EPiC Series in Computing

Volume 107, 2025, Pages 108–116

Proceedings of EUNIS 2025 annual congress in Belfast

Cost Management for the Use of Generative AI in Higher Education: A Case Study from North Rhine-Westphalia

Bernd Decker[1] and Denise Dittrich[1]
[1] RWTH Aachen University, Germany
decker@itc.rwth-aachen.de, dittrich@itc.rwth-aachen.de

Abstract

Generative Artificial Intelligence (AI) models, like systems such as ChatGPT, have catalyzed a transformative shift in higher education, prompting universities to explore innovative applications across research, teaching, and administration. This paper discusses the KI:connect.nrw project, which aims to provide universities in North Rhine-Westphalia (NRW), Germany, with equitable access to commercial and open-source AI services while addressing the challenges associated with cost management. The project introduces a centralized web interface that facilitates user-friendly access to multiple AI systems through API integration. Key objectives include transparent cost allocation tailored to specific user groups or projects and flexible budget control mechanisms that accommodate diverse funding sources within participating HEI. The paper further examines the complexities of billing methods, ranging from account-based to token-based models, and their implications for the central provider and using universities. Initial experiences at RWTH Aachen University highlight the project's operational framework and its potential impact on managing AI-related costs effectively as usage among employees and students increases.

1 Introduction

Since the introduction of ChatGPT in November 2022 and the subsequent discussions on the impact of Artificial Intelligence (AI) on research, teaching, and administration, it has become evident that this technology will have a lasting influence on all areas of higher education. The demand for powerful AI solutions at universities has significantly increased since then.

The availability of increasingly powerful large language models with extensive functionalities, as utilized in systems like OpenAI's ChatGPT, Google Gemini, or Claude by Anthropic, has become almost indispensable for critical and innovation-promoting tasks at universities. At the same time the implementation and utilization of generative AI models pose significant challenges. There are ongoing concerns regarding the reuse of prompts, uploaded data or even personal user data for training or other purposes. Concurrently, open-source AI services provide an opportunity to implement generative AI applications with full data protection control within universities.

The rapid technological development and integration of these technologies into diverse university processes are complex fields of endeavor.

Additionally, there is uncertainty regarding the cost structure of such implementations. This refers to many aspects of cost management: the basis for calculating the costs (e.g. tokens, not words) and the resulting lack of transparency for the user, the different costs for different AI models (including open source models), and ultimately also to the distribution of costs among different users or projects.

1.1 Project KI:connect.nrw

The KI:connect.nrw project aims to provide universities in the state of North Rhine-Westphalia (NRW) in Germany with access to commercial and open source AI services.

This will be achieved through a model that enables individual cost allocation, budgeting and offers flexible and customizable options for budget control while maintaining data-friendly settings. Different language models like ChatGPT-4o can be integrated via API.

The model can be used via a centrally provided web interface which is deployed and developed at the RWTH Aachen University, or via self-deployment of the web interface using the code provided under open-source license.

The main goals of the project are:

- providing user-friendly access to multiple AI systems: one central web interface for various generative AI services and access via a single programming interface
- data-friendly settings: The additional layer of abstraction ensures that no personal user data is transferred to the AI models, other than the prompts.
- centralized operation & development of the webinterface
- flexible cost allocation and budget control: individual cost allocation for universities as well as flexible and customizable budget control options to provide different user groups with individual offers

In the following sections we explain details regarding the cost allocation and budget control.

2 Challenges in cost management with different AI models

The project KI:connect provides a central user interface for all 41 HEI in NRW. Concerning cost management the following aspects were addressed:

- billing method
- cost allocation to a specific university and different user groups or projects
- settlement of costs
- cost control regarding fixed budgets for different user groups

2.1 Billing method

The billing method is closely related to the licensing model. Most commercial AI models offer two types of billing models, each with its own challenges:

- Billing on an account basis: A fixed sum is charged for each user in an institution for a specified period of time, usually on a monthly basis. The sum does not depend on actual usage of the AI model and usually refers only to one vendor (e.g. OpenAI). With typical costs around €10-20 per user even smaller universities generate considerable total costs that usually significantly exceed the university budget. Without a university-wide license, license and user management for all 41 HEI in one interface is a big challenge including monitoring the number of licensed users, license lifecycle management and allocating licenses to users. Furthermore, account-based billing may require the transfer of personal user data to the vendor.
- Billing on a token basis: a token is typically referred to as the "the smallest unit into which text data can be broken down for an AI model to process" (Lighton - The Magic of Tokens in Generative AI: A Deep Dive). Depending on the language one token corresponds to approx. 0.65 to 0.75 words. Some manufacturers differentiate between costs for tokens sent by users and the response tokens generated by the AI model. The cost per token varies between different models. In this billing method only the actual usage of the correspondent language model is billed, though the actual costs for a prompt are not intuitively understandable for the user.

As the goal of KI:connect is to integrate different AI models for all university users, regardless of the extent of use, token based billing is used here. Another benefit is that this billing method is usually offered via API access which can easily be integrated in the central web interface.

The costs for open-source language models are of course depending on the costs of the hardware used. Nevertheless, similar billing metrics can be used here as with commercial models.

In NRW, there will be centrally offered open-source models in the future, which will be provided by another funded project. In the meantime, some universities already use open source models provided by other community members, like universities from other states in Germany.

2.2 Cost allocation

A central requirement is the transparent and causation-based allocation of incurred costs. It must be ensured that all costs can be unequivocally attributed to the respective cost drivers, such as a specific user group, a particular research project or a university on a bigger scale.

Challenges arise from data protection regulations: personal data often cannot be transmitted to commercial providers, making the direct identification of individual users for cost allocation purposes impossible. Moreover, commercial providers generally lack the ability to differentiate between various user groups within an institution - such as employees, students, or research projects.

To address these requirements, cost allocation is implemented within the KI:connect system. User authentication is carried out through Shibboleth, which transmits defined metadata of this user from the associated university without violating data protection regulations. Based on this information, KI:connect automatically assigns the underlying access to the corresponding university and its predefined university groups. Consequently, the incurred costs can be recorded internally in a cause-related manner and later reported on an aggregated basis—by, for example, cost centers or projects—without having to forward any personal data to the commercial provider.

A prerequisite is that each university one the one hand defines the set of projects or user groups needed for billing and on the other hand is able to transfer this data through the Shibboleth interface.

2.3 Settlement of costs

The KI:connect projects provides central funding for the development and central operation of the web interface and corresponding backend, but not for using AI models. Consequently, the costs for using commercial AI models must be borne by the universities themselves. Taken together with the centralized interface, this may become a challenge.

Using the cost allocation possibilities explained before one option is to settle the costs for each university. This would mean that RWTH Aachen University, as the central provider, would have to pay the costs in advance and invoice them afterwards. Considering the high number of possible users (~160.000 employees (Federal Statistical Office of Germany - Institutions of higher education by state) and ~700.000 students (Department for statistics NRW - students in HEI)) and the resulting possible costs as a central provider it is not possible for one university to pay in advance for all 41 HEI in NRW.

Furthermore, this may pose another challenge regarding the procurement of the used commercial AI models. This would have to be done centrally by one university as well while regarding legal regulations and the (maybe different) wishes and requirements of all participating universities.

The solution here is that each university is responsible for providing the API access to the desired AI model themselves. This means that the procurement process and maybe even the administration or configuration of the API access (e.g. using OpenAI on Microsoft Azure) has to be done by the respective universities.

In the central interface the API access is added and assigned to the respective university, ensuring that only users with the corresponding metadata can use this API.

2.4 Cost control

To ensure that KI:connect is integrated as optimally as possible into the university's processes, various sources of funding also need to be considered. Additionally there may be budgets corresponding to the cost centers or limited funding for specific projects.

For the KI:connect system this poses the following requirements:

- limitations: there has to be a technical limitation so that costs do not get out of control. As the billing method refers to prompts, a logical implementation is to restrict the number of prompts per time (hour, day) that a user of a certain group can make. Hereby different costs for different AI models could be referred to by setting different prompt limits for them.
- regarding different budgets for different user groups and projects: if a university has different sources of funding for specific user groups or projects, this can be considered in two different ways. One way is to use the cost allocation methods to internally settle the costs to the corresponding cost center. The other possibility is using different API accesses for different users. Thereby the billing is relocated to the vendor.

