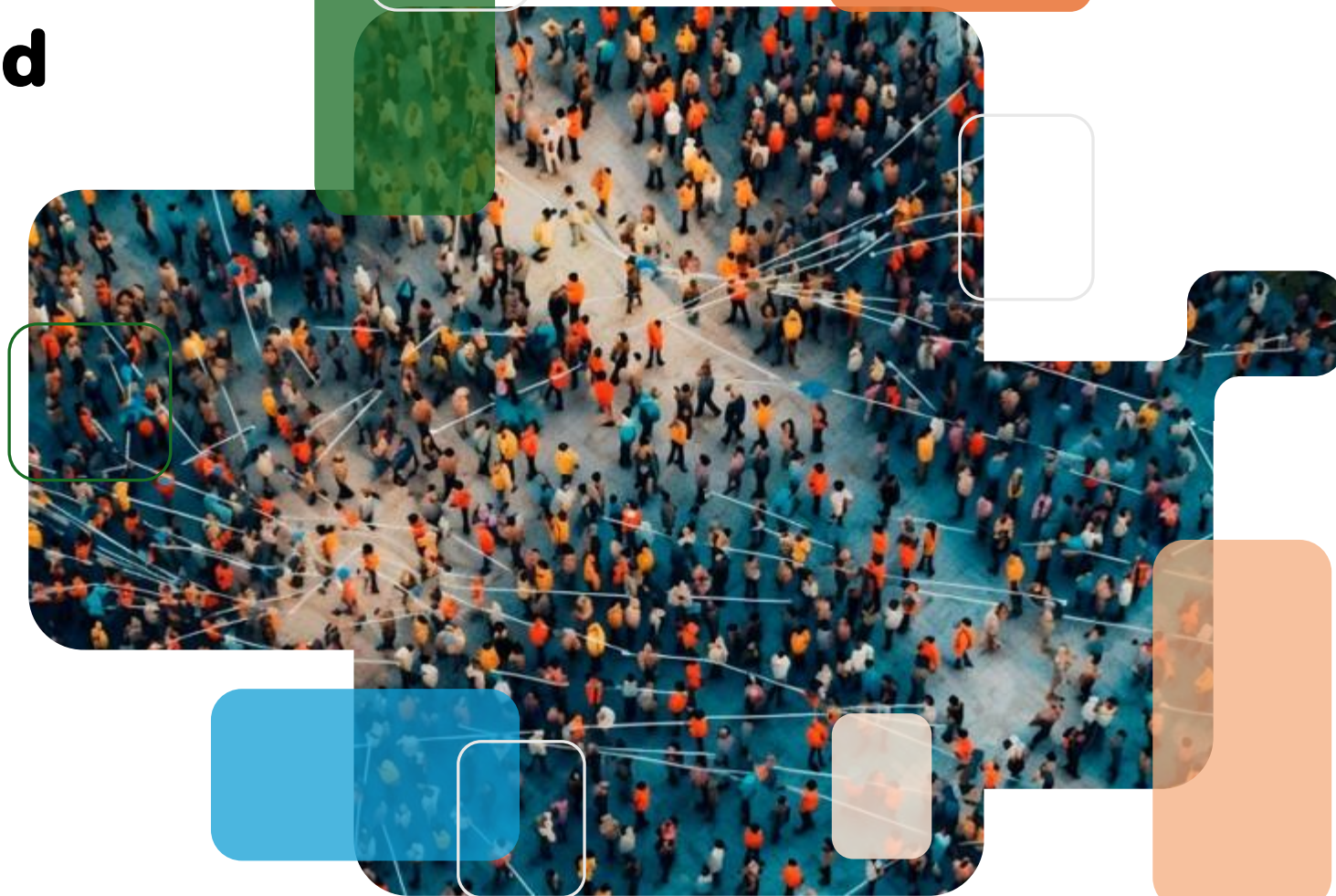


Standaardisatieraad



Npuls

| Soevereine AI?



