

The background of the image is a digital landscape. On the left and right sides, there are rows of server racks or data center units, receding into the distance. The floor is a smooth, reflective surface that mirrors the racks and the sky above. In the center of the image, a large, dark, and billowing cloud formation rises. A bright, jagged lightning bolt strikes down from the base of this cloud towards the center of the floor. The overall color palette is dominated by deep blues and greys, with the white of the lightning providing a sharp contrast.

Lokale KI



# Regularien & Gesetzgebungen

AI-Act  
DSGVO  
Compliance  
Urheberrecht

# AI-Act



[Link zur RTR AI-Act-Seite](#)



## Unzulässige KI-Anwendungen

Anwendungen, die aufgrund von Sicherheits- und ethischen Bedenken verboten sind.

zB.: Social-Scoring, biometrische Überwachung...



## Hochrisiko-KI-Anwendungen

Anwendungen, die strengen Anforderungen unterliegen, um Fairness und Transparenz zu gewährleisten.

zB.: Gesundheitswesen, HR/Recruiting, Strafverfolgung...



## Geringes Risiko-KI-Anwendungen

Anwendungen, die grundlegende Sicherheitsstandards erfüllen, aber weniger reguliert sind.

zB.: Chatbots, Deepfakes...

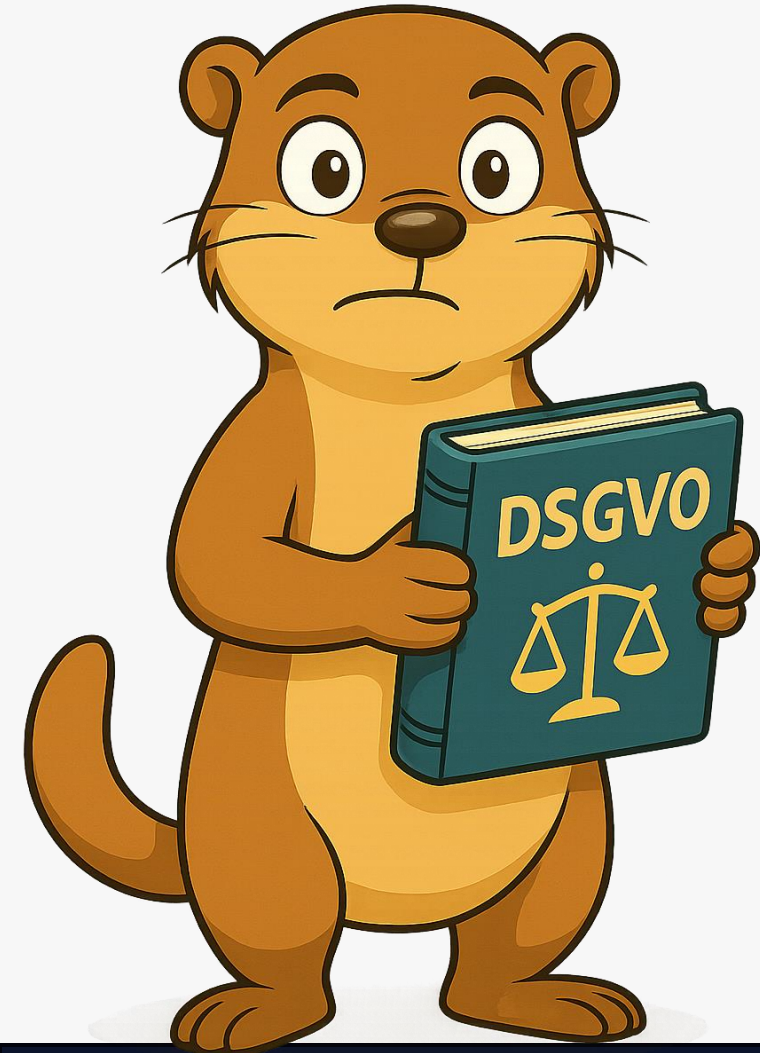


## Minimalrisiko-KI-Anwendungen

Anwendungen, die kaum reguliert sind, da sie keine signifikanten Risiken darstellen.

zB.: Entertainment, Datenanalyse, Spam-Filter...

