0 | Front-Matter

Field	Details
Document	*RHOAI v3 UI – AI Engineer (AIE) Persona Workshop Report*
Session	"Re-organising RHOAI for GenAI" — Day 2 (Persona & Pain-point deep-dive)
Date & time	28 May 2025 · 08:00-12:03 ET (243 min)
Venue / modality	Google Meet (hybrid), live Miro board
Transcript source	transcript (.txt) — archived in project Drive
Facilitators	Dash Copeland (PM/UX) · Peter Double (UX)
Participants	Adel Zaalouk · Andy Braren · Ann Marie Fred · Burr Sutter · Dash Copeland · Eder Ignatowicz · Jason Greene · Jenn Giardino · Jon Nemargut · Peter Double · Tony Kay
Authors of this report	Gemini (draft) → to be reviewed by Dash Copeland & Peter Double
Document status	Working draft — v0.1 (29 May 2025)
Purpose	Capture Day-2 outcomes (persona, pains, assets) in a share-ready format for Engineering & PM leads.

---

1 | Executive Summary

Day 2 of the RHOAI v3 design sprint transformed **raw discussion into concrete artefacts**. Key outcomes:

- **Persona re-framing:** consensus to standardise on the industry-recognised term **"AI Engineer"** for the pro-code power user building agentic apps.
- **Validated pain matrix:** four high-leverage pains (model choice, eval toil, data plumbing, opaque guard-rails) mapped to matching design bets (Registry, Auto-Eval pipeline, Data connector, Policy-aware scaffolds).
- **Journey map v C:** a five-stage flow (Discover → Operate) embellished with feelings, key actions and ~40 capability stickies harvested live on Miro.

- **Capability backlog:** 11 Stage-2 items (Playground, Data Manager, …) and 15 Stage-3 items (Auto-Evals, LLM-Judge, …) now prioritised for spec drafting.
- **Process insight:** hybrid workflow (morning talk → afternoon Gen-AI synthesis → next-day review) is working; transcript fidelity is critical (issue encountered with Gemini Notes).

Next-step actions

1. Finish sections 3-9 of this report; circulate for inline comments by 30 May.
2. Feed validated pain-points into feature-discovery session (Day 3 agenda).
3. Create first-cut PR-FAQ and Kick-start templates ahead of Tech-Stack day.
4. Establish secure, loss-proof transcript workflow (ditch Gemini Notes; keep local Tactiq).

2 | Workshop Highlights

Theme	Highlight
Opening context	*"We're trying to support people building agentic solutions… give them the power to make any of their dreams come true."* – Peter Double, 06:58 ET
Persona naming debate	lively discussion on "AI Builder vs AI Engineer"; consensus to adopt **AI Engineer** as the externally understood term.
Process win	Team validated that overnight Gen-AI synthesis of sticky notes → persona report saved hours; plan to iterate this "morning review / evening AI" rhythm.
Transcript scare	Discovery that enabling Gemini Notes deleted the Meet transcript; fallback Tactiq recorder saved the day – action to formalise capture.
Quote of the day	*"Everybody's guessing… we need to hand them sane defaults, not infinite options."*
Duration & energy	4-hour sprint, one 15-min break; notable engagement (no drop-offs) across all disciplines.

3 | Persona Refinement — "AI Engineer"

3.1 Persona Card

Field	Snapshot (validated 28 May 2025)
Archetype	**Pro-code AI Engineer** — builds end-to-end Gen-AI systems (model ↔ RAG ↔ agent ↔ UI), customises prompts & workflows.
Environment	Starts with a *Golden-Path Wizard* that spins up **VS Code / Cursor + GitHub CI**. Needs strong *"models-as-a-service"* and opinionated tools; **doesn't want to touch raw OpenShift YAML**.
Role / day-job	Staff / Lead engineer embedded in product or innovation squad; chartered with shipping PoCs that can harden into prod.
Signature quote	"**I want a sandbox that behaves like prod.**"
One-liner	*"Give me sane defaults, then get out of my way so I can ship."*
Critical success metric	**Median Time-to-First-Working-Agent ≤ 30 min** in Dev Sandbox.

