PUBLIC-FACING PDF | REVIEW BEFORE PRODUCTION | FOLIUM SYSTEMS

RUNTIME CAPACITY ENGINEERING

AI Runtime And Capacity Engineering

AI becomes fragile when every task is forced through the same runtime. Folium maps each workload by privacy, cost, latency, source truth, support, availability, and business risk, then chooses the right route: cloud API, private endpoint, local model, container service, GPU lane, CPU lane, retrieval-first path, database-backed workflow, or fallback route.

AUDIENCE	PURPOSE	UPDATED
Executives, technical buyers, operations leaders, AI owners, cost reviewers, and security reviewers	Show how Folium makes AI runtime placement an operating decision instead of a vendor default	May 2026

Runtime placement should be chosen by workload class, data boundary, cost, latency, fallback, portability, FinOps, and support.

Local, cloud, private, hybrid, GPU, CPU, retrieval, and database routes can coexist under one control model.

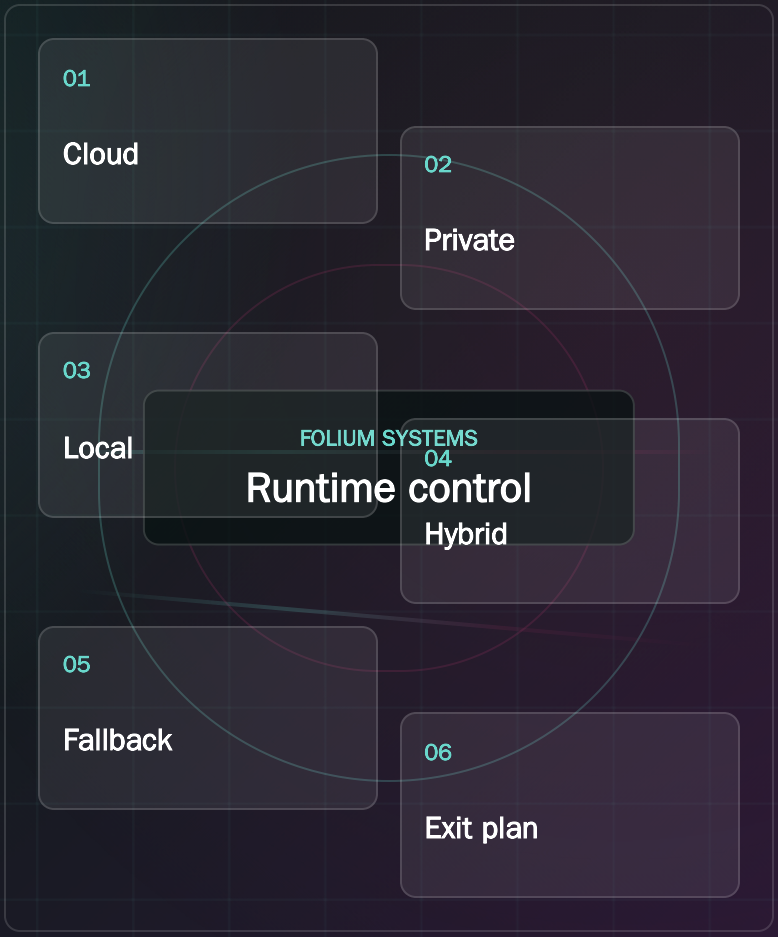
Capacity, token-budget, cache, quota, provider-spend, and saturation signals tell the business when to promote, park, split, fail over, or re-architect a workload.

RUNTIME PLACEMENT

Local, private, and hybrid AI choices should be business decisions.

The runtime guide compares cloud APIs, private endpoints, local models, hybrid routes, RAG placement, agent lanes, costs, latency, custody, and fallback.

RUNTIME CONTROL



01
Shows Folium can reason beyond one provider.

02
Makes privacy, cost, latency, and control tradeoffs visible.

03
Protects data custody before architecture locks in.

R

NAVIGATION MAP

Choose the review route before reading cover to cover.

This packet is meant to support a real decision meeting. Different reviewers should enter through different routes, then come back together around the same controlled next step.

