Serving Inference
UNIBE Survey Results

GenAI Infrastructure Access, Usage, and Satisfaction Analysis

WHAT AND WHY

- running GenAI independently of commercial providers
- use Open Source/Open Weights LLMs/VLMs for sensitive data
- rely on cost-efficient/local approaches

→ digitally sovereign

- HPCs exist but are not meant to serve inference constantly

METHOD

→ run a survey, collecting insights from stakeholders

ORGANIZATIONAL DEMOGRAPHICS

n = 36

University Sector

89%

Majority of respondents are from Universities

Location

100%

Respondents are based in Europe

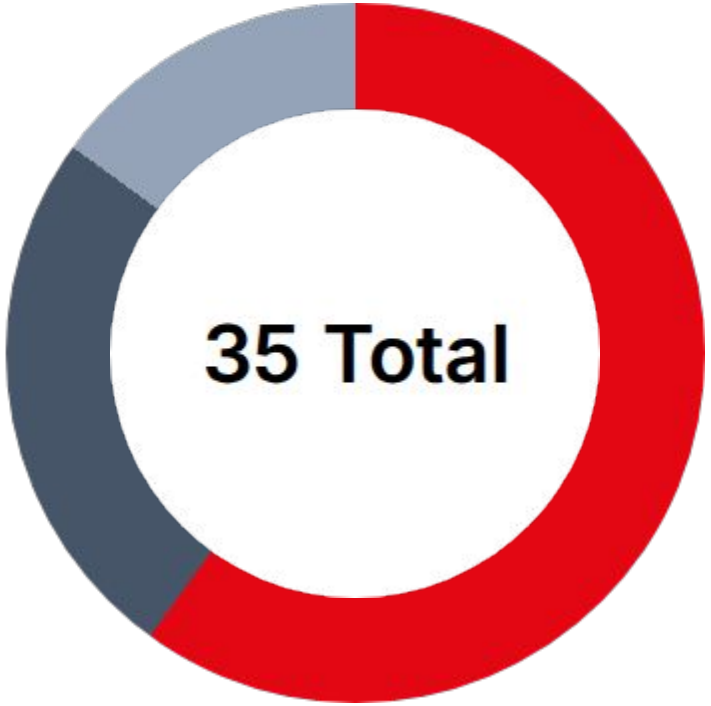
INFRASTRUCTURE ACCESSIBILITY

UNIVERSITÄT BERN

Types of infrastructure respondents have access to (Multiple Selection):

Total Respondents: 35. High Performance Computing (HPC) remains the primary accessible resource for AI deployment.

WHO PROVIDES THESE SERVICES?



- University (28 responses)
- Lab/Research Group (12 responses)
- State/Agency (7 responses)

INFRASTRUCTURE UTILIZATION

UNIVERSITÄT BERN

80%

Current Usage Rate

The vast majority of researchers with access are actively utilizing the provided infrastructure for generative AI tasks.

Usage Status Breakdown

Active Users (Yes) **28**

Non-Users (No) **4**

Interested ("I would if I could") **3**

PRIMARY GENAI TASKS

OTHER GENAI TASKS

UNIVERSITÄT BERN

“I do not use AI for any task I do. I have a brain and can do those tasks myself.” **Respondent**

Alternative Tasks

- image and sound restoration
- speech to text
- translation
- annotate datasets
- music generation