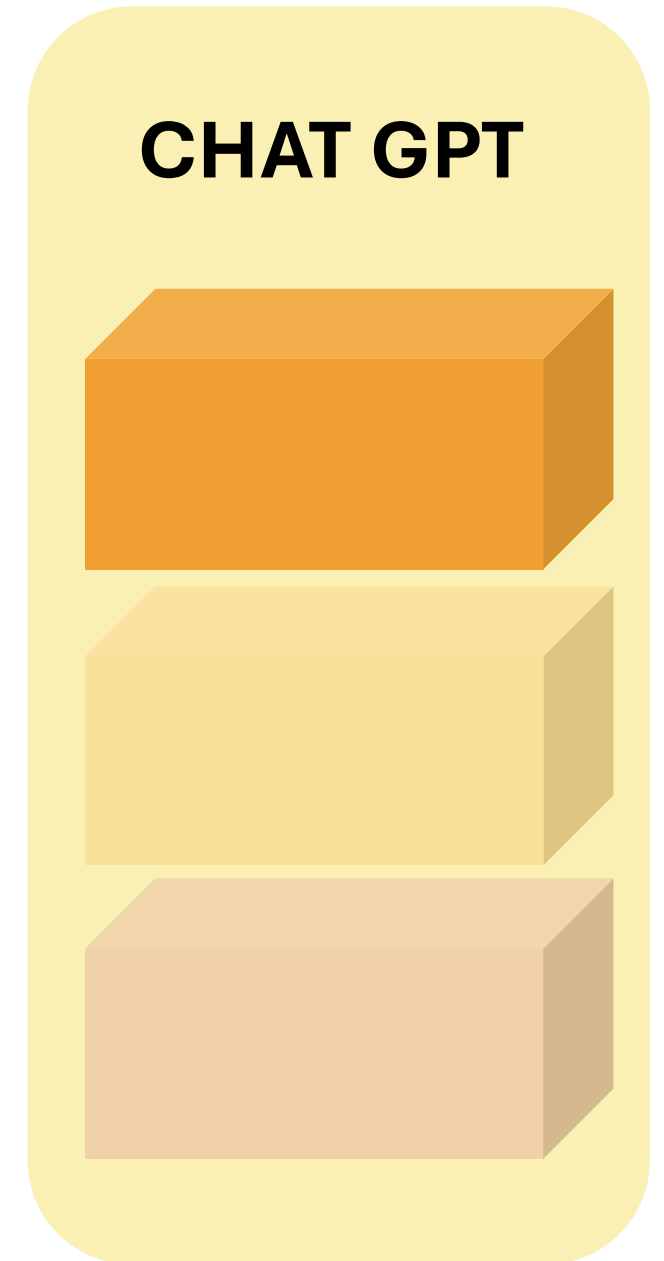
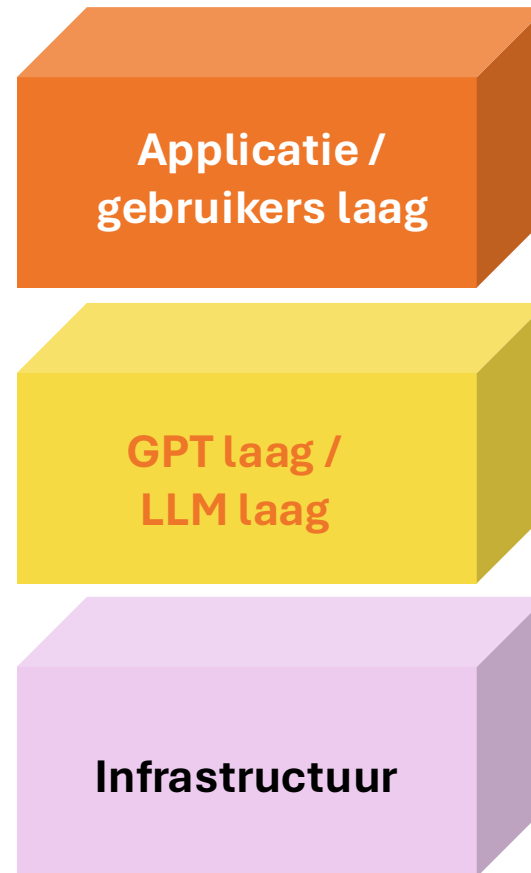
Npuls

| Visie AI

We want to get the most out of AI in a trustworthy way

Trustworthy staat voor voor wettige, ethische en robuuste oplossingen. We respecteren alle toepasselijke wet- en regelgeving, gebruiken en ontwerpen AI-systemen die gericht zijn op eerlijkheid, transparantie en veiligheid, met respect voor mensenrechten en ethische overwegingen.

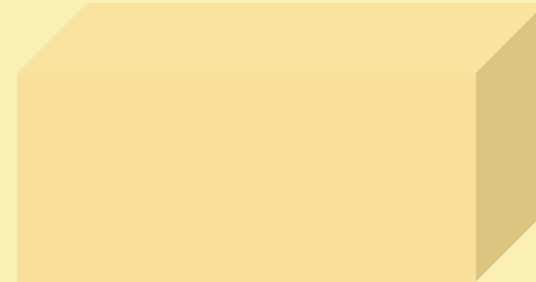
Commerciële systemen bestaan uit drie lagen die als één product worden aangeboden