3.2 Updated Key Traits *(ordered by discussion emphasis)*

1. **Need for speed** — dopamine hit from seeing "it works!" in minutes.
2. **Framework-first** — LangChain / LangGraph preferred over raw model APIs.
3. **IDE-native** — lives in VS Code; despises notebook drift.
4. **Low-toil bias** — loves less setup but keeps full control.
5. **Security-aware (not auth expert)** — wants baked-in guardrails.
6. **Continuous learner** — AI-Dev YouTube, Discord, X at night.
7. **Collaborative** — shares PRs/snippets in Slack & GitHub gists.
8. **Seeks quick peer validation** — rapid share for feedback loops.

3.3 Goals & Motivations

#	Statement
1	**Rapidly iterate safely** — fast loop for prompt/model/data tweaks with auto-evals
2	**Ship MVP fast** that solves a real business pain
3	**Own the full stack** without begging Platform Ops
4	**Leverage open standards & community** ecosystems
5	**Win Infosec & Ops trust** so PoCs reach prod

3.4 Validated Pain Points & Design Opportunities

Pain Point (verbatim)	High-Leverage Design Opportunity
“I need help to access and choose the right model / MCPs / artifacts.”	**Model / Agent / MCP Registry** with HW-fit guidance
Manual / unclear evaluation workflows	**Auto-Eval pipeline** with default test suites
“Data Data Data / Context”	**Standardised data connectors + feature store**
Security / guardrails feel opaque	**Policy-aware scaffolds** (auth & guardrails by default)

3.5 Representative Quotes

- "The only problem we need to solve first is 'which model can I use?' I cannot get started because I don't have a model." — Burr Sutter
- "You can rapidly iterate — we'll set up your project and then get out of the way." — Jason Greene
- "Everybody's guessing… we need to give clients good practices that narrow the options and provide guidance." — Peter Double
- "We want to give them a sane default to start with, and then they can customise." — Adel Zaalouk

3.6 **Confidence & Gaps** Confidence & Gaps** Confidence & Gaps**

- **Alignment:** High convergence between transcript quotes and sticky-note votes (e.g., speed, model choice).
- **Gaps:** Need 5 external AI-Engineer interviews to validate pain-ranking; missing quantitative TTFWA telemetry.

---

4 | Validated Pain-Points & Design Opportunities

This section distills the four pain themes that surfaced repeatedly across the day-2 transcript and the live Miro synthesis. For each pain we capture: the user voice (verbatim snippets), the underlying need, and the design stance we agreed to pursue.

4.1 At-a-Glance Matrix

Pain Theme	What AI Engineers Experience	Design Opportunity	Why This Matters
P-1 Model / MCP Choice	"There's only one problem we need to solve right now — **what model can I use?** I'm dead in the water until I know." — Burr Sutter 03:47 ET	**Curated Model / Agent / MCP Registry** with hardware-fit, cost and capability filters	Unblocks first 10 minutes; keeps users inside RHOAI instead of Bedrock / Hugging Face
P-2 Evaluation Toil	"Evals are a big blocker … everybody is doing manual evals today." — Burr Sutter & Peter Double 03:31 ET	**Auto-Eval Pipeline** (preset tests, CI hooks, dashboard)	Makes quality measurable; speeds demo-to-prod hand-off
P-3 Data / Context Plumbing	"Every lost client was because we didn't have a proper **document-ingestion**	**Standardised Data-Connector + Feature-Store** wizard	Converts static & dynamic sources into RAG/context with < 10 clicks

Pain Theme	What AI Engineers Experience	Design Opportunity	Why This Matters
	pipeline." – Peter Double 03:37 ET		
P-4 Opaque Safety & Guardrails	"Is it safety? security? guardrails? We need to untangle that or nobody will trust the app." – team dialogue 02:46-02:49 ET	**Policy-Aware Scaffolds** (prompt & model guardrails, SBOM, risk scores)	Gives Infosec a single pane; lets Engineers focus on code

---

4.2 Deep Dive – P-1 "Which Model Can I Use?"

User voice

- "I cannot get started because I don't have a model." – Burr Sutter
- "The very first thing our clients need is centralised, validated models." – Peter Double

Root causes

- Fragmented discovery (Hugging Face, Bedrock, private registries).
- Unclear fit-for-purpose metadata (tool-calling, reasoning, GPU RAM, licence).
- Fear of future lock-in or surprise cloud bills.

