Mini Report: AI Lab Infrastructure & Strategy

Christiaan Verhoef

2025-06-01

Table of contents

1 Executive Summary

> "Progress isn't always visible — but it compounds."

During the week of May 26–30, the AI Lab made tangible progress toward enabling scalable, local-first AI adoption at Windesheim. Key efforts focused on three fronts: overcoming hardware limitations in student-facing LLM deployments (Mira), delivering a complete automated reporting prototype for internal use (Ronald), and aligning infrastructure direction through practical work and stakeholder conversations.

We successfully ran small-scale LLMs locally on a Windesheim-issued laptop using the MSTY stack and specialized RAGs, demonstrating that lightweight, modular AI tools are possible even under strict policy and hardware constraints. Although QEMU setup is still pending, initial VirtualBox testing confirmed the need for more efficient simulation methods.

Meanwhile, Ronald's reporting system was delivered using Python, Quarto, and Outlook, and received with enthusiasm. It supports repeatable, low-friction reporting for survey data, and illustrates how open-source tools can replace heavier enterprise dashboards when thoughtfully applied.

Finally, infrastructure planning continued through direct collaboration with Tommy and brief sessions with Maxime and Rea. While Nextcloud will not be adopted, a GitLab-hosted stack was confirmed for the upcoming semester. Student engagement also remained a focus, with a dedicated 3-hour session designed to reinforce motivation and ownership in applying AI responsibly.

Together, these activities advanced all four core objectives of the AI Lab initiative: scalable local-first AI, reproducible toolchains, human-centered design, and facilitative—not prescriptive—adoption of technology.

2 Weekly Focus

- Ran local LLMs on a Windesheim student laptop using **MSTY** and **specialized RAGs**
- Cleaned and optimized the machine to reduce memory use and enable model inference
- Attempted VirtualBox-based simulation for Mira's setup (too slow due to Hyper-V & Docker); QEMU pending
- Built and delivered **Ronald's automated reporting system** (Python + Quarto + Outlook)
- Met with Maxime & Rea to align on infrastructure (GitHub cleanup, GitLab hosting confirmed)
- Continued **server setup and deployment tasks** with Tommy
- Facilitated a **3-hour group coaching session** to support student motivation and AI ownership

---

3 Time Spent (Actual)

Task	Time Spent
Mira – LLM testing, MSTY + RAGs, laptop cleanup, VirtualBox	~1 day
Ronald – Automated reporting system	~2 full days
Infra: Server work with Tommy + Git strategy	~0.5 day
Students: Group coaching session	~0.5 day
Maxime & Rea discussion	< 0.25 day

Total Time Invested: ~**4.25 days**

4 Successes

- Successfully ran **local LLMs** on a Windesheim-issued student laptop using **MSTY**
- Used **specialized RAGs** to simulate context-aware behavior with minimal memory usage

- Cleaned and optimized the student laptop to allow LLM inference under tight hardware limits
- Built 3 modular Python scripts for Ronald's reporting flow:

  - `clean_data.py` – prepares raw survey data
  - `generate_report.py` – creates branded PDFs via Quarto + LaTeX
  - `send_email.py` – dispatches reports via local Outlook

- Delivered a **working automated reporting system** now ready for iterative improvements
- Ronald reviewed and approved the solution as **scalable, maintainable, and practical**
- Avoided unnecessary complexity by moving away from Grafana and Tableau
- Confirmed GitLab-hosted stack for AI and tooling infrastructure (replacing Nextcloud idea)
- Motivated students through a **live 3-hour group session** on responsible AI usage and self-directed experimentation
- Continued **technical server work with Tommy**, supporting future infrastructure expansion

5 Contribution to Overall Objectives

Objective	Progress This Week
1. Scalable, secure, local-first AI usage	Ran local LLMs using MSTY on student laptops Cleaned system for optimal performance QEMU simulation pending
2. Reproducible tools and templates	Delivered modular Python + Quarto reporting prototype for Ronald Created reusable LaTeX PDF template
3. Make technology a lever, not a barrier	Avoided over-engineered BI tools (Grafana/Tableau) Chose lean, open-source stack for automation
4. Facilitate—not impose—AI adoption	Supported Mira with constrained-friendly stack Ran a live 3-hour student coaching session Refined workshop strategy to modular, use-case tools