**Daardoor
is er geen
invloed op de
verschillende
lagen**



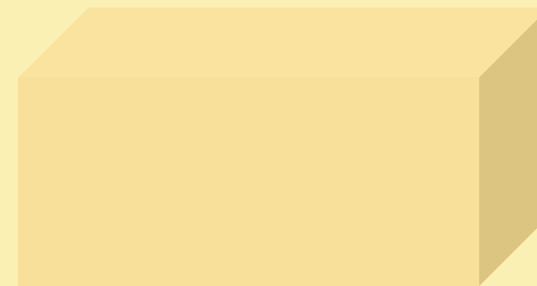
CHAT GPT



**Dat breken we
open met een
tussenlaag:
de AI hub**



AI hub



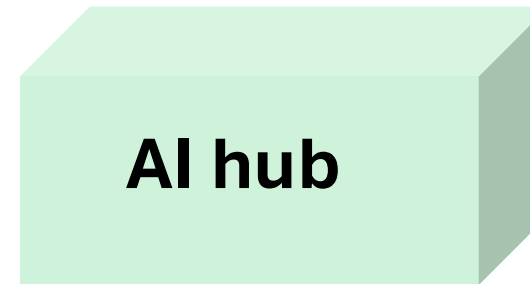
Hoe ziet dat eruit in de praktijk?

De AI hub is een vertrouwd koppelpunt op basis van gedeelde waarden, waar applicaties/gebruikers kunnen kiezen en koppelen met GPT en LLM

Dit betekent maximale flexibiliteit bij gebruikers en minimale macht bij aanbieders



Bijvoorbeeld
eduGenAI



Ontwikkeld door
SURF/Npuls. Wordt
open source



GPT-NL

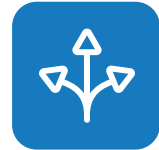
Bijvoorbeeld
GPT-NL



| Wat zijn de uitdagingen zonder AI-Hub?



Afhankelijkheid van
bigtech



Data Privacy



Onzekerheid over
reproduceerbaarheid



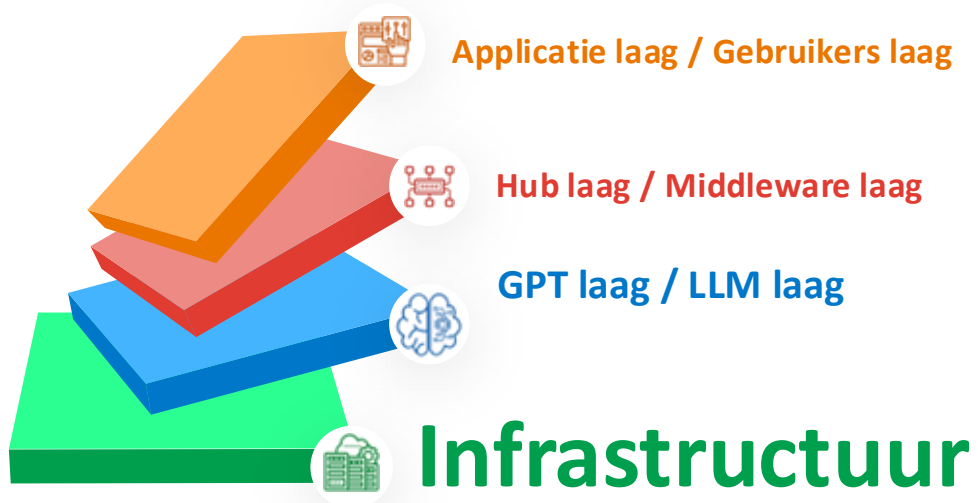
Hoge investeringen in
hardware



Tekort aan geschoold
personeel



Efficiënt gebruik van
resources



Latency

Latency mode

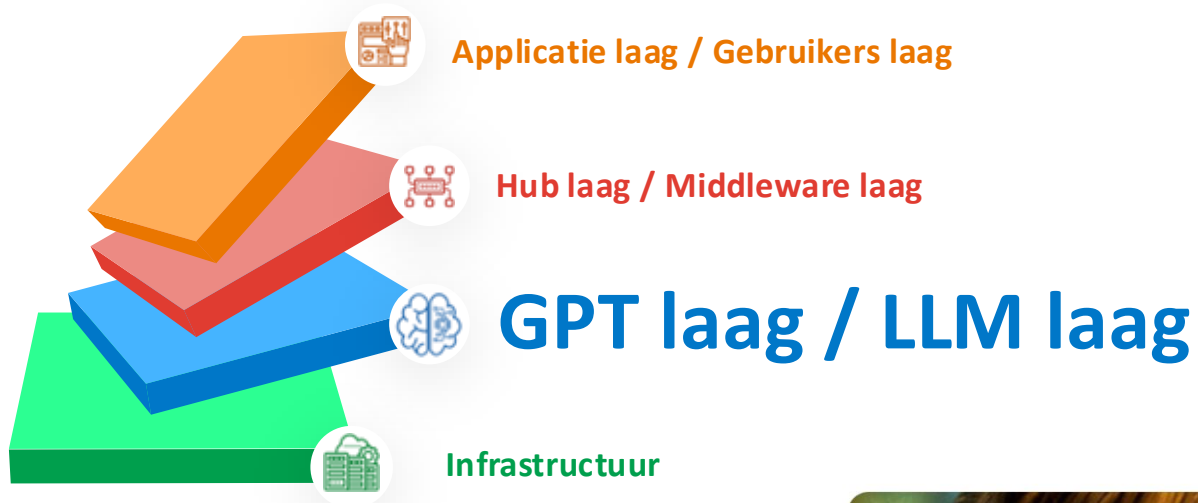
The API supports different latency modes.