Design stance

- Ship a **Registry** that merges Red Hat-validated models, user-added models and MCP servers.
- Surfaced facets: modality, agent-compatibility, cost/$1k tokens, vLLM readiness, licence.
- Provide starter "Top 3 for X" recommendations and side-by-side compare UI.

4.3 Deep Dive – P-2 "Evals Are Hard"

User voice

- "People are doing evaluation *oakley* today; there's **no streamlined way**." – Adel Zaalouk
- "We could write 20 more tickets just on evals." – Burr Sutter

Root causes

- Manual copy-paste / vibe-check culture.
- Lack of reusable test harnesses for factuality, cost, latency.
- No integration with GitHub/GitLab CI; drift checks missing.

Design stance

- **Auto-Eval Pipeline** (YAML + dashboard) shipping with default metrics.
- GitHub Action & Tekton tasks; green/red badge gates merges.
- Extensible judge-model plug-in system (supports LLM-as-Judge).

4.4 Deep Dive – P-3 "Data / I Need Context"

User voice

- "Clients are totally blocked without an **easy document pipeline**." — Peter Double
- "Data-related pain points, from cleaning to access, have been a big learning." — Andy Braren

Root causes

- Enterprise SoRs (SQL, PDFs, APIs) require bespoke ETL each time.
- Confusion between RAG vs dynamic data; vector vs graph vs SQL retrieval.
- Insufficient UI to preview, chunk, embed and validate.

Design stance

- **Data Manager Wizard** with connectors, chunk/embedding presets, preview & reindex.
- Generates reproducible pipelines (DAG) and registers datasets in a **Knowledge Base**.
- Integrates with evals to grade recall & citation accuracy.

4.5 Deep Dive – P-4 "Is It Safe?" (Guardrails & Security)

User voice

- "Safety is larger than guardrails… we must decouple safety and security." — Ann Marie Fred & Adel Zaalouk
- "Prompt guardrails vs model guardrails need different handling." — Peter Double

Root causes

- Ambiguity between safety (toxicity, PII) and infra security (auth, RBAC).
- No default guardrail when tool-calling external MCP services.
- Infosec approval late in cycle → last-minute blockers.

Design stance

- **Policy-Aware Scaffolds** auto-attach prompt & model guardrails at Dev Sandbox time.
- Generate SBOM + risk score; integrate Trusty AI checks.
- One-click security pipeline template for deployment stage.

4.6 Prioritisation & Stage Fit

Stage	Pain(s) Tackled First	Rationale
Discover / Ideate	P-1 Model choice	Unblocks “Hello World” (< 10 min)
Prototype / Build	P-3 Data plumbing + P-2 Eval toil	Needed before credible PoC demo
Evaluate / Iterate	P-2 Eval toil + P-4 Guardrails	Quality & safety gates to prod
Deploy / Operate	P-4 Guardrails + P-1 model updates	Infosec sign-off & model refresh

Early design spikes will focus on **Registry MVP** + **Auto-Eval v0.1**, as they de-risk both adoption and production readiness.

5 | AI-Engineer Journey Map (Rev C)

The journey map traces one AI Engineer from the moment a business idea lands to steady-state production. Each stage shows triggers, goals, actions, feelings, pain themes, and the RHOAI UI v3 capabilities that resolve them.