DECISION ROUTE | OPERATING ROUTE | TRUST ROUTE

EXECUTIVE ROUTE

Decision first

Start with the cover, visual summary, executive read, controls, first ninety days, and handoff. This route helps leaders decide whether the next move is education, audit, first build, pilot, or operations.

- Outcome
- Risk
- Owner
- Next gate

OPERATIONS ROUTE

How the work will run

Read the workflow map, procedures, operating roles, metrics, first sprint, and buyer worksheet. This route shows whether staff can actually use, review, and improve the future process.

- Workflow
- Staff
- Support
- Improve

TECHNICAL AND TRUST ROUTE

Where the boundaries live

Focus on records and work products, controls, risk assumptions, reference work products, source truth, runtime placement, and launch conditions before any private access expands.

- Source
- Access
- Runtime
- Rollback

BUYER SESSION ROUTE

Turn reading into a working session

Use the discovery questions, role review route, buyer worksheet, and engagement fit ladder to prepare one process, one owner, one source map, and one next decision.

- Process
- Examples
- Questions
- Decision

> Best use: bring one workflow, the people who own it, the systems it touches, the data classes involved, and the decision this packet should help leadership make.

01

EXECUTIVE READ

Runtime capacity engineering in plain language.

AI becomes fragile when every task is forced through the same runtime. Folium maps each workload by privacy, cost, latency, source truth, support, availability, and business risk, then chooses the right route: cloud API, private endpoint, local model, container service, GPU lane, CPU lane, retrieval-first path, database-backed workflow, or fallback route.

RECORD | BOUNDARY | ACTION

CLASSIFY

Know the workload first

Separate private, public, retrieval-heavy, high-speed, lightweight, batch, customer-facing, and state-changing work.

- Data
- Speed
- Authority

PLACE

Choose the right runtime

Use cloud, private, local, GPU, CPU, container, edge, database, or hybrid placement by evidence.

- Cloud
- Local
- Hybrid

FALLBACK

Design degraded mode

Plan what happens when a provider, model, queue, source, or runtime is unavailable.

- Pause
- Route
- Recover

WATCH

Capacity stays visible

Monitor cost, latency, queue depth, failures, fallback use, saturation, and support burden.

- Cost
- Latency
- Health

> This packet is public-facing. It is written for serious review without exposing private infrastructure, customer data, credentials, live provider wiring, or internal project labels.

02

WORKFLOW MAP

The operating path should be visible before anyone trusts the outcome.

Folium uses workflow maps to turn broad AI ambition into inspectable work. Each phase names the procedure, the visible output, and the decision gate that prevents excitement from outrunning control.

DECISION GRID | REVIEW LENS | NEXT STEP

PHASE	PROCEDURE	VISIBLE OUTPUT	DECISION GATE
Inventory	List workloads, data classes, users, providers, sources, models, latency needs, and action authority.	Runtime inventory.	The route map starts from real work.
Classify	Group tasks by privacy, speed, retrieval, compute weight, customer exposure, cost, and support expectation.	Workload class map.	Each workload has a reason.
Place	Choose cloud, private, local, GPU, CPU, container, edge, database, RAG, or hybrid route.	Placement matrix.	The runtime matches the workload.
Protect	Define fallback, degraded mode, queue behavior, rate limits, and stop conditions.	Fallback plan.	The system can fail safely.
Operate	Track latency, cost, token budget, cache value, quota use, saturation, route health, source freshness, incidents, and support ownership.	Capacity cockpit.	Expansion is based on operating signals.

03

RECORDS AND WORK PRODUCTS

The work should leave behind material a buyer can inspect.

A serious engagement should produce more than conversation. Folium packages records, diagrams, checklists, routes, system surfaces, launch gates, and handoff material so the buyer can keep control after the first win.