3 Implementation

3.1 Overview of the System Architecture

The KI:connect web interface serves as the central point of contact for authorized users. Among other things, it offers features such as a selection of various LLMs, a chat history, and the ability to upload documents. The web interface forwards incoming prompts to the selected AI models and displays their responses in the browser. User authentication is handled via Shibboleth, ensuring that authorization remains within the respective universities and that no personal user data is transmitted to external providers. Only those metadata required for budgeting and cost accounting are only used within KI:connect system.

The backend acts as a central control component that receives user queries - known as prompts - processes them, and forwards them to the connected LLMs. The LLM's response is then prepared and displayed in the user interface, while the backend logs the tokens and costs incurred, assigning them to the appropriate project, institute, or user account.

By adopting this "orchestration approach," the backend can process any request centrally and support multiple AI services without requiring each university to implement its own frontend or accounting logic. Figure 1 illustrates, by way of example, how prompts are relayed from the web interface to the chosen model and how the AI's output is then sent back to the frontend.

In designing the backend architecture, special attention was paid to supporting both commercially and locally hosted LLMs while allowing all other components to be operated on-premises. Thanks to the modular design, additional AI services can be integrated in the future via appropriate connectors or APIs without necessitating any fundamental changes to the existing infrastructure.

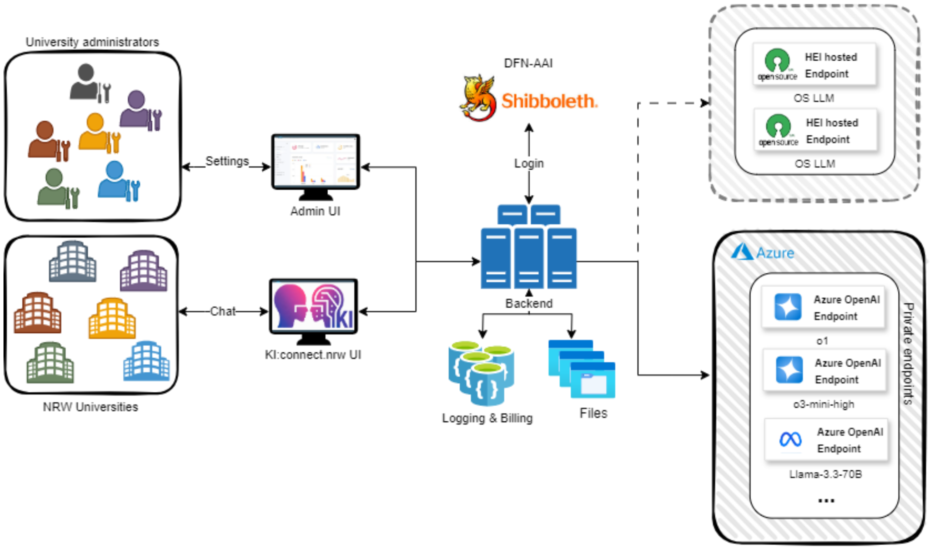

Figure 1 KI:connect system architecture

3.2 Billing Model and Cost Management

In order to transparently represent the costs incurred by using generative AI services and to enable the flexible distribution of expenditures across various universities or projects, KI:connect employs a token-based billing system. Each request (prompt) is tracked in terms of the tokens processed and is charged via the corresponding APIs. This approach ensures that costs are allocated according to actual usage, allowing user groups to pay only for the resources they genuinely consume rather than relying on rigid flat-rate licenses.

From a technical standpoint, KI:connect's central backend processes all incoming prompts, determines the inbound and outbound tokens, and automatically assigns them to individual universities, institutes, or projects. Billing is based on the respective API cost structures of the connected AI services and is continuously logged. Using this data, reports can be generated to provide project managers and university administrators with an overview of current token consumption and associated costs. Additionally, these reports support proactive resource planning by enabling early budget adjustments or the regulation of usage frequency.

