# 2025-05-28 | RHOAI Round-table (Wien)

Inhaltsverzeichnis:

# Teilnehmer

**Red Hat:**

- Stephan Kraft
- Mahmuthan Bastug (Basti)
- Matthias Rettl (Remote)
- Max Murakami
- Detlef Knierim
- Szabolcs Gleszer
- Philipp Bergsmann
- David Hanacek
- Thomas Ettenauer

Verschiedene **Unternehmen** waren durch ihre **Teilnehmer** an der Diskussion vertreten:

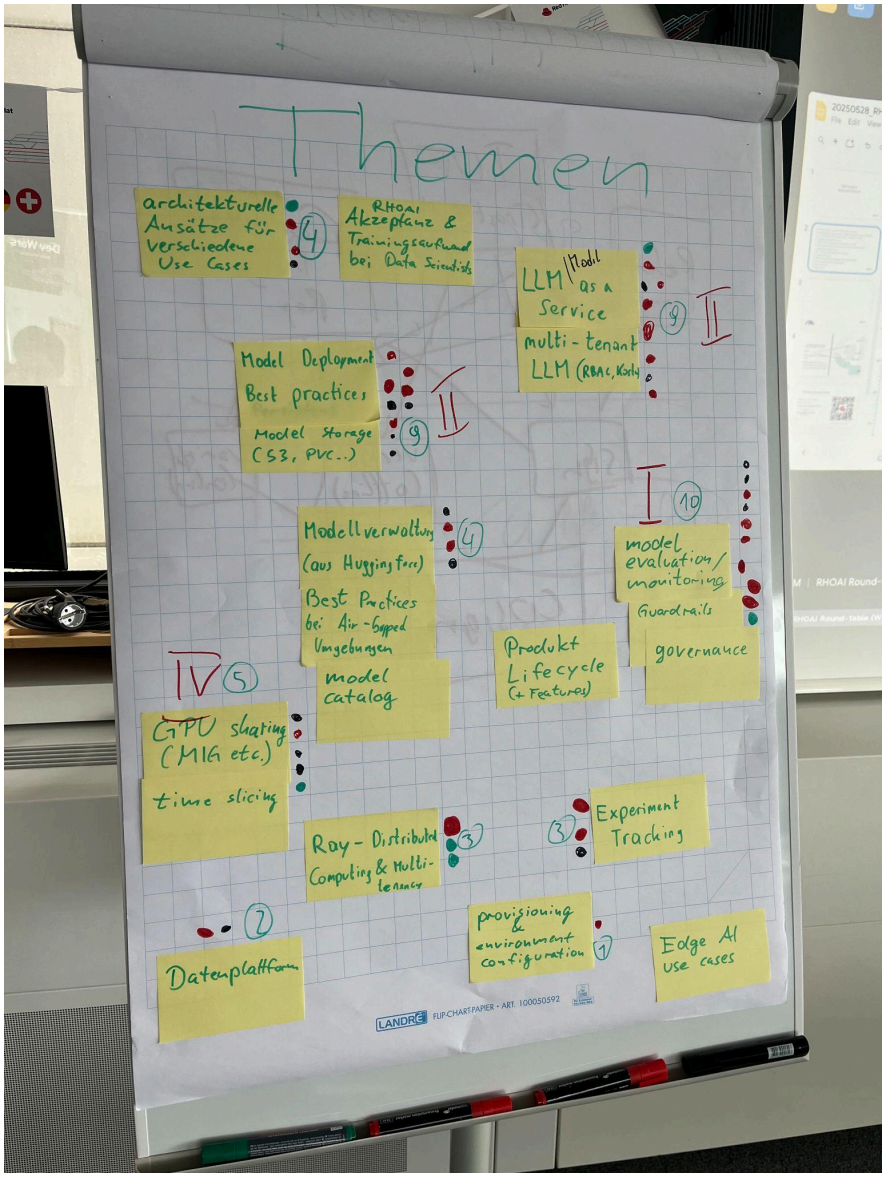
- Data Lab Hell
- BMI
- Wien Digital
- BRZ
- Infineon (Remote)
- PV
- NTS
- Fronius
- IBM Consulting