FREQUENCY OF INFRASTRUCTURE USE

UNIVERSITÄT BERN

Weekly

11

Respondents

Irregularly

11

Respondents

Daily

8

Respondents

Monthly

2

Respondents

CLOSED SOURCE ADOPTION

UNIVERSITÄT BERN

83%

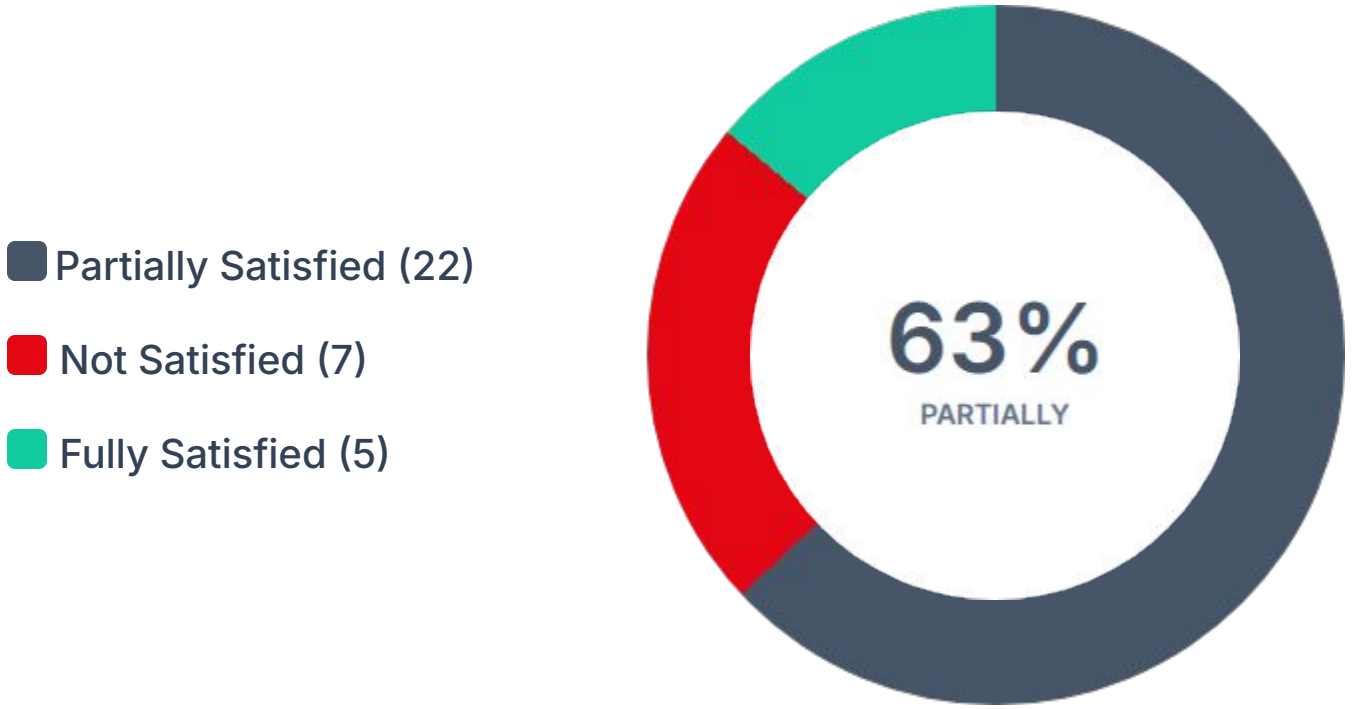
Use Commercial Models

29 out of 35 respondents use commercial/closed source models besides open source weights.

Key Observations

- Commercial providers outperform open-weights models in performance and convenience.
- Commercial models are often used for non-sensitive tasks due to better access and output quality.
- Infrastructure is preferred when data privacy or institute-only storage is required.

INFRASTRUCTURE SATISFACTION



IDENTIFIED PAIN POINTS

- **Performance Gap:** Open-weights models often lack the "near-frontier" performance of commercial counterparts.
- **Technical Barriers:** Complexity of infrastructure (e.g., SLURM) and lack of intuitive GUIs for non-tech-savvy users.
- **Resource Constraints:** Need for more high-end GPU nodes and larger memory (VRAM) for running larger models.
- **Service Stability:** Reports of models being removed without explanation or lack of long-term support for tools.
- **Management:** Missing features for resource assignment, monitoring, and statistics.

THE CHALLENGE OF SCALE

"To approach near-frontier-model performance, one needs to run the largest open-weights models and update them continuously, which is often not feasible." **Respondent**

VOICES: IT & STRATEGY DISCONNECT

"I'm IT Supporter, but I'm generally not able to provide a satisfying AI environment to my people (financial and organisational reasons)"

"The University wants digitalisation but can't offer it by itself... So probably too many small projects instead of a few big ones?"

VOICES: USABILITY & STABILITY

"Models I have been using one day are being taken away over night without an explanation... And I don't have the option to fine tune an existing model."

"More large GPUs. More training or support. E.g. I'm not super proficient with SLURM and I think I could configure experiments better with more knowledge of that."

VOICES: THE IDEAL FUTURE

UNIVERSITÄT BERN

> *I want to host own models and make them available to external users (and services) via some standard gateway provided by our infrastructure.*
>
> **Respondent**

UNIVERSITÄT BERN

Thank you

Questions?

Serving Inference – UNIBE Survey Results

Digital Humanities | University of Bern