# DSGVO & URHEBERRECHT



- DSGVO bleibt bei KI vollständig anwendbar. AI Act ergänzt, ersetzt nicht.
- KI-Bilder haben kein Urheberrecht
- Kein exklusives Recht an KI-Bilder (daher kann es jeder nutzen)
- Vorsicht beim Verwenden urheberrechtlich geschützten Bildern/Dokumenten (klären ob erlaubt)



# Local vs Cloud

Die “eigene” KI

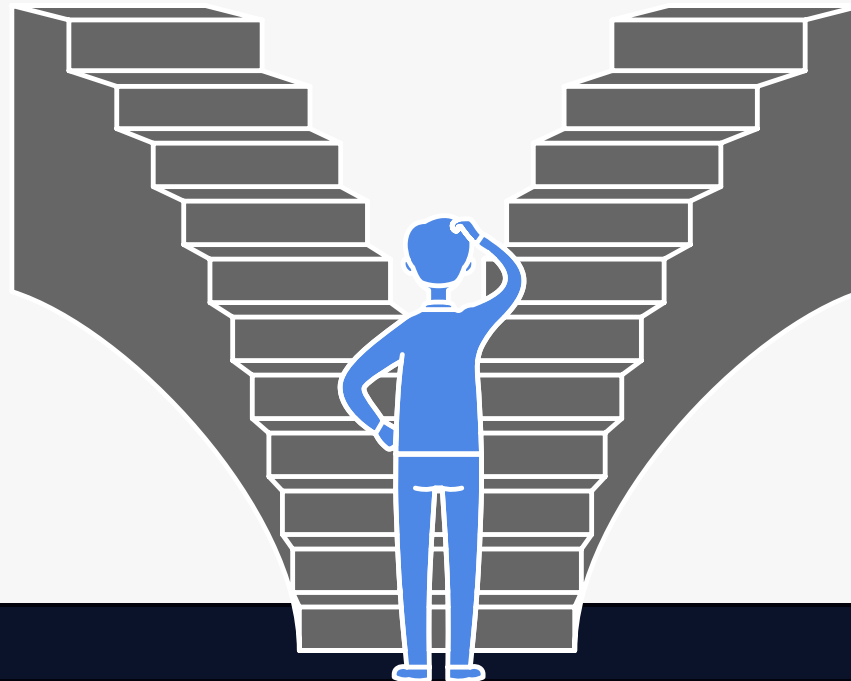
# Cloud vs Lokal

## Cloud- Implementierung

Einfache Einrichtung.  
Bedenken hinsichtlich  
Datenverarbeitung. API  
kostet Extra.

## Lokale Implementierung

Bietet Datenkontrolle und  
große Modellauswahl.  
System-ressourcen.



# Cloud vs Lokal

## Cloud

### ■ Vorteile:

- Schnell und „einfach“ umgesetzt
- Hardware ausgelagert

### ■ Nachteile:

- Es ist nicht klar, was mit den Daten passiert. Wie werden die Daten weiterverarbeitet?
- keine Daten-Souveränität

## Lokal

zB: Ollama mit Mistral oder Llama3, Deepseek usw.

### ■ Vorteile:

- Alle Daten liegen im Unternehmen
- Volle Kontrolle über Daten und KI-Modelle

### ■ Nachteile:

- Systemanforderungen
- ggf. trainieren der KI

# Anforderungsmanagement

## ■ Randbedingungen erheben:

- Welche Technologien werden bereits im Unternehmen verwendet? Gibt es KI-Tools, die mit diesen Technologien gut zusammenarbeiten?
- Wie sieht die Infrastruktur des Unternehmens aus? Eigene Server und Infrastruktur oder ausgelagert in die Cloud?
- Wie steht es um Security oder Datenschutzrichtlinien?

## ■ Ziele definieren:

- Was für ein Ziel soll mit der KI-Automatisierung erreicht werden?
- Kosten- und/oder Zeitersparnis?