## Ablauf

1. Opening Attila Virag (Sales Lead, Red Hat Österreich)
2. Einleitung Stephan Kraft - Code of Conduct, Timeline
3. Vorstellungsrunde aller Teilnehmer
4. Themensammlung und Bewertung
5. Besprechung und Ideenaustausch zu den vier bestbewerteten Themen

# Themensammlung

Folgende Themen wurden von den Teilnehmern vorgeschlagen und im nächsten Schritt bewertet:




## Themen

- Model evaluation / monitoring, Guardrails, governance
- LLM  / Model as a Service multi- tenant, LLM (RBAC, Kosten)
- Model Deployment Best practices - Persistente Modelle, Storage (S3/PVC)
- Time Slicing, MIG (GPU Sharing)
- Modellverwaltung (aus Huggingface) Best Pactices bei Air-gapped Umgebungen, Model catalog

- Architektonische Ansätze für verschiedene Use Cases
- Ray - Distributed Computing & Multi-tenancy
- Experiment Tracking
- RHOAI Akzeptanz & Trainingsaufwand bei Data Scientists
- Provisioning & environment configuration
- Edge AI Use Cases
- Datenplattform
- Product - Features und Lifecycle

### Bestbewertete Themen

1. Model evaluation / monitoring, Guardrails, governance
2. LLM  / Model as a Service multi- tenant, LLM (RBAC, Kosten)
3. Model Deployment Best practices - Persistente Modelle, Storage (S3/PVC)
4. Time Slicing, MIG (GPU Sharing)

Die Teilnehmenden haben sich unter der Moderation von Red Hat mit den bestbewerteten Themen auseinandergesetzt.

## Thema 1: Model evaluation / monitoring, Guardrails, governance

Guardrails - was ist es, was wird benötigt, erste Erfahrungen

- Model davor schalten - für ethische Filterung, sensibel für öffentliche Themen
- Bei Azure mit Regler, aber als Blackbox
- Default Regeln, gemischte Erfahrung
- Bei internem Use Case - Qualität der Antworten soll nicht abfallen / passend sein, mit Daumen rauf / runter
- vLLM user namespace monitoring? - gibt's da was out of the box, etwas besseres
- Metriken über RHOAI, kserve, im OpenShift in observability
- Wird es was eigenes in RHOAI geben? -> Red Hat mit der Plattform agnostisch, Plan für Data Scientist reduzierte Sicht zu bieten
- Welche Metriken für Performance / Auslastung: Time to first token,...

RHOAI - Testen von Modellen

- TrustyAI zur Evaluierung von LLMs, existierende Fragenkataloge, in Zukunft mit eigenen Fragen auf einer multiple choice Basis, um die eigenen Anforderungen zu verfizieren - ja community basierend
- Reranking use case für Suche - training / optimierung außerhalb mit RAG?

Predictive AI

- Drift Detection mit Trusty AI
- Pipeline, Test laufen durch, Evaluationsfragen - Ground-Truth ermitteln

## Thema 2: LLM  / Model as a Service multi- tenant, LLM (RBAC, Kosten)

- Proxy Service
- Open Telemetry
- Quality of Service - Performance Garantie:
    - Rate limiting, wrapper API (python FastAPI, authentication und authorization, parameter für das gewünschte Model, inkl. Rate-limiting - unterschied PoC und Produktion)
    - Endpoints: Dev/Test, Prod,... ? Produktiv 2 Instanzen, separat für Load Balancing
- vLLM bezüglich Abrechnung unkompliziert
- S3 als externer eigener Service - YAML und PVCs werden selbst erstellt (außerhalb RHOAI als RWX)
- Öffentliches ChatGPT gesperrt -> hohe Anzahl User bis zu 6000 tgl.
- Management?
    - Eigener Shop für neue Use Cases - GPT for … Anpassungen, neue Dienste, erstmal zeitlich limitiert (30 Tage)
    - Über einen chat werden lifecycle und ähnliche Themen mit den Nutzern geteilt, wie z.B. Update Fenster und Verfügbarkeit
- Sprachmodelle und Predictive AI?
    - Telefon-Assistenz: kurze Antworten, Qualitätsfragen, Shadow ChatGPT Nutzung
    - Weiterentwicklung als Angebot -> lieber intern unterstützen, damit es im Haus auch mit sensiblen Daten funktioniert
- Nachts Batchläufe, um Verbesserungen herbeizuführen
- Dutzende von Applikationen - LLM, embedding oder spezifische Modelle -> viele User
    - Ray? GPUs im Ray Cluster?