DECISION GRID | REVIEW LENS | NEXT STEP

WORK PRODUCT	WHAT IT CONTAINS	HOW THE REVIEWER USES IT
Runtime placement matrix	Cloud, private, local, GPU, CPU, container, edge, RAG, and database options scored by workload.	Shows why each route exists.
Capacity dashboard	Latency, queue, cost, fallback, saturation, incidents, and source freshness.	Keeps AI operations visible.
Fallback contract	Primary route, relief route, degraded mode, blocked actions, and rollback trigger.	Prevents improvisation during outages.
AI FinOps ledger	Token budgets, semantic caching, quota limits, repeated-prompt waste, provider spend, and route-cost decisions.	Keeps AI savings from becoming unmanaged AI overhead.
Cost and latency ledger	Expected cost, current cost, latency bands, and review thresholds.	Lets leaders adjust before spend or delay becomes damage.
Support route map	Owners, incident paths, service boundaries, and handoff records.	Makes runtime support accountable.

04

PROCEDURES

The procedure is the product as much as the technology.

The goal is not to make AI look impressive for one meeting. The goal is to make the operating path repeatable, explainable, reviewable, and safe enough to improve.

CHECKLIST	OWNER PATH	RELEASE SIGNAL

- Classify every workload before choosing a model or provider.
- Separate private, public, customer-facing, retrieval-heavy, and state-changing work.
- Choose runtime placement by risk, cost, latency, privacy, resilience, support, and ownership.
- Define fallback and degraded-mode behavior before launch.
- Track capacity, cost, failures, queue depth, and source freshness after launch.
- Use semantic caching, quota alerts, token budgets, and route-cost reviews where repeated AI work creates waste.
- Avoid locking the business into one runtime when a hybrid route better fits the work.
- Document promotion, parking, failover, rollback, and support ownership.

05

CONTROLS

Governance, quality, and launch gates keep speed honest.

Folium keeps the buyer's next decision tied to observable gates: source truth, authority, access, testing, ownership, support, rollback, and improvement cadence.

DECISION GRID | REVIEW LENS | NEXT STEP

GATE	WHAT MUST BE TRUE	STOP OR REFINE SIGNAL
Placement gate	Workload class, data boundary, cost, latency, and support needs are named.	One runtime is assumed for every job.
Fallback gate	Primary, fallback, degraded, and blocked modes are defined.	The team will improvise during failure.
Cost gate	Expected spend, token budgets, cache strategy, quota alerts, monitoring, and review thresholds are visible.	Token, provider, or infrastructure cost is unknown.
Support gate	Every route has owner, incident path, and release record.	The runtime becomes orphaned.
Expansion gate	Capacity and quality signals justify more users, data, or authority.	Expansion is based on excitement alone.

06

DISCOVERY QUESTIONS

The right questions expose the real project.

These prompts help a buyer and Folium decide whether the next step should be education, audit, first build, security review, pilot, or an operating support path.

CHECKLIST	OWNER PATH	RELEASE SIGNAL

- Which tasks are private, public, high-speed, retrieval-heavy, batch, or state-changing?
- Which workloads need GPU capacity and which can run on CPU or a lighter route?
- What happens when the primary model, provider, source, or queue fails?
- Which sources need freshness checks before the AI answers?
- What cost signal would force a route change?
- Which repeated prompts, retrieval calls, or model routes deserve caching, quota controls, or cheaper placement?
- Who owns runtime incidents and fallback decisions?

07

VISUAL DIGESTION

Diagrams, charts, and overlays make the work easier to review.

Dense AI work should not only be explained in paragraphs. The reviewer should be able to inspect maps, scorecards, matrices, lanes, and before-after views that reveal where the value and risk live.

RECORD | BOUNDARY | ACTION

Runtime lane map

Shows each workload and its best-fit cloud, private, local, GPU, CPU, retrieval, or hybrid route.

- Class
- Route
- Fallback
- Owner

Capacity cockpit

Tracks latency, cost, queue depth, fallback use, saturation, and incidents.

- Latency
- Cost
- Queue
- Health

Fallback tree

Names pause, degrade, reroute, manual review, rollback, and relaunch options.

- Pause
- Degrade
- Reroute
- Relaunch

Placement scorecard

Compares privacy, speed, cost, resilience, and supportability.

- Privacy
- Speed
- Cost
- Support

08

OPERATING ROLES

Every serious AI path needs named owners before it becomes dependency.

The same technology can be safe or unsafe depending on who owns the workflow, data, quality, launch authority, support, and improvement loop. Folium makes those responsibilities explicit so no buyer inherits an orphaned system.