6 Deliverables

6.1 Local AI Stack (Mira)

- Deployed **local LLMs on Windesheim student laptop** using the MSTY stack

- Created and tested **specialized RAGs** to modularize context injection for small models
- Performed system-level cleanup and optimization to reduce memory footprint (~6.7GB baseline RAM usage now manageable)
- Initiated VirtualBox test for environment replication (too slow); **QEMU setup planned**

6.2 Reporting Pipeline (Ronald)

- Designed and implemented a **modular reporting prototype** using:
  - `clean_data.py`: cleans and structures CSV input
  - `generate_report.py`: renders branded Quarto + LaTeX PDF reports
  - `send_email.py`: dispatches PDF reports via local Outlook instance

- Delivered and demoed the tool to Ronald, who confirmed it meets his operational needs
- Validated the system for **scalability and future reuse**
- Defined follow-up tasks: visual polish + packaging into a deployable application

6.3 Infrastructure & Strategy

- Reached agreement to **host GitLab** internally for code and data handling
- Officially dropped **Nextcloud migration**, aligned with stakeholder constraints
- Supported **Tommy on server configuration**, laying groundwork for shared deployment infrastructure
- Initiated planning for **GitHub cleanup** and project consolidation

6.4 AI Enablement & Facilitation

- Conducted a **3-hour group session** with students to:
  - Motivate independent AI exploration
  - Contextualize tooling constraints
  - Encourage responsible, creative experimentation

- Adjusted workshop format: moving from full-stack demos to **small, focused, token-free AI tools**

7   Problems Tackled

7.1   Student Hardware Limitations

- Windesheim-issued laptops (~8GB RAM) are unable to run standard local LLM stacks
- The previously developed full-stack AI infrastructure could not even boot on these devices
- Required redesigning the approach: minimal models + extreme memory optimization

7.2   Data Handling Constraints

- Windesheim's policy prohibits transferring student data off-device (e.g., to the lab server)
- Mira preferred to remain fully compliant and avoid any workaround that might risk policy violations
- This eliminated many otherwise viable server-based solutions, forcing a local-only architecture

7.3   Simulation Barriers

- Attempted to replicate Mira's environment via VirtualBox, but performance was too poor due to Docker + Hyper-V conflicts
- QEMU setup is pending, limiting current ability to debug and simulate her setup remotely

7.4   Tooling Overhead

- Initial experiments with Grafana and Tableau were overly complex and inflexible
- These tools introduced unnecessary friction for Ronald's needs and did not support PDF/email automation cleanly

7.5   Platform Misalignment

- Stakeholders (Maxime & Rea) confirmed that while they're excited about the lab's tooling direction, they cannot migrate to open-source platforms like Nextcloud unless the entire lectorate moves with them
- Required reframing infra strategy: align with existing ecosystems, and build modular tools that complement—not replace—core platforms

7.6 Motivation & Ownership Gaps

- Students showed hesitation around advanced AI tools and did not always feel ownership over digital workflows
- Addressed through direct engagement and facilitative teaching (3-hour group session focused on confidence and accessibility)

8 Reflections & Insights

- **Lightweight AI is a necessity, not a nice-to-have**:
  The constraints of Windesheim's student laptops and data policies are not temporary hurdles—they define the baseline. MSTY and modular RAGs are not compromises, but enablers of access.

- **"It works" is not enough — it must work here**:
  The gap between server-based success and local usability is larger than expected. What runs in a lab isn't automatically usable in the classroom. Context-aware, resource-sensitive tooling is essential.

- **Simplicity is a feature**:
  Grafana and Tableau provided bells and whistles, but Quarto delivered what Ronald actually needed — with far less friction. Tool choice should always reflect context, not trend.

- **People support what they help shape**:
  Mira's insistence on staying within institutional boundaries led to more ethical, thoughtful design. Student conversations reminded me that confidence and inclusion matter as much as code.

- **Build for modularity, not just delivery**:
  The shift toward small, task-specific tools isn't just more scalable — it's more teachable, maintainable, and adaptable for different user skill levels.

- **We need space to simulate constraints**:
  Without working virtualization (QEMU), it's difficult to predict whether our solutions will succeed. The ability to replicate and test within constrained conditions is now a core priority.

