn from the public record; the first weekly review overwrites this table with the inside view. An entry with no flow, no date, or no action is commentary, and commentary does not make the radar.

3. Artifact two — the portfolio, sorted; one charter, filled in

Initiatives, as a sorted table rather than a method. The 100+ mapped processes are the raw list; the selection triangle — cost concentration × AI feasibility × control payoff — sorts it. Twelve rows shown; scoring is illustrative° on published criteria, so a disagreement becomes an argument about evidence rather than taste. Gross sizing for these flows lives in the capacity map and is not re-argued here.

Flow	Cost conc.	AI feasibility	Control payoff	Tier
Cross-border payment exceptions	●●●	●●●	●●●	**Lighthouse — chartered**
SDLC & legacy modernization	●●●	●●●	●●○	**Lighthouse — scaling**
KYC periodic refresh	●●●	●●○	●●●	Pipeline · Q4 charter
Sanctions screening false positives	●●●	●●○	●●●	Pipeline · behind exceptions
Reconciliations & breaks — Markets	●●●	●●○	●●○	Pipeline
Control testing & evidence generation	●●○	●●○	●●●	Pipeline · Q1 2027
Client onboarding narratives	●●○	●●●	●●○	Pipeline
Corporate-actions processing	●●○	●●○	●●○	Watch
Servicing inquiries — middle office	●●○	●●●	●○○	Watch
Credit-memo drafting — Banking	●●○	●●●	●○○	Watch
Regulatory-reporting lineage	●●●	●○○	●●●	Watch · data dependency
Collateral management	●●○	●○○	●●○	Watch

●●● high · ●●○ medium · ●○○ low. Three-of-three makes a lighthouse; two, pipeline; the rest, watch list.

The charter — what the top of the portfolio looks like signed. No charter, no build. The document below is the artifact the pod and the business would actually sign; in the real version the owner is a name, not a function.

CHARTER № 2026-07	Cross-border payment exceptions · autonomy tier 2 — draft-and-approve, human sign-off on every action
1 · Bottleneck map	Where the handle time sits°: intake & classification 15% · context gathering 35% · repair drafting & rationale 30% · approval & release 12% · client communication 8%. The agent targets the middle 65% — context and drafting. Baseline°: ~45 minutes and ~$28 per case; aging tail beyond 48h on the hardest corridor.
2 · The bridge — dated	Q3 2026: assisted triage live in one corridor pod, ~20% of flow volume; measure handle time, override rate, aging. Q4 2026: scale to ~60% of volume; target cost per case −40%°; first freed-hours log. Q1 2027: redeploy-or-release decision on freed hours, finance co-signed; banked run-rate recorded as a share of the capacity map's payments row — not an addition to it.
3 · Evidence plan	Agreed with Model Risk up front: eval suite on historical cases before go-live · error taxonomy with severity tiers · human-override log reviewed weekly · tested rollback path · tier-3 autonomy earned only against this record.
4 · Pattern & reuse	Reuses the agent-assisted investigation pattern and the eval harness proven on the SDLC lighthouse; contributes the corridor-triage variant back to the library for screening and reconciliations to inherit.
Signatures	Business owner: Services — cross-border operations (named in the real version) · Delivery: Payments pod · Bridge co-sign: Finance · Autonomy tier: Model Risk.

4. Artifact three — the scorecard at Q2 close, and the reconciliation shell

Targets, as a filled table rather than a taxonomy. Three altitudes, one instrument; the full economic design lives in the operating-leverage review — this is it, installed. An honest quarter-close for a machine proposed in July shows exactly this: public figures where they exist, baselines being set where they do not, and first entries dated to Q3.

Altitude	Metric	Q2 2026 close	Status
Impact	Efficiency ratio → the <55% bar	57.4% — record-revenue quarter	on track · cyclical
	Operating leverage — productivity-led share	>9 pts total; split not yet measured	**baseline being set**
	RoTCE bridge contribution	13.0% firmwide; the engine's line starts at zero	first entry Q3
Results	Cost per cross-border exception	~$28° · target −40% by Q4	**baseline being set**
	STP rate by flow	Baselining across corridors	**baseline being set**
	Onboarding time-to-revenue	~4–6 weeks° · pilot targets days	charter Q4
	Hours redeployed vs. released	~100K dev-hrs/wk capacity exists; routing unlogged	first log Q3
Resources	Investment drawn vs. ~$5B envelope	On plan per Investor Day pacing	on track
	Adoption — in-workflow daily use	~9 in 10 have access; daily-use baseline TBD	**baseline being set**
	Model-risk approval cycle time	SLA to be agreed with the evidence-pack template	SLA set Q3
	Severance conversion — backfill rate	>$800M YTD taken; the structural	tracking from Q3

Altitude	Metric	Q2 2026 close	Status
		test is the backfill rate	

The quarter-close reconciliation — the shell the Q3 close will fill. Capacity claimed this cycle: ~100K developer-hours per week (disclosed) ≈ 2,500 FTE-equivalents. Routed — redeployed into the change backlog (operating leverage): destination log starts Q3. Routed — released as avoided backfill and vendor reduction (expense): finance co-sign starts Q3. Banked vs. bridge: first entries land with charter № 2026-07. Claims retired — capacity asserted but never routed: the honesty line, starts Q3.

> **Why a shell is the honest artifact.** A machine proposed in July cannot show a July close it never ran — and inventing one would break the trust every other number here depends on. What it can show is the exact table the Q3 close will fill. That is the difference between a promise and a template.