DECISION GRID | REVIEW LENS | NEXT STEP

ROLE	OWNS	RECORD TO INSPECT
Executive sponsor	Priority, budget, risk tolerance, stop/continue decision, and expansion timing.	Decision note, value hypothesis, and approval boundary.
Business process owner	The day-to-day work, acceptance criteria, staff impact, and operational usefulness.	Workflow map, user feedback, and adoption notes.
Technical owner	Systems, APIs, databases, runtime placement, deployment, monitoring, and fallback.	Architecture map, integration log, and support route.
Knowledge owner	Source truth, document freshness, policies, retrieval scope, and correction workflow.	Source inventory, freshness cadence, and review exceptions.
Security or risk reviewer	Data classes, credentials, access, logs, retention, blocked actions, and incident path.	Boundary map, permission table, and rollback trigger.
Folium delivery lead	Build coordination, review file, known limits, quality checks, and handoff completeness.	Launch room, eval record, and improvement backlog.

09

QUALITY SCORECARD

A max-detail packet should tell reviewers how to judge the work.

Folium uses scorecards to make a subjective AI conversation more inspectable. The score is not a substitute for judgment; it helps leadership see whether the next step is education, repair, sandbox, pilot, or operations.

DECISION GRID | REVIEW LENS | NEXT STEP

SCORE AREA	STRONG SIGNAL	WEAK SIGNAL
Business fit	The workflow is specific, painful, owned, and tied to measurable operational improvement.	The project is framed as adding AI generally.
Source truth	Approved sources are known, fresh, classified, and connected to the answer path.	The system mixes stale, unknown, or unapproved sources.
Behavior quality	Representative tasks pass, wrong-answer behavior is known, and edge cases are recorded.	The review build only shows a polished happy path.
Authority control	AI actions are separated into draft, retrieve, recommend, route, execute, block, and escalate.	The system can act without visible permission.
Staff readiness	Users can explain the tool, correct it, escalate, and understand their role.	Staff feel replaced, confused, or unsupported.
Operations readiness	Support, monitoring, rollback, release rhythm, and source refresh are owned.	No one knows who maintains the system after launch.

10

THIRTY / SIXTY / NINETY

The work should have a believable first ninety days.

A controlled first ninety days keeps ambition high without turning uncertainty into production risk. Folium uses the period to move from understanding into a narrow working example, then into reviewable operating rhythm.

DECISION GRID | REVIEW LENS | NEXT STEP

WINDOW	FOCUS	EXPECTED OUTPUT
First 30 days	Discovery, source inventory, first-lane selection, staff interviews, data boundary, and build plan.	Process map, owner map, first-build scope, source list, and launch blockers.
Days 31-60	Working surface, RAG or agent behavior, integration stub, evaluation cases, browser checks, and staff review.	Sandbox, evaluation file, screenshots, known limits, and repair list.
Days 61-90	Architecture review, pilot conditions, governance layer, training guide, support path, and improvement cadence.	Launch room, go/no-go record, operations guide, and next-stage recommendation.

11

RISK AND ASSUMPTION REGISTER

The hidden assumptions should be visible before they become expensive.

Every AI engagement contains assumptions about data, people, systems, cost, behavior, and authority. Folium treats those assumptions as review material, not background noise.

DECISION GRID | REVIEW LENS | NEXT STEP

ASSUMPTION	WHY IT MATTERS	HOW FOLIUM REVIEWS IT
The source is authoritative	AI can only be as reliable as the sources and business rules it is allowed to use.	Source inventory, owner confirmation, retrieval tests, freshness cadence.
The process is ready	A broken process can become a faster broken process when AI is added too early.	Workflow mapping, bottleneck review, owner interview, first-lane narrowing.
The runtime fits the data	Cloud, private, local, and hybrid routes carry different privacy, cost, latency, and support tradeoffs.	Runtime matrix, data classification, provider review, fallback plan.
Staff will adopt the tool	Adoption fails when users do not understand, trust, correct, or benefit from the system.	Training notes, staff review, feedback loop, manager visibility.
Authority is clear	The system can create harm if it sends, updates, approves, or routes without permission.	Permission table, blocked actions, human review, audit trail.
The system can be supported	A useful first build becomes fragile if nobody owns incidents, source updates, or cost review.	Support guide, owner map, release rhythm, rollback trigger.