Mode	latency	Models	Example use-cases
Always-on	< 10 sec	Only models marked with as always-on	Chat interfaces
On-demand	< 30 min	All models available to you	Playground
Batch	< 24 hrs	All models available to you	Research, automatic workflows



Npuls

We geven inzicht in het energieverbruik door de leden



Normeren en controleren

Vraag: hoe doen we DPIA's op taalmodellen?



SURF adviseert onderwijs- en onderzoekinstellingen terughoudend om te gaan met de inzet van Microsoft 365 Copilot. Dat is de conclusie van een nieuwe DPIA (Data Protection Impact Assessment) op Microsoft 365 Copilot. Ondanks dat er verbeteringen zijn doorgevoerd in de AI-toepassing, zijn niet alle risico's overtuigend geadresseerd.



GPT-NL

Towards a fair data value chain

- We believe data artists, journalists and other creators should be paid for their work
- We pay 50% of the revenue from the commercial license to the owners of the data.
- The other 50% will solely be used for continuation of GPT-NL

	Content
	Umbrella dutch media
	Dutch news medium
	Dutch news medium
	Student thesis'
	Financial documents
	Societal research
	ICT law documents
se Taal	Dutch linguistic research/data
	Social reports
or de	Medical articles
	Travel blogs
	Frysian linguistics

GPT-NL Corpus	Content / Source Description
Collected Data (GPT-NL Curated)	
Openraadsinformatie	Municipal council documentation
Officiële bekendmakingen	Government announcements
Woogole	Open Dutch government documents
Koninklijke Bibliotheek	Public domain Dutch texts
De Rechtspraak	Judicial cases
Tweede Kamer	Dutch parliamentary documents
Nationaal Archief	Dutch archive
Belgian Journal	Belgian company bylaws (Flemish foc
Utrechts Archief	Dutch archive
Noord-Hollands Archief	Dutch archive
Zeeuws Archief	Dutch archive
Dienst Publiek en Communicatie	Dutch public communication docs
Wikiwijs	Dutch school content
PBL	Planbureau Leefomgeving docs
Naturalis	Biological publications
European Parliament	Multilingual EU documents
DANS-KNAW	Dutch archaeology descriptions
Auditedienst Rijk	Dutch audit publications

Europees AI-bedrijf Mistral komt met nieuwe AI-modellen en maakt ze opensource



Door **Daan van Monsjou**

Nieuwsredacteur

[Feedback](#) • 02-12-2025 17:29 85 • submitter: [Westpjojr](#)

Het Franse AI-bedrijf Mistral komt met verschillende nieuwe Mistral 3-llm's. Daaronder valt een groot 'frontiermodel' dat multimodaal werkt en meerdere talen ondersteunt, naast een stel kleinere modellen die lokaal gedraaid kunnen worden. Ze komen allemaal opensource beschikbaar.

Mistral Large 3 wordt [volgens de ontwikkelaars](#) 'een van de beste openweightmodellen ter wereld'. De Franse start-up trainde het model zelf met 3000 Nvidia H200-gpu's. Het gaat om een [mixture-of-experts](#)-model dat bestaat uit 47 miljard actieve en 675 totale parameters.

Het nieuwe flagshipmodel kan volgens Mistral meekomen met andere populaire openweightmodellen zoals Kimi K2 en Deepseek V3.1. Daarnaast voegt het ook verschillende functies toe die al langer in gesloten AI-modellen zitten, zoals multimodale functies en ondersteuning voor meerdere talen. Het model komt beschikbaar onder een Apache 2.0-licentie.

Naast het Mistral Large 3-model komt Mistral ook met verschillende kleinere llm's. Deze krijgen de naam Ministral 3 en worden geleverd in modellen met 3, 8 en 14 miljard parameters. Van ieder model komen bovendien base-, instruct- en reasoningvarianten, ieder met de mogelijkheid om afbeeldingen te begrijpen. Volgens de makers zijn deze modellen bedoeld voor lokaal gebruik. Ook al deze varianten komen beschikbaar onder een Apache 2.0-licentie.