- **Infrastructure is as much about relationships as code**:
  Conversations with Tommy, Maxime, and Rea were just as important as any config file — they shape what is possible, and what will be supported long-term.

9 Next Steps

9.1 Mira & Local AI Tooling

☐ Complete **QEMU setup** to fully simulate student laptop constraints

☐ Create a documented baseline for MSTY-based setups on underpowered devices

☐ Evaluate additional lightweight models or distillation strategies

☐ Provide feedback loop to determine which tools Mira can realistically deploy

9.2 Ronald's Reporting Pipeline

☐ Improve visual design of generated PDFs (Quarto + LaTeX styling)

☐ Wrap reporting scripts into a **deployable app or CLI tool**

☐ Write basic user documentation for non-technical handover

☐ Explore optional hooks for Supabase or direct survey integration

9.3 Git Strategy & Infrastructure

☐ Clean up GitHub (Maxime & VCH repositories)

☐ Begin GitLab deployment on internal server

☐ Consolidate Obsidian files across Windesheim & project spaces

☐ Document infrastructure map for the AI stack (OpenWebUI, n8n, pgvector, etc.)

9.4 Workshop & Student Tools

☐ Define **first set of modular, use-case-based AI tools** for student pilots

☐ Link workshop formats to real classroom/staff needs

☐ Allocate tokens for development only — not general student use

☐ Develop example workflows that require **no server or cloud dependency**

9.5   Support & Engagement

☐ Finalize and distribute **automation support survey** for Michiel

☐ Prepare summary of insights from last student group session

☐ Check in with Fred for curriculum planning alignment

☐ Continue hands-on support for Mira + student-led AI projects

Absolutely — let's **refactor and improve your agenda for June 2–7** so it aligns more tightly with your next steps. The goal is to structure the week **around deliverables**, **open tasks**, and your **Windesheim work rhythm** (Mon, Tue, Thu = Windesheim | Wed = Pangea | Fri–Sun = wrap-up + personal).

---

10   Weekly Agenda: June 2–7

10.1   Monday (Windesheim) — *Build + Handover*

Tags: `#dev` `#admin` `#infra` - [ ] GitHub cleanup for Maxime (start consolidation) `#infra` - [ ] Install Quarto prototype with Ronald `#dev` - [ ] Polish visuals for reporting PDFs `#dev` - [ ] Update Michiel's reports `#admin` - [ ] Attend Windesheim + VCH standups `#collab`

10.2   Tuesday (Windesheim) — *LLM Lab & Infra Day*

Tags: `#dev` `#infra` `#educ` - [ ] Support Mira: test MSTY + micro-RAG setup `#dev` - [ ] Begin QEMU installation for replication `#dev` - [ ] Obsidian workspace consolidation `#infra` - [ ] Continue GitHub cleanup (if needed) `#infra` - [ ] Test Beer Game setup `#educ`

10.3   Wednesday (Pangea) — *Strategy & Planning*

Tags: `#admin` `#collab` `#educ` - [ ] Attend Pangea board meeting `#collab` - [ ] Draft automation support survey (Michiel) `#admin` - [ ] Clean GPT chats + archive last week's report `#admin` - [ ] Outline modular AI workshop formats `#educ` - [ ] Optional: test QEMU if time allows `#dev`

10.4   Thursday (Windesheim) — *Student & Collaboration*

Tags: `#collab` `#educ` `#infra` - [ ] VCH student progress meeting `#collab` - [ ] Meet with Fred: align curriculum `#collab` `#educ` - [ ] Refine use-case-based modular AI tool concepts `#educ` `#dev` - [ ] Finalize consolidated Obsidian workspace `#infra`

10.5   Friday — *Documentation & Activation*

Tags: `#admin` `#dev` `#collab` - [ ] Attend Beer Game final meeting `#educ` - [ ] Start automation interview draft `#admin` - [ ] Write Ronald prototype handover notes `#admin` `#collab` - [ ] Publish this week's mini-report `#admin`

10.6   Weekend (Sat–Sun) — *Catch-up & Cleanup*

Tags: `#admin` `#infra` `#educ` - [ ] Finalize + send Michiel's automation survey `#admin` - [ ] Prepare workshop templates & student guide drafts `#educ` - [ ] Complete GPT chat cleanup & server backups `#admin` `#infra` - [ ] Write GitLab deployment playbook `#infra` - [ ] Rest + reflect