## Thema 3: Model Deployment Best practices - Persistente Modelle, Storage (S3/PVC)

- Triton - komplizierter als vLLM, speziell mit verschiedenen GPU Modellen
- vLLM - spezifisch mit LLMs
- Andere Modelle sind mit Triton relevant auch als certified runtime für RHOAI
- Tekton Pipeplines mit docker / podman push
- Model registry verwenden
- Security scanning der Modelle? Quay (generisch)?
    - Sonatype (über Huggingface Mirror?) und JFrog Artifactory
- Trusted application pipeline als Lösung: aktuelles gap scanning des Modells!
- Huggingface - enterprise safety platform?
- Red Hat validated models angekündigt beim Red Hat Summit 2025

## Thema 4: Time Slicing, MIG (GPU Sharing)

- MIG und vLLM -> funktioniert mit single GPU - bei mehreren großen scheint es nicht zuverlässig / möglich zu sein
- Empfehlung bei großer GPU?
    - GPU Klassen?
    - RHOAI Accelerator profiles existieren bereits
- Time Slicing
    - MIG nicht immer möglich (speziell bei älteren GPUs) und daher eine Option, bisher kein Impact im Bezug auf Performance
    - Nur mit virtualisierten Nodes? - vGPUs!
- Hardware accelerators | OpenShift Container Platform | 4.18 | Red Hat Documentation

## Verschiedene Themen

Architektonische Ansätze für verschiedene Use Cases:

- PoC & Produktion - selbe HW oder nicht - bereits bei der vorherigen Diskussion angesprochen und aktuell noch am Anfang, das Richtung GPU teilen (MIG) / Time Slicing
    - Firewall?
    - Übers Frontend, public cloud zum Testen neuer Modelle, bevor es selbst zu betreiben

Fine Tuning von Foundation Models:

- Fehlt die HW
- Österreichische Amtssprache… wie. z.B. Pension / Rente, Stichtag Definition -> derzeitig Merging (EU AI Act?!)
- InstructLab von Red Hat bekannt?
    - Granite und Mistral aktuell

## Q&A

### References

- explainable AI mit Trusty AI: Github und RHOAI Doku Evaluierung von LLMs,
- llm-d: article
- GPU enablement: article
- verteiltes Training mit Ray: article
- ai-on-openshift.io Community Website
- OpenShift AI Produktdokumentation

## Red Hat Trainings und Zertifizierungen zum Thema AI

Hier findet ihr aktuell verfügbare Red Hat Trainings zum Thema AI. An einem MLOps Workshop mit dem Namen “ML500” wird aktuell gearbeitet. Wenn ernsthafte Interesse besteht, einfach mal bei einem der Red Hatter des Vertrauens bekanntgeben.

**Red Hat Artificial Intelligence (AI) Training und Zertifizierung:**

Kostenfreie Inhalte mit kurzen 2-3h Videos:

- Red Hat OpenShift AI Technical Overview | AI067
- Red Hat Enterprise Linux AI Technical Overview | AI096

Kostenpflichtige Kurse als 4-Tageskurs (onsite) oder 5-Tageskurs (on demand / online):

- Introduction to Python Programming and to Red Hat OpenShift AI | AI252
- Creating Machine Learning Models with Python and Red Hat OpenShift AI | AI253

Kostenpflichtige Kurse als 3-Tageskurs (onsite) oder 4-Tageskurs (on demand / online):

- Developing and Deploying AI/ML Applications on Red Hat OpenShift AI | AI267
- Red Hat Certified Specialist in OpenShift AI | EX267

Die Kurse können über eine bestehende Red Hat Learning Subscription abgerufen werden oder auch einzeln gebucht werden.