

AI Offers in Academic Cloud

AI Offers in Academic Cloud



Hosting

Open Weight Models

Hosted by GWDG in Göttingen

Cost Model

Prepaid

Flexible

Free

Variants

Full Service or API-Only



Hosting

commercial Models

Hosted by external cloud provider

Cost Model

Prepaid

Flexible

Variants

Full Service or API-Only

AI Service Cost Models

ChatAI

Free

access for testing

Einschränkungen

reduced list of AI-models

no access to Image AI, Voice AI

best-effort availability

reduced request-rate-limits

only individual API-Keys

no user management

ChatAI & Externe Modelle

Prepaid

fixed budget

ChatAI & Externe Modelle

Flexible

pay-as-you-go

Billing

annual payment

monthly billing

billed token-exact

billed token-exac

unused budget carries over to next month

Billing starts when consumption exceeds 1.000 €

Budgets & Limits

default budget for all organization members

default budget for all organization members

Individual and group budgets override the organization-wide default

Individual and group budgets override the organization-wide default

Internal Models – Cost Model



Base value: compute time

Compute time for a request

- Depends on compute time per token (performance)
- Depends on number of Input-/Output-Tokens

Costs for the compute time

- Depends on system / GPU
- GPU seconds as base unit
- » comput time per token (I/O) x GPU time cost

Internal Models - Calculation



Factors in Full-Cost Accounting

- All information is based on standards and empirical data from operational practice.
- Percentage figures are specific to each GPU.
- Typical values are provided here as examples.

(74%) Infrastructure

- (66%) Hardware Investment
- (8%) Maintenance, Network, Racks
- (26%) Energy costs (incl. Cooling)
- Downtime factor

(17%) Personel

- (67%) Devs, Sys-Admins, Consulting, Sales, Support
- (33%) Personell overhead

(9%) Sonstige Kosten

- Administrative-Overhead

Calculation in GPU/s

Infrastructure costs specific to GPU-Typ

Personell costs distributed on GPUs

Cluster of 300 GPUs



A100	0,0005991 €/s
H100	0,0007408 €/s
H200	0,0008847 €/s
B200	0,0013898 €/s

Internal Models - Pricing



Token as consumption unit	
Example - Qwen 3.6 35b	
GPU x amount:	H100 x 1 » 0,0007408 €/s
Time / Input-Token:	0,000162604 s
Time / Output-Token:	0,001648210 s
Calculation Input:	$0,0007408 \text{ €/s} * 1 \text{ M} * 0,000162604 \text{ s}$
Calculation Output:	$0,0007408 \text{ €/s} * 1 \text{ M} * 0,001648210 \text{ s}$
Price Input Token:	0,11 € (pro 1 M Token)
Price Output: Token	1,11 € (pro 1 M Token)

Standard Budgets for Prepaid



AC Kontingent	AC Description*	Prepaid €/year
AC 1.3	Für bis zu 30 Nutzende	1.000,-
AC 2.1	Für bis zu 100 Nutzende	2.000,-
AC 2.3	Für bis zu 300 Nutzende	3.750,-
AC 3.1	Für bis zu 1.000 Nutzende	7.500,-
AC 3.3	Für bis zu 3.000 Nutzende	15.000,-
AC Campus	Für bis zu 10.000 Nutzende	30.000,-
AC Campus Plus	Für bis zu 30.000 Nutzende	60.000,-
AC Campus Max	Für bis zu 60.000 Nutzende	100.000,-

*Notice for AC Zuordnung

The quota model is required for a service to be included in the AC.

The prices listed are based on a cost model that uses a mixed costing approach and reflects empirical data on service usage.

For ChatAI, this is based on GWDG's experience with the hundreds of thousands of users who currently use ChatAI. Actual usage may vary for individual universities.

Internal Models - Prepaid

Prepaid Budgeting

Kontingent

Fixed Budgets from Academic Cloud Kontingenten

Optional: Budget-Top-Off possible

Calculating organization budgets

Annual = Kontingent-Fixed amount + 20%

Monthly = Annual / 12

Daily = Monthly / 15 (not 30 to maintain flexibility)

Implementation

Budget enforced as Budget-Rate-Limits (see following Slides)

Carry-over of unused budget to next month



Example: AC Campus+

Kontingent

Annual plan:	60.000,- €
--------------	------------

Calculating organization budgets

Annual budget:	72.000,- €
----------------	------------

Monthly budget:	6.000,- €
-----------------	-----------

Daily budget:	400,- €
---------------	---------

Consumption and carry-over

Consumption this month:	5.000,- €
-------------------------	-----------

Budgt next month:	7.000,- €
-------------------	-----------

External Models - Prepaid

Prepaid Budgeting

Calculating organization budgets

Annual = Kontingent-Fixed Amount (**No bonus here**)

Monthly = Annual / 12

Daily = Monthly / 15 (not 30 to maintain flexibility)

Kontingents » identical to Chat AI prepaid

Implementation » identical to Chat AI prepaid

ChatAI Externe Modelle

Example: AC Campus+

Kontingent

Annual Payment:	60.000,- €
-----------------	------------

Calculating organization budgets

Annual:	60.000,- €
---------	------------

Monthly:	5.000,- €
----------	-----------

Daily:	333,33- €
--------	-----------

Consumption and carry-over

Consumption this montht:	4.500,- €
--------------------------	-----------

Budgt next month:	5.500,- €
-------------------	-----------

Consumption Control



Budget Control

Budget-Rate-Limits

Limits consumption in € per time

Hour, day, month

Adjustment via IDM by organization admins

For user groupes

For individual accounts

Request-Limits

Overload Protection

To avoid noverload there is a Request-Limit in place

- Limits requests per time
- 10s-window and hourly

Calculation

For ChatAI based on statistics:

minimal RL = 10/10s (!)

- 10s Request Limit = **Daily budget** / 0,01 € / 360
- Example Daily budget 400,- €
10s-Rate-Limit = 111 requests

For ChatAI ExterneModelle

- Default 10s Rate Limit = 100, or contact support@gwdg.de

Cost Model Flexibel



Flexible Budgeting

Define Organizations-Limits

- Monthly Hard-Limit (recommended)
- Monthly Soft-Limit (Alert limit)

Billing

- Billing starts at consumption of min. 1.000,- €
- Monthly billing over this amount
- Billing is
 - Usage last month
(or 1.000,- € when hitting the minimum)
 - Plus usage vor next month based on last month as advanced pay
- Carryover of credit to the following month