



Connect

Red Hat AI

1000 PoCs – lessons learned, road ahead

Andreas Bergqvist

Red Hat AI EMEA



Background



Initiated by Red Hat leadership driven by

A Need for deeper insights

Customers were asking for more help to get started

We had something to prove



Where have we spent our time

Categories

Amplify Model Performance

44%



Optimize Model Inference

56%



Use Cases

14%



Models as a Service

10%



GPUs as a Service

51%



Expert Assistant

10%



CLI/Code Assistant

9%



Hybrid Serving

6%



Automation



Ex nr 1 – Swedish purchasing bot

Finetuning of model w InstructLab

Many iterations of trainings

3 days to complete finetuning

Successful demo - good accuracy

BUT:

Still not part of methodology - no clear team to own the task

What should we have done instead?



Ex nr 2

Automotive customer going AI ops

Big Openshift user. Wanted to leverage OCP Lightspeed

On prem

Air gapped

Limited HW (nVidia T4 GPUs)

Model that fits (Granite 3.1 2B)- bad answers.

- ▶ However, Mistral 7B Instruct v0.3 and Granite 3.2 8B are too large

POC

Quantized model (compression to make it smaller)

Test behaviour vs benchmark

Test speed

How can we help them? 🙌 Quantization

Results: Happy users. Case implemented as a standard solution. Rolled out in all factories over the coming year.

Why? Business case identified. Team to roll it out identified and ready



Ex nr 3

vLLM vs proprietary model serving

Customer: Large bank.

Wanted Red Hat to perform benchmark of specific models on vLLM

To be compared to nVidia runtimes

Took 2 weeks to perform. Including optimizing models/vLLM integration.

Results came in rolling over these weeks.

Some benchmarks vLLM faster, some same, some few percent slower.

seen as slow?

However: Feeling in the Red Hat team was that we maybe were

Reality: 2 weeks was ultra fast compared to own testing. To get there team had spent 6 months (!).

Result: Decided to implement vLLM.

Why: Stack **Security compliant**. Easier to set up. Easier to optimize



Lessons learned

Let's not just do PoCs

! Don't just do PoCs - don't prove what is already proven. (If we have documented it it is supported.)

Instead: **Do demos + workshops + MVPs**

Align request for **PoC with business case** - no case no PoC/MVP.

Work with your eco system from day 1 (hw, SI, apps, developers, business stakeholders)

Keep agreeing on the next step clear in collaboration with your ecosystem (ITTT)



PoC vs. MVP

Proof of Concept

- product validation,
- focus on product fit
- drives sales discussion
- not directly tied to business use case
- requires modest commitment and investment from decision makers
- “quick-and-dirty” implementation,
- short timeframe (days-weeks),
- short-lived environment

Minimal Viable Product

- use case realization,
- focus on solution creation
- drives adoption
- delivers business outcomes
- requires and ensures significant commitment and investment from customer (incl. decision-makers),
- production-grade implementation,
- longer timeframe (weeks-months),
- long-lived environment,

Red Hat AI Incubator

Co-creating MVPs with Red Hat Experts

The **Red Hat AI Incubator** is a hands-on, collaborative engagement designed to help organizations rapidly develop and deploy AI solutions. Spanning **4-12 weeks**, the Incubator focuses on enabling teams through guided co-creation on real-world use-cases to create AI services that solve business problems and function in your environment.



Goals

- ▶ **Prototype** a custom AI solution using your data
- ▶ **Focus** on automation and production readiness
- ▶ **Deploy** to your environment for immediate business value
- ▶ **Upskill your team** with best practices and MLOps skills
- ▶ **Break down silos** to establish effective AI workflows



Engagement Principles

- ▶ **Mission Clarity** - Align on the problem and business goals
- ▶ **Experimentation** - Rapid feedback cycles to iterate quickly
- ▶ **Culture** - Foster collaboration across business, data and IT
- ▶ **Speed** - Rapidly iterate on solutions
- ▶ **Open Source** - Drive innovation with collaboration

1000 PoCs

Wonderful but easily get's stuck

Let's continue to ask for **proof** but find easiest way to **prove quickly**

Let's build business case w your ROI

Let's share findings together for others to iterate on

Let's make AI great anywhere



Connect

Thank you



[linkedin.com/company/red-hat](https://www.linkedin.com/company/red-hat)



[facebook.com/redhatinc](https://www.facebook.com/redhatinc)



[youtube.com/user/RedHatVideos](https://www.youtube.com/user/RedHatVideos)



twitter.com/RedHat