5. The structure behind the artifacts

A small spine owns what must be consistent everywhere — the radar, the pattern library and its evals, the scorecard and evidence packs. Delivery lives in pods embedded in the flows; pods reuse patterns, the spine hardens what pods prove, and the engine posts its own unit costs on the same scorecard it holds others to. The seam with the businesses is explicit:

Decision	Who holds it
Select the flows	Joint — spine + business, on the triangle's published criteria; disagreements logged, not smoothed
Charter an initiative	The pod, with the named business owner — no charter without both signatures
Set the autonomy tier	Model Risk, against a pre-agreed evidence bar; approval cycle time a managed SLA on the scorecard
Sign the capacity→P&L bridge	Finance — the co-sign is what makes a claim bankable
Own the flow and its clients	The business — always; the engine optimizes, it never expropriates
Own patterns, evals, the scorecard	The spine — one version of the truth, reused everywhere

Every cross-seam decision is logged in one of three states — **acceptance, rejection, resistance** — the third recorded, not euphemized, because resistance mapped early is delivery risk priced early. And the sunset test is stated up front: when the unit-cost curve holds for four consecutive quarters without central intervention, the spine shrinks and the pods dissolve into the line. An engine that plans its own obsolescence stays honest about why it exists.

6. Installing the machine — a first-90-days frame

Weeks 1–4 — adopt, don't add. Wrap the weekly review unchanged; overwrite the radar's eight sample signals with the real ones; re-sort the portfolio with real scores on the published criteria and retrofit two in-flight initiatives into charters. If this feels like new bureaucracy in month one, it is being installed wrong.

Weeks 5–8 — run one full loop. Ship charter № 1 — the exceptions corridor, or whatever the re-sort puts on top; log the first redeploy-or-release decisions with finance at the table; agree evidence pack v1 with Model Risk so the next three charters inherit it.

Weeks 9–12 — first reconciliation. Close the quarter on the shell — capacity claimed, routed, banked, retired; agree the decision-rights memo across the seam in writing; set the 2027 rhythm. From here the machine is not a proposal — it is how the quarter runs.

Illustrative by design — the real sequence and the real values belong to you and the team, and much of this is likely already in flight.

7. The questions this demonstration should be asked

The system rule: a serious review should attack this. The expected challenges, answered plainly — the first two conceded before they are asked.

“Half these cells are sample values — this is theater.”

Conceded, and flagged cell by cell with the ° convention. Public figures are public; everything else is a placeholder shaped exactly like the real artifact — a radar entry with a flow and a date, a charter with a bridge and signatures, a reconciliation with an honesty line. The alternative was another framework diagram. A demonstration that refuses to show filled-in artifacts is a philosophy; this one can be overwritten by the first weekly review, which is the point.

“An operating model proposed from outside is presumptuous.”

Conceded, fully. That is why every artifact is a reference to react against — no org charts, no names, no headcounts, no new sizing. Organizations converge fastest on a real operating model when a concrete, wrong-in-detail one is on the table; this is that artifact, and it expects to be corrected.

“A weekly review of 100+ processes already runs — this exists.”

Partly true, and the design leans on it: the weekly review is the heartbeat. What the machine adds is visible in the artifacts — a radar in front of the list so it points at where capability and regulation will be; charters that make selection criteria and bridges explicit; and a reconciliation behind it that turns claims into P&L-grade records. A list plus a meeting becomes a portfolio when those three artifacts exist.

“Central transformation offices calcify.”

They do — when the center hoards capability and permanence becomes the goal. Hence a spine deliberately small, delivery embedded in pods inside the flows, and a sunset test stated up front: four consecutive quarters of the unit-cost curve holding without central intervention, and the engine dissolves into the line.

“FIT comes from a strategy department at a development bank — a different world.”

The provenance is both the point and the risk, so it is stated rather than hidden. What transfers is the artifact discipline itself — a mandate in one sentence, a dated radar, chartered initiatives, a dashboard reconciling promises with results — proven while organizing a 57-country institution. What is re-weighted is emphasis: sensing shrinks to one table, delivery and banking grow. The method survives translation because it is about promises and reconciliation, not sector.

“Where do consent orders and Model Risk sit in this?”

Visibly inside the artifacts: the radar carries the regulatory horizon as a first-class lens with a dated action; the charter carries the evidence plan and an autonomy tier earned against a bar Model Risk defines; the scorecard carries approval cycle time as a managed SLA. The data foundations remediation requires are the same ones these use cases need — complementary programs, as the capacity map argues.

Where outside help fits

The artifact discipline in this demonstration is not borrowed: the same machine — one-sentence mandate, dated radar, chartered initiatives, a dashboard reconciling promises with results — was built and run at a AAA-rated multilateral development bank, organizing a strategy-and-transformation function across 57 country programs. Pairing that craft with hands-on agentic-AI delivery (an agentic platform built end to end; AI-augmented SDLC workflows designed for one of the largest U.S. banks; an LLM customer-experience analytics solution on Google Vertex AI) and bank-regulatory grounding (transformation work at the U.S. OCC and SEC; a certified-banker credential) is the specific combination an engine like this needs at stand-up. The offer is unchanged across the system: react against these artifacts in a working session — hours, not weeks — and keep whatever survives contact with the inside view.

Basis: an operating-model demonstration built on the public record. Values marked ° are illustrative placeholders for the inside view to overwrite; unmarked figures are from Citi's public disclosures (Q2 2026 results and earnings-call remarks, Investor Day 2026 materials, press releases). The loop's Q2 2026 instruments — "Citi — Wholesale AI & TOM Opportunity Map" (the capacity map) and "Citi — Business Transformation & Operating-Leverage Map" (the operating-leverage review) — carry all sizing and economics; no new sizing is introduced here. The method's origin institution is referenced without attribution by design. Not official Citi material.