12

FIRST SPRINT PROCEDURE

The first sprint should produce something real and reviewable.

Folium prefers a narrow first sprint that creates a working surface or review file the buyer can challenge. The first sprint is not the final system; it is the safest way to make the future visible.

CHECKLIST	OWNER PATH	RELEASE SIGNAL

- Confirm the single process and the decision the sprint must support.
- Collect approved example material, redacted review records, public references, screenshots, workflow notes, and source rules.
- Define what will be built: portal, dashboard, RAG assistant, agent route, integration adapter, audit file, or launch room.
- Create the visual workflow: intake, source, model or agent route, human review, output, record, and next gate.
- Run representative tasks, edge cases, bad input, missing data, and blocked-action tests.
- Prepare browser screenshots, known limits, support questions, and next-stage blockers.
- Review with staff and leadership before expanding data, access, authority, or dependency.
- End with a decision: stop, refine, rebuild, pilot, or prepare an operating plan.

13

REFERENCE WORK PRODUCTS

The packet should make the invisible work tangible.

AI work often fails because the important pieces are invisible until something breaks. Folium turns those pieces into work products the buyer can open, print, challenge, and improve.

RECORD | BOUNDARY | ACTION

Process map

A before-and-after workflow showing people, systems, data, decision points, blockers, and expected output.

- Before
- After
- Owner
- Gate

Data boundary map

A map of source classes, approved use, blocked use, retention, provider exposure, and custody.

- Public
- Internal
- Private
- Blocked

Model and agent route

A path showing which model, tool, retrieval source, or agent lane is used and where humans approve.

- Route
- Tool
- Review
- Escalate

Evaluation file

A record of tasks, expected outcomes, failures, repairs, known limits, and acceptance criteria.

- Cases
- Failures
- Repairs
- Limits

Launch room

A board for owners, support, training, rollback, incidents, go/no-go, and improvement backlog.

- Owner
- Support
- Rollback
- Backlog

Handoff guide

A plain-language guide staff can use to understand what the system does, cannot do, and how to report problems.

- Use
- Limit
- Correct
- Report

14

METRICS AND REVIEW RHYTHM

The business should know how improvement will be measured.

Folium keeps measurement practical. The first goal is not a perfect dashboard; it is a clear set of signals that shows whether the process is saving time, reducing risk, strengthening staff, or improving customer outcomes.

DECISION GRID | REVIEW LENS | NEXT STEP

SIGNAL	WHAT TO WATCH	DECISION IT SUPPORTS
Time recovered	Manual steps removed, average handling time, repeated work reduced, faster routing.	Should this workflow expand to more users or adjacent processes?
Quality improved	Wrong answers, missing sources, correction rate, review exceptions, customer rework.	Is behavior strong enough for pilot or does it need repair?
Risk reduced	Blocked unsafe actions, escalations, data-boundary violations avoided, rollback readiness.	Can authority expand or should controls remain tight?
Staff confidence	Training completion, feedback volume, adoption friction, override rate, manager notes.	Does the workforce need more support before launch?
Cost and runtime	Provider cost, local infrastructure cost, latency, uptime, fallback use, subscription sprawl.	Should runtime placement change?
Customer impact	Response speed, consistency, issue resolution, conversion support, satisfaction signals.	Is the capability improving the business outcome?

15

ROLE REVIEW ROUTE

Each reviewer should know what to inspect first.

A max-detail packet is only useful when different reviewers can find their lane quickly. Folium separates executive, operations, technical, security, finance, and staff questions so the buyer can bring the right people into the right part of the review.

DECISION GRID | REVIEW LENS | NEXT STEP

REVIEWER	START WITH	DECISION THEY SUPPORT
Executive sponsor	Value hypothesis, launch gate, first ninety days, and stop/refine/continue choices.	Whether the process deserves a controlled engagement.
Operations lead	Workflow map, operating roles, support rhythm, and staff feedback loop.	Whether the future process can be run by the team.
Technical lead	Runtime placement, data path, integration surface, monitoring, and fallback.	Whether the architecture can be supported safely.
Security or risk reviewer	Data classes, permissions, blocked actions, logs, retention, and rollback.	Whether access can expand beyond public review.
Finance or owner	Cost signals, subscription overlap, runtime tradeoffs, labor impact, and support burden.	Whether the first build has a practical business case.
Staff user	Plain-language use, limits, escalation, correction path, and training expectations.	Whether the tool strengthens the job instead of confusing it.