Bron: <https://tweakers.net/nieuws/242176/europees-ai-bedrijf-mistral-komt-met-nieuwe-ai-modellen-en-maakt-ze-opensource.html>

Behoeftte aan framework om te bepalen wanneer modellen compliance zijn

Initiatief gemeente Amsterdam

× Gemeente
× Amsterdam

EN | NL

LLM Overzicht in de Nederlandse Taal

Om het verantwoord gebruik van LLM's te bevorderen, bieden wij een overzicht van modellen die ingezet kunnen worden binnen de Nederlandse overheid. Omdat modellen zich anders gedragen per taal, zijn alle experimenten direct in het Nederlands uitgevoerd.

Model	Aanbieder	Licentie	Training Data	Energieverbruik	Kosten	Bias	Feitelijkheid	Eerlijkheid
TinyLlama-1.1B-Chat-v1.0	StatNLP	Open	Open	0.63Wh	€0.01	? ? ? ?	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
Phi-4-mini-instruct	Microsoft	Open	Beschreven	2.79Wh	€0.03	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
SmolLM3-3B	HuggingFace	Open	Open	3.14Wh	€0.04	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
c4ai-command-r7b-12-2024	Cohere	Beperkt	Beschreven	5.09Wh	€0.06	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
Qwen3-8B	Alibaba Cloud	Open	Gesloten	5.71Wh	€0.07	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
Llama-3.1-8B-Instruct	Meta	Beperkt	Gesloten	5.8Wh	€0.07	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
Falcon3-7B-Instruct	TII	Beperkt	Gesloten	5.92Wh	€0.07	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
Apertus-8B-Instruct-2509	Swiss National AI Institute	Open	Open	6.19Wh	€0.07	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
EuroLLM-9B-Instruct	UTTER Project	Open	Open	6.33Wh	€0.08	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
Mistral-7B-Instruct-v0.3	Mistral AI	Open	Gesloten	6.39Wh	€0.08	⚠️ ⚠️ ⚠️ ⚠️	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●
gpt-oss-20b	OpenAI	Open	Gesloten	9.33Wh	€0.11	⚠️ ⚠️ ⚠️ ?	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ●



Bron: <https://amsterdam.github.io/grip-on-llms/nl/>



Applicatie laag / Gebruikers laag

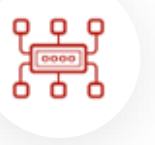
Hub laag / Middleware laag

GPT laag / LLM laag

Infrastructuur

API-keys conform OpenAI standaard

 1. Applicatie laag / Gebruikers laag
"de interface"

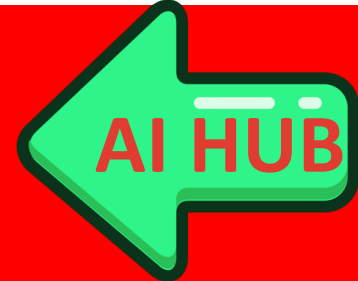
 2. Hub laag / Middleware laag
SURF "Trusted AI HUB"