# MCP

## Model-Context- Protokoll

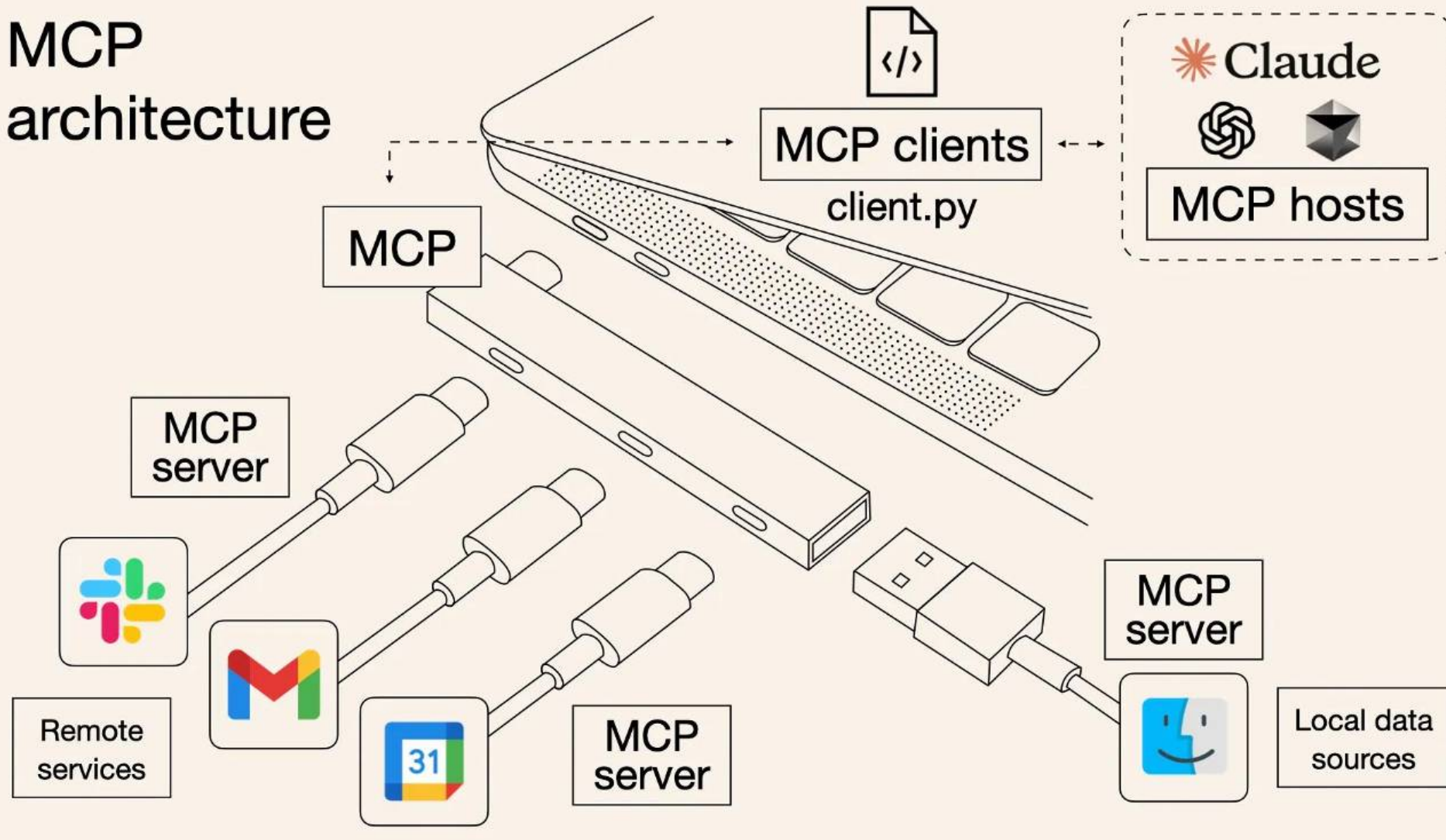
Entwickelt von  
Anthropic.

„Konnektor“ für  
diverse Anbindungen.  
zB.: Claude kann  
Blender steuern und  
Renderings erstellen,  
mittels natürlicher  
Sprache.

MCP-Markt:

<https://mcpmarket.com>

## MCP architecture



The background of the slide features a photograph of a modern train station. The image is split diagonally from the top-left to the bottom-right. The upper-left portion shows the station's ceiling with curved, metallic structural elements and recessed lighting. The lower-left portion shows the platform with rows of blue and grey seats. The right half of the slide is a solid white area.

# Modelle

# Aktuell 3 Möglichkeiten

- Proprietäre Modelle (Anthropic, OpenAI, Google uä. bekannt Anbieter von großen Multimodalen Modellen)
- Open-Source Modelle (Llama, Deepseek, Mistral)
- Vortrainierte/spezifische Open-Source Modelle (1,832,356 bei Huggingface)

# Huggingface & Ollama

 **Hugging Face**

Models

Datasets

Spaces

Tasks

Libraries

Datasets

Languages

Licenses

Other

Multimodal

Audio-Text-to-Text

Image-Text-to-Text

Visual Question Answering

Document Question Answering

Video-Text-to-Text

Visual Document Retrieval

Any-to-Any

Computer Vision

Depth Estimation

Image Classification

Object Detection

Image Segmentation

Text-to-Image

Image-to-Text

Image-to-Image

Image-to-Video

Unconditional Image Generation

Video Classification

Text-to-Video

Zero-Shot Image Classification

Mask Generation

Zero-Shot Object Detection

Text-to-3D

Image-to-3D

Image Feature Extraction

Models **1,776,273**

Filter by name

deepseek-ai/DeepSeek-R1-0528

Text Generation • Updated 11 days ago • 97.1k • 1.88k

google/gemma-3n-E4B-it-litert-preview

Image-Text-to-Text • Updated 14 days ago • 1.05k

Qwen/Qwen3-Embedding-0.6B-GGUF

Updated about 14 hours ago • 8.45k • 253

ResembleAI/chatterbox

Text-to-Speech • Updated 10 days