16

BUYER WORKSHEET

The packet should turn into a working session, not only reading material.

Before a call, Folium wants the buyer to gather the real operating pieces that make the review useful. The worksheet keeps the conversation grounded in one process, one owner, one source map, and one next decision.

CHECKLIST	OWNER PATH	RELEASE SIGNAL

- Bring one workflow that is slow, risky, expensive, repetitive, customer-visible, or staff-heavy.
- Name the systems touched by the workflow: store, CRM, ERP, inbox, spreadsheet, database, portal, document folder, or legacy application.
- Separate approved public material from internal, customer, regulated, confidential, credential, and blocked material.
- Write down who owns the work today, who reviews exceptions, and who will own the AI-assisted version.
- List the decisions AI may draft, retrieve, recommend, route, block, or escalate, and the decisions that stay human-owned.
- Bring examples of good output, bad output, common exceptions, missing data, and customer-facing risk.
- Name the first useful working surface: dashboard, portal, assistant, queue, control room, commerce lane, integration, or review file.
- Decide what record would make leadership comfortable with the next stage.

17

ENGAGEMENT FIT LADDER

The next step should match the maturity of the record.

Folium does not need every buyer to start at the same altitude. The right offer depends on how much process clarity, source truth, owner alignment, and launch readiness already exists.

DECISION GRID | REVIEW LENS | NEXT STEP

IF THE BUYER HAS	BEST NEXT FOLIUM MOVE	OUTPUT TO EXPECT
AI interest but no clear process	AI systems audit or first workflow finder.	Pressure map, source inventory, first-lane recommendation, and risk view.
A clear process but no working surface	Forward engineering first sprint.	Clickable surface, route map, known limits, and next-stage blockers.
A tool that works in parts but not in operations	Architecture and launch readiness review.	Permission map, runtime decision, support model, and go/no-go record.
A failed or frightening rollout	AI recovery and staff enablement path.	Issue register, staff training plan, repair roadmap, and confidence loop.
Sensitive data or cost pressure	Local, private, or hybrid AI placement review.	Runtime matrix, data custody plan, fallback route, and vendor-exit view.
A useful pilot that needs care	AI operations support.	Monitoring rhythm, source refresh, release notes, incident path, and improvement backlog.

18

HANDOFF

The last page of a packet should create the next controlled move.

Folium's handoff view separates what can be done now, what needs customer records, what needs approval, and what should wait until the review file is stronger.

DECISION GRID | REVIEW LENS | NEXT STEP

HANDOFF LANE	OWNER	NEXT RECORD
Executive sponsor	Priority, budget, stop/continue decision, and expansion timing.	Decision memo, value hypothesis, and next-stage gate.
Business process owner	Daily workflow, user acceptance, staff impact, and usefulness.	Workflow map, exception list, and adoption notes.
Technical owner	Runtime, integrations, APIs, databases, deployment, monitoring, and fallback.	Architecture map, route contracts, and support guide.
Risk or security owner	Data classes, permissions, logs, blocked actions, incident path, and rollback.	Boundary map, permission table, and rollback record.
Folium delivery lead	Build coordination, evaluation, known limits, launch room, and handoff completeness.	Review file, release notes, and improvement backlog.

> The strongest next step is narrow: one process, one owner, one source map, one working surface, one review file, and one decision gate.

19

NEXT STEP

Runtime placement is where AI becomes operational.

Use this packet when AI needs to run across more than one tool, provider, model, route, or infrastructure pattern without losing control.

Bring the process

Name the business process, the systems involved, the people affected, and the decision this PDF should support.

Separate review from production

Keep public examples, sandbox review, pilot access, and production dependency in separate stages with clear owners.

Ask for the record

Request screenshots, browser checks, known limits, launch blockers, support plans, and the next approval path.