 3. GPT laag / LLM laag

Sandboxed

 4. Infrastructuur


SURF Infrastructuur



In datacenter SURF onder beheer van SURF

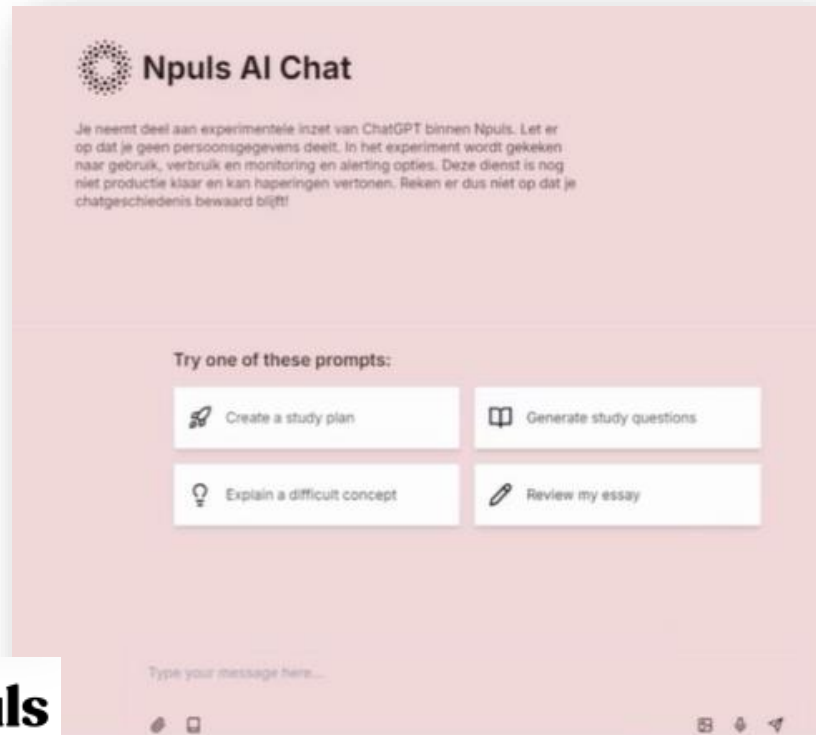


Applicatie laag / Gebruikers laag

Hub laag / Middleware laag

GPT laag / LLM laag

Infrastructuur



- ✓ Veilig gebruik van foundation modellen
- ✓ Userinterface voor het onderwijs (door Uva/HvA)
- ✓ Keuze uit verschillende LLM'S
- ✓ Chatten met eigen data/documenten
- ✓ Promptbibliotheek
- ✓ Persona's gebruik in het onderwijs
- ✓ Leren over en met AI – verantwoorde inzet
- ✓ Ecosysteem van instellingen

Applicatie laag / Gebruikers laag

Hub laag / Middleware laag

GPT laag / LLM laag

Infrastructuur




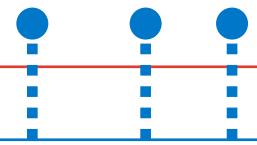
The screenshot displays the LibreChat web interface. On the left, there is a sidebar with a search bar and a list of chat sessions under the heading "Previous 30 days". The main area shows a chat window with a message input field and a dropdown menu for selecting AI models. The dropdown menu is open, showing a search bar "Search Willma models..." and a list of models including LLaMa-2 13B Chat, LLaMa-2 7B Chat, ChatGPT-3.5, Zephyr 7B, Mixtral Instruct AWQ, Llama 3.1 8B Instruct, Phi-3.5 mini, Qwen2.5 7B, Qwen2.5-Coder-1.5B-Instruct, Qwen2.5-Coder-7B-Instruct, stabilityai/stable-code-instruct-3b, DeepSeek Distilled Llama 70B, Llama 3.3 70b Instruct AWQ (selected), Qwen 2.5 VL 32B Instruct AWQ, and Qwen 2.5 Coder 32B Instruct AWQ. The chat window also shows a "Message Willma" button and a "Code Interpreter" button. The footer of the interface includes the user name "Corno Vromans" and the version "LibreChat v0.7.8 - Every AI for Everyone." along with links to "Privacy policy" and "Terms of service".


SURFice


Nextcloud en AI

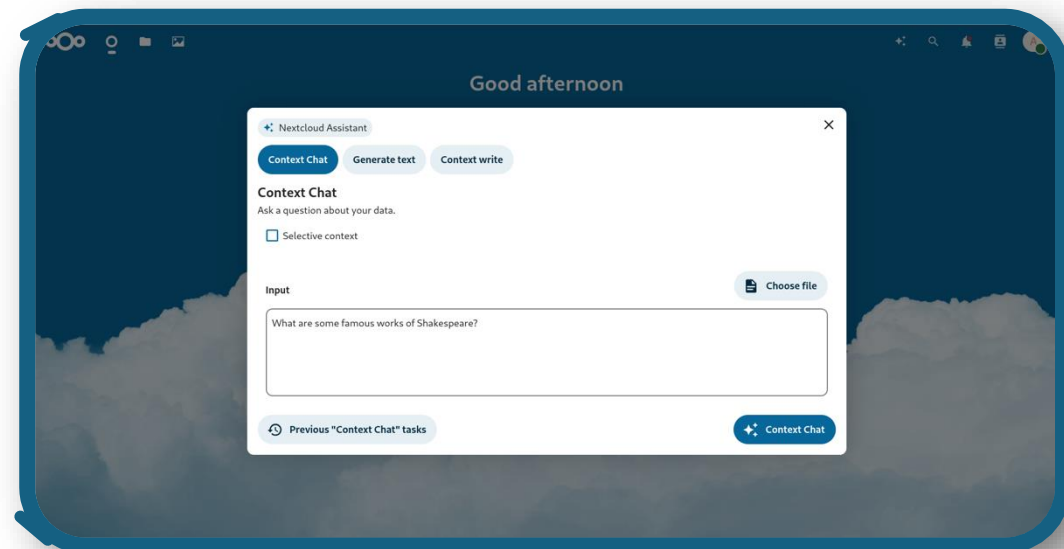
Koppeling van AI assistent aan AI-hub gelukt! (op test omgeving)

 **1. Applicatie laag / Gebruikers laag**
"de interface" 

 **2. Hub laag / Middleware laag**
SURF "Trusted AI HUB" 

 **3. GPT laag / LLM laag**





Integratie met bestaande toepassingen

The AI-Hub can be used as a backend for various applications that support the OpenAI API standard. We have some documentation to get started, but it's up to you to enroll this responsibly within your institute:

<p>OpenWebUI: https://openwebui.com/</p> <p>browser-based interface for running and managing local or remote large language models.</p> <p>> Getting started (behind login)</p>	<p>Nvidia Nemo Guardrails: https://github.com/NVIDIA-NeMo/Guardrails</p> <p>NeMo Guardrails is an open-source toolkit for easily adding programmable guardrails to LLM-based conversational applications.</p> <p>> Getting started (behind login)</p>	<p>PyCharm: https://www.jetbrains.com/pycharm/</p> <p>A popular IDE from JetBrains designed specifically for Python development, with strong debugging and productivity tools.</p> <p>> Getting started (behind login)</p>
<p>Elastic: https://www.elastic.co/security</p> <p>Using the ai-hub as an inference provider, you can also use the Elastic Security AI Assistant.</p> <p>> Getting started</p>	<p>LibreChat: https://www.librechat.ai/</p> <p>An open-source, self-hostable alternative to ChatGPT that lets you run and customize conversational AI.</p> <p>> Getting started (behind login)</p>	

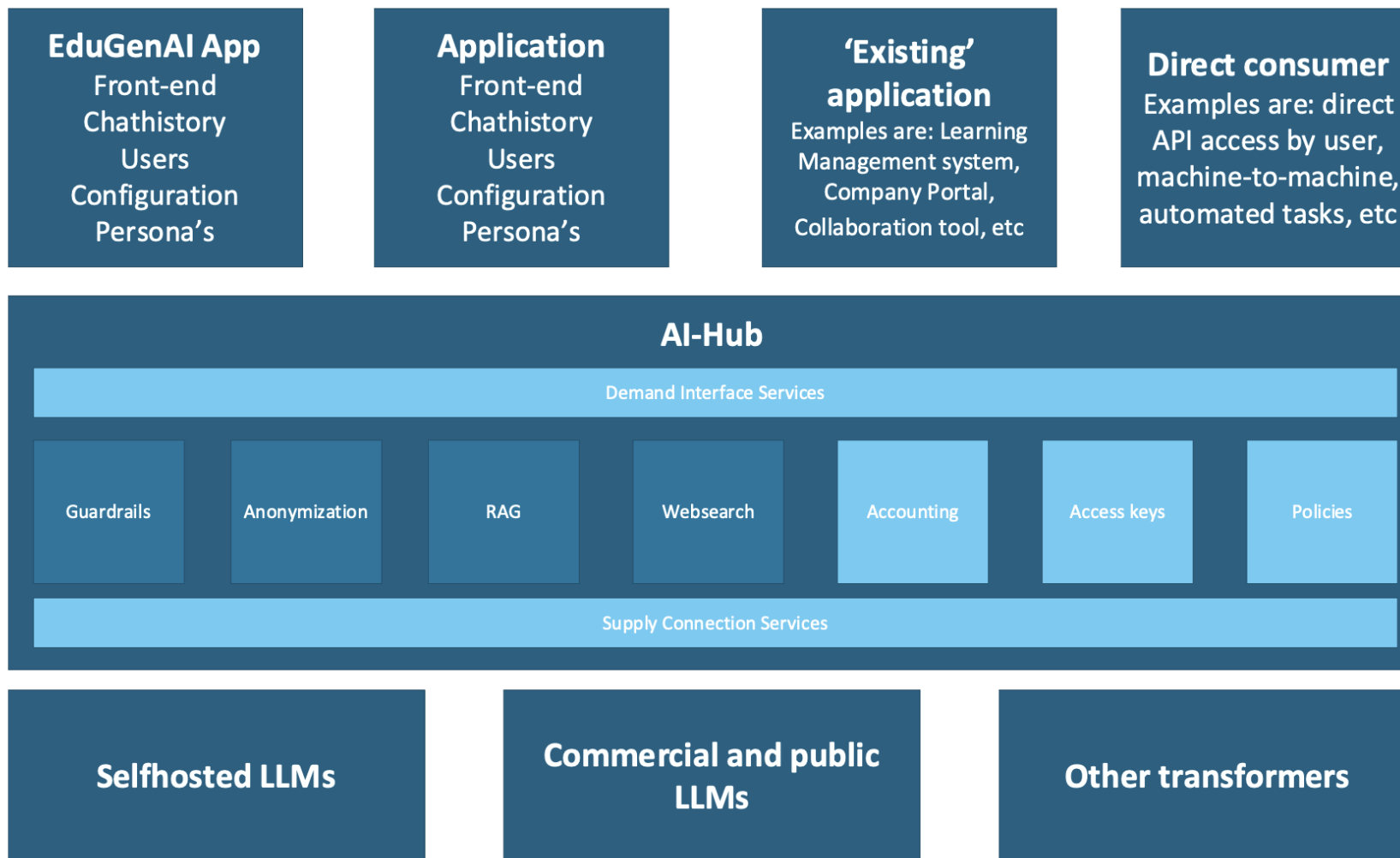
NextCloud has also successfully tested in an experimental setup, but you are welcome to try other integrations as well.