For cross-university and cross-project allocation, only the metadata required to associate tokens and costs is retrieved from the Shibboleth authentication process. No personal information is transferred to external providers, thus ensuring compliance with data protection regulations. Due to this flexible architecture, multiple universities can simultaneously use KI:connect and store various cost or project accounts without necessitating additional system components.

To avoid cost-intensive utilization or uncontrolled budget overruns, KI:connect allows for the definition of individual thresholds. Once a specified limit is reached, further requests are blocked - a particularly relevant feature for large user groups such as students or research teams. In this way, the system promotes the sustainable and transparent use of generative AI services and ensures that each institution or project remains accountable solely for those costs that arise from its own requests.

4 Experiences at RWTH Aachen University

At RWTH Aachen University the KI:connect system has been offered since July 2024. With around 10.000 employees as potential users the actual use and thus system load and costs were not predictable Therefore KI:connect started with gradual access for small groups of employees. Due to the comprehensive cost management features we were able to slowly increase the number of users adapted to the factual usage until in November all employees were granted access without restrictions. In figure 2 you can see the gradual growth.

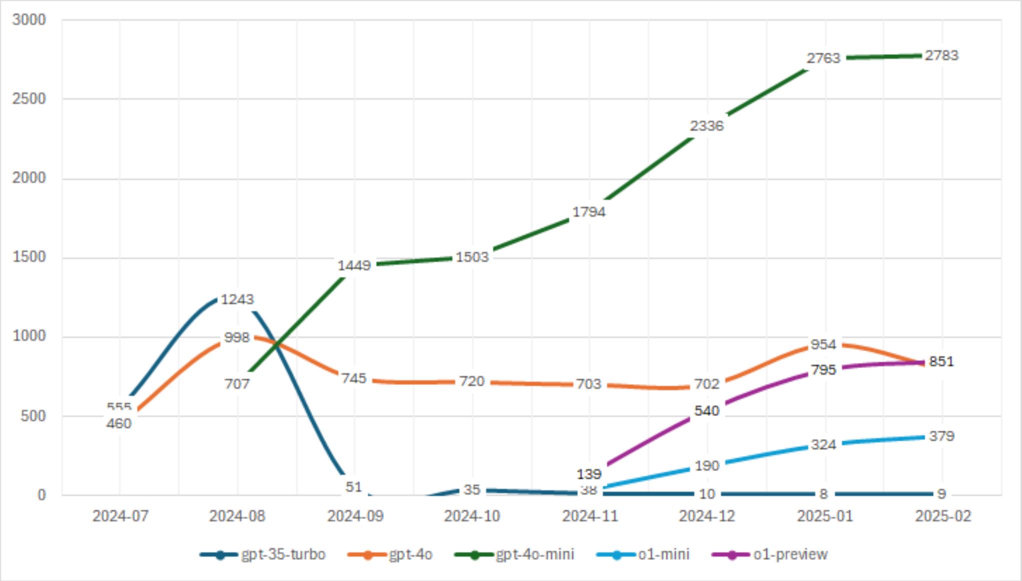

Figure 2 Number of RWTH employees per LLM

In December 2024 additional access for the students was added. To differentiate costs between the groups and due to different sources of funding separate API access are used here. Thanks to the experiences made with employee cost management we were able to grant all students access at once.

Right now, the following LLMs are available:

- GPT-4o
- GPT-4o mini
- GPT-o1-preview
- GPT-o1 (only employees)
- GPT-o3-mini-high
- GPT-o3-mini
- Llama-3.3-70B (available soon)

Each model has its own limits per token / hour according to the costs and the prompting behavior of the user group.

We were able to observe an increasing usage especially for the o1-preview model during the beginning of February. As the usage was rising, so were the belonging costs for the employees (see figure 3). By mid-February figures showed a forecast for this user group to hit the total budget set by the University by the end of the month. This made it necessary to adjust the limitations for the allowed tokens from 25 per user / hour to 15 per user / 3 hours.

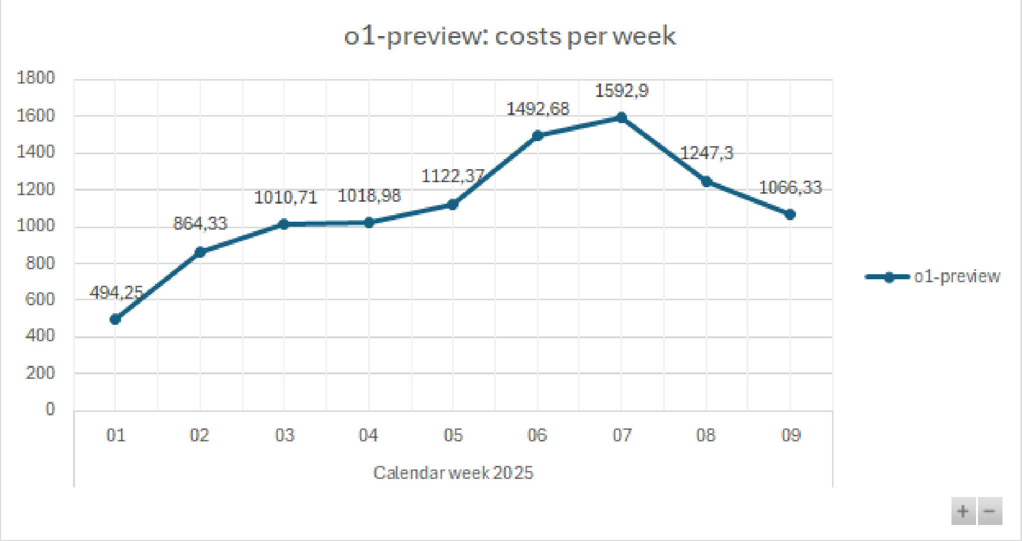

Figure 3 costs for o1-preview (2025)

Right now only commercial models are available for RWTH Aachen University, but the methods for cost control are also applicable to open source models, as the following example shows. We are offered access to an open source model operated by another University in Germany. The access is granted via API and thus integrateable into the KI:connect interface to make it accessible for specific user groups. There is a total limit of 8000 prompts / day for this API access. Since KI:connect is able to monitor both token and prompts consumed, we can set prompt limits for each user as well as a total limit / day.

We are currently investigating whether an open source model operated in the public cloud would be a good addition to the commercial LLMs. This will be an additional challenge in terms of cost management as hosted infrastructure for the model has to be translated into known methods and limits for the users.

5 References

Department for statistics NRW - students in HEI. (2025, 03 04). Retrieved from https://statistik.nrw/gesellschaft-und-staat/bildung-und-kultur/hochschulen/studierende-nach-hochschularten

Federal Statistical Office of Germany - Institutions of higher education by state. (2025, 03 04). Retrieved from https://www.destatis.de/EN/Themes/Society-Environment/Education-Research-Culture/Institutions-Higher-Education/Tables/total-states-further-indicated.html

Lighton - The Magic of Tokens in Generative AI: A Deep Dive. (2025, 03 04). Retrieved from https://www.lighton.ai/lighton-blogs/the-magic-of-tokens-in-generative-ai-a-deep-dive#

6 Author biographies

Bernd Decker is deputy head of the Department "Process Management and Digitalization in Learning & Teaching" at the IT Center of RWTH Aachen University since 2011. From 2006 to 2009, he worked at the IT Center as a Software Developer, and since 2009 he is leading the development group. His work focuses on IT solutions for processes in the fields of Learning Management Systems, E-Services, and Generative AI. (CRediT: Conceptualization, Software, Writing – original draft)

Denise Dittrich has been working at the RWTH Aachen University's IT Center since 2005. She received her Master Degree in Artificial Intelligence from Maastricht University in 2009. Since 2016 she is deputy head of the department for Systems&Operation. With a background in providing large-scale central services like E-Learning, Identity Management and Collaboration platforms she now leads the university's cloud coordination. Within EUNIS she leads the SIG Cloud Management. (CRediT: Conceptualization, Writing – original draft)