5.1 End-to-End Timeline

```
Unset
Discover/Ideate ─▶ Prototype/Build ─▶ Evaluate/Iterate ─▶ Deploy/Hand-off ─▶ Operate/Monitor/Scale
(hrs)          (days)         (days)         (days->wks)           (ongoing)
```

Moments-that-Matter ★:

1. ★ First PoC runs (< 10 min)
2. ★ Auto-Eval passes
3. ★ Security pipeline green-lights
4. ★ Drift alert auto-resolved

5.2 Stage-by-Stage Detail

Stage 1 – Discover / Ideate

Trigger	New Jira ticket / Slack ask: “Can we add AI for X?”
Goal	Validate idea and pick a sane starting point.
Key actions	Skim brief → Browse *Use-Case Gallery* → Run 5-min *QuickStart Wizard* → Book 30-min call.
Touch-points	Slack · Jira · Confluence playbook · **RHOAI UI v3 Gallery + Wizard**
Emotion	😀 curiosity → 😵‍💫 overwhelm (too many choices)
Top pains	Unaware Red Hat does AI · No golden path
Key capabilities	Free Community Resources (Bootstrapper Tool, MCP Hub) · Getting-Started kits (Kickstarts, Your First Agent) · Marketing / Enablement (YouTube, blogs, workshops) · UX / Docs (workflow-based IA, one-click model endpoint)

Stage 2 – Prototype / Build (Playground)

Trigger	Sandbox created; GPU quota assigned.
Goal	Run a working PoC agent on proprietary data.
Key actions	Launch Dev-Sandbox Wizard → Clone Golden-Path repo → Connect Postgres via Data-Manager → Add guardrails → Share Live Share.
Touch-points	VS Code ext · RHOAI CLI ais up · Playground UI · GitHub Live Share
Emotion	🤩 in-flow → 🧐 mild frustration (data wiring) → 😃 win when PoC answers query
Top pains	Which model? · Data plumbing · Manual eval · Guardrail clarity
Key capabilities	Playground · Data Manager · Good Model Search & Better Model Info · Bootstrapper Templates · Prompt Management · Out-of-Box Guardrails · Easy Inference · Cost/Efficiency Analyzer

Stage 3 – Evaluate / Iterate

Trigger	Demo date set; quality & safety must pass.
Goal	Quantify accuracy, latency, cost; minimise hallucinations.
Key actions	Wire Auto-Eval YAML → Run GitHub Action → Review Eval Dashboard & LLM-Judge scores → Slack SME feedback → Re-tune prompt & RAG chunks.
Touch-points	GitHub Actions · Eval Dashboard · Slack DM · VS Code terminal
Emotion	😖 anxious KPIs → 😌 relief when metrics green
Top pains	Eval toil · Guardrail gaps · Trace debugging
Key capabilities	Auto Evals · LLM-as-Judge · Eval Feedback & Creation Help · Tracing · Human-in-the-Loop · Version Tracking · Advanced RAG · Safety Guardrails · CI helpers

Stage 4 – Deploy / Hand-off

Trigger	MVP accepted; must hit prod cluster.
Goal	Push agent to production with minimal re-work.
Key actions	Run ais deploy (Helm/OCP manifests) → Execute Security Pipeline (Snyk + SBOM) → Open Change-Request ticket → Use Cost Estimator.
Touch-points	GitHub PR · OCP Pipelines UI · Security portal · Jira CR · Cost dashboard
Emotion	😬 compliance stress → 🙂 confidence after approvals
Top pains	CVE/SBOM workload · RBAC confusion · Budget anxiety
Key capabilities	Secure Artifact · Security & Policy Automated Testing · Auth / Access · Governance · Scaling / Deployment · CLI & Build Flow · Monitoring / Observability bootstrap · Cost estimator · Red-Team Safety tests

Stage 5 – Operate / Monitor & Scale

Trigger	Agent live; traffic & data drift begin.
Goal	Keep app fast, cheap, accurate; spot drift/cost spikes early.
Key actions	Watch Grafana & Ops UI → Respond to PagerDuty drift alert → Bump model version → Schedule weekly Auto-Eval → Export ROI dashboard.
Touch-points	RHOAI Ops UI · Grafana · PagerDuty · Slack alerts · BI dashboards
Emotion	😎 confidence when dashboards green → 😤 stress during spikes
Top pains	Blind-spot metrics · Drift & hallucinations · GPU cost
Key capabilities	Monitoring / Observability · Automated Eval & Alerting · Workload Logging

---

5.3 Capability Heat-Map

Stage	**Capability clusters with highest user value↔feasibility**
Discover	Model/Agent Registry → Use-Case Gallery → QuickStart Wizard
Prototype	Playground → Data Manager → Out-of-Box Guardrails
Evaluate	Auto-Eval Pipeline → LLM-Judge → Human-in-the-Loop share
Deploy	Secure Artifact + Security Pipeline → Cost Estimator
Operate	Monitoring Dashboard → Drift Alerting → Auto-Eval cron

(Used MoSCoW + t-shirt sizing from workshop whiteboard to determine "highest value ↔ feasible by November" clusters.)

---

6 | Capability Backlog by Stage (MVP → GA)

This backlog aggregates every capability sticky captured on the Miro board and clusters them by journey stage. Each item carries:

- **Priority** – MoSCoW label agreed in the workshop. *Must* = MVP-critical • *Should* = Nov release target • *Could* = stretch / GA.
- **Effort** – rough T-shirt sizing from the engineering huddle (S ≤ 2 sprints, M ≈ 4, L > 4).

- **EPIC tag** – shorthand for cross-linking in Jira.

6.1 Stage 1 – Discover / Ideate

EPIC	Capability	Priority	Effort
UC-GALLERY	*Use-Case Gallery* with filterable real-world templates	**Must**	M
QSTART-WIZ	5-min *QuickStart Wizard* (Golden-Path video + workspace setup)	Must	M
REG-MCP	*MCP Hub* — browse model/agent servers, HW fit	Should	L
BOOT-GEN	*Gen-AI Bootstrapper CLI* (model+server picker, code snippet)	Should	M
DOCS-IA	Workflow-based IA overhaul of docs + landing page	Could	S
COMM-YT	YouTube explainer series & SEO blogs	Could	S

6.2 Stage 2 – Prototype / Build

EPIC	Capability Cluster	Priority	Effort
PLAY-CORE	*Playground* – chat UI, side-by-side model compare, export-to-.py	**Must**	L
DM-PIPE	*Data Manager* – ingestion wizard, KB registry, preview	Must	L
REG-MODEL	*Model Registry* – RH-validated + user-import, model cards	Must	L
SEARCH-MODEL	Advanced model search / filter by use-case, HW, licence	Should	M
TMP-GOLD	Golden-Path repo & *Bootstrapper Templates*	Must	M

EPIC	Capability Cluster	Priority	Effort
PM-WORK	Prompt-Management workspace with versioning & eval hooks	Should	M
GUARD-OOB	Out-of-Box prompt & output guardrails	Must	S
MaaS-EASY	Shared Inference (MaaS) endpoints + “host-my-own” guide	Should	M
COST-ANA	Tokens/s, $/1k tokens, tokens/Watt panel	Could	S

6.3 Stage 3 – Evaluate / Iterate

EPIC	Capability Cluster	Priority	Effort
EVAL-PIPE	*Auto-Eval Pipeline* (YAML spec, CI hooks, dashboard)	**Must**	L
LLM-JUDGE	Pluggable *LLM-as-Judge* scorer library	Must	M
TRACE-VIS	Trace viewer with good/bad/ugly tagging	Should	M
HITL-SHARE	Human-in-the-Loop share & feedback pane	Should	S
VER-TRACK	Version diff & rollback across model / prompt / eval runs	Should	M
ADV-RAG	GraphRAG, re-ranking, advanced chunking + retrieval metrics	Could	L
SAF-GRDL	Eval-time safety guardrail checks	Must	S
CI-KIT	Tekton / GitHub Actions templates (build, test, eval)	Must	S
DATA-LABEL	SME word-level data-labeling micro-app	Could	M

6.4 Stage 4 – Deploy / Hand-off

EPIC	Capability Cluster	Priority	Effort
SEC-ART	*Secure Artifact* pipeline – SBOM, CVE scan, sign & push	**Must**	M
POL-TEST	Policy & security automated tests in CI	Must	M
AUTH-EZ	End-to-end Auth/RBAC autoconfig (Keycloak gateway)	Should	L
COST-EST	GPU & token Cost Estimator pre-deploy	Should	S
DEP-ANY	"Deploy Anywhere" Helm/OCP manifest generator	Should	M
MON-BOOT	Bootstrap Monitoring & Observability packs	Should	M
SCALE-GUIDE	Autoscaling guide & capacity planner	Could	S
REDTEAM	Red-Team safety test harness (prod gating)	Could	M

6.5 Stage 5 – Operate / Monitor & Scale

EPIC	Capability Cluster	Priority	Effort
MON-DASH	Unified Ops dashboard (TTFT, drift, tokens/Watt, GPU usage)	**Must**	M
EVAL-CRON	Scheduled Auto-Eval jobs + drift alerts	Must	S
ALERT-RISK	Hallucination & cost spike alert rules	Must	S
LOG-WL	Workload & convo logging with user-feedback thumbs	Should	M
ROI-BI	Exportable BI dashboard (ROI, adoption)	Could	S

6.6 Road-mapping Notes

- MVP (Q4 CY25) focuses on **PLAY-CORE, DM-PIPE, EVAL-PIPE, SEC-ART, MON-DASH** — de-risk adoption & prod readiness.
- Parallel spikes on **REG-MODEL** and **AUTH-EZ** start ASAP to align with infra teams.
- *Should* items are November target to wow Kubecon demos; *Could* roll into GA or fed to community.

7 | Proposed Next Steps (90-Day Action Plan)

Action Item	Owner(s)	Due	Notes / Deliverable
Publish *AI Engineer Persona v1* (Section 3) to Confluence, solicit cross-BU feedback	Dash Copeland	**3 Jun**	Include traits, goals, pains as tables; link to Miro board captures
Convert Capability Backlog (Section 6) into **Jira EPICs & Stories**	Peter Double, Adel Zaalouk	5 Jun	Use EPIC tags (PLAY-CORE, DM-PIPE…) and MoSCoW priorities
Schedule **5 external AI-Engineer validation interviews** (pain ranking)	Andy Braren (UX)	14 Jun	Target partners who built on Bedrock / Vertex; feed findings into backlog refinement
Draft **PR-FAQ v0.1** for exec review (Scope: Registry + Auto-Eval MVP)	Jenn Giardino	10 Jun	Align messaging with Sales / Enablement; embed screenshots of Wizard mock-ups
Complete **Architecture Spike**: Model Registry metadata & Auth-Gateway flow	Eder Ignatowicz, Tony Kay	12 Jun	Decision doc comparing Open-Model-Zoo vs in-house schema; PoC Keycloak flow
Stand-up **Dev Sandbox Prototype Environment** (Llama-3-B8 preset, 1 GPU)	Jason Greene (Ops)	17 Jun	Enables PLAY-CORE demo; tied to Time-to-First-Working-Agent metric

Action Item	**Owner(s)**	**Due**	**Notes / Deliverable**
Produce **Kickstart Template & Demo Video** (Document summarisation use-case)	Burr Sutter	21 Jun	2-min voice-over, code repo link; will be Gallery item #1
Define **Telemetry & KPIs**: TTFWA, Auto-Eval pass-rate, tokens/Watt	Adel Zaalouk + Observability guild	24 Jun	Add to MON-DASH backlog; publish metric spec
Create **Security Pipeline Template** (SBOM + Snyk + Trusty)	Security Eng (Jenn G.)	28 Jun	Integrate into ais deploy flow; gate prod promotion
Prepare **Kubecon Demo Storyboard** using MVP feature set	Burr Sutter, Dash Copeland	10 Jul	Needs working Registry, Playground, Auto-Eval; draft talk-track

Milestones & Review Cadence

- *M0* (3 Jun) — Persona & backlog frozen
- *M1* (17 Jun) — Dev Sandbox prototype up; spike decisions documented
- *M2* (1 Jul) — MVP feature code complete for internal dog-food
- *M3* (15 Jul) — Demo freeze for Kubecon dry-run

Weekly “Wednesday Window” stand-ups (30 min) will track progress and unblock owners.