Development

The AI-Hub exposes an API to develop your applications or workflows against.

<p>Onboarding</p> <p>Check our onboarding documentation to get from zero to your first successful request.</p> <p>> Onboarding</p>	<p>Python code examples</p> <p>Shows examples when using different modalities and functions such as tool calling.</p> <p>> Python code examples (behind login)</p>	<p>API documentation</p> <p>The API-documentation can be used for detailed information on the AI-Hub endpoints</p> <p>> API Swagger Documentation (behind login)</p>
--	--	--

Ontwikkeling AI-Hub

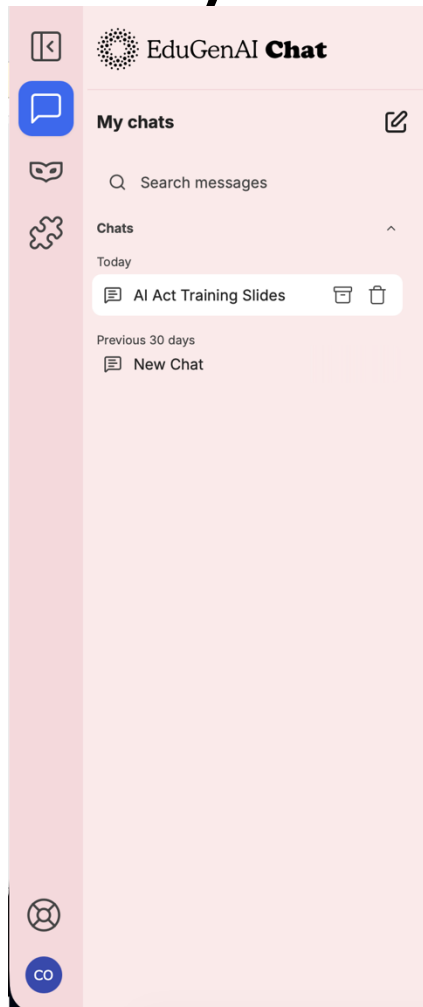


Integratie met andere systemen

MCP

Toetsen van guardrailsmodellen op Nederlands

Voorbeeld gebruik MCP (model context protocol)



Corno

Ik zoek onderwijs materiaal (in pptx-format) voor een training over de ai-act. Kun je me hiermee helpen?

EduGenAI Chat

- ✓ Ran get_terms_and_conditions
- ✓ Ran get_filter_categories
- ✓ Ran get_filter_values
- ✓ Ran get_filter_values
- ✓ Ran advanced_search_materials

Hieronder vind je een passend les-/presentatiemateriaal dat je direct kunt downloaden en gebruiken voor een training over de (EU) AI-Act.

Titel	Beschrijving	Bestand (pptx)	Licentie	Publicatiedatum
GenAI Geletterdheid – “Verantwoord en effectief gebruik van GenAI” (werk-/ hoorcollege)	Een PowerPoint-presentatie (plus bijbehorend lesplan) die de basis van AI-literacy behandelt. Het materiaal refereert expliciet aan de AI-Act (art. 4) en is gericht op universitaire studenten, maar kan gemakkelijk aangepast worden voor een bredere training.	PPT-college-GenAI literacy-OER-feb 2025	CC-BY-NC-SA 4.0 (Open Access)	27-mei-2025

Waarom dit materiaal geschikt is

Formaat – Het is een PowerPoint-presentatie (pptx), zodat je de slides direct kunt bewerken of

Ask your question

+ : edusources x

gpt-oss-120b v



Regie bij de leden!

Enable/disable models

By CO / Group

To enable models, ensure you are in the 'Willma Model Administrator' in your CO.

Select CO:

surf_rsc:advanceddutchllm

View by CO View by Group

All highlighted models on the right are **enabled** for the selected scope. Use the arrows or double-click to move.

Available models

- Zephyr 7B (text)
- openai/gpt-oss-120b (text)

Enabled models

- Stable Diffusion XL 1.0 (image)
- Massively Multilingual Speech Dutch (tts)
- Massively Multilingual Speech English (tts)
- RAG document indexer (index)
- Stable Diffusion 3 Medium (image)
- Mixtral Instruct AWQ (text)
- Llama 3.1 8B Instruct (text)
- Phi-3.5 mini (text)
- gemma-2 9b (text)
- Qwen2.5 7B (text)
- Qwen2.5-Coder-1.5B-Instruct (text)
- Qwen2.5-Coder-7B-Instruct (text)

»

>

<

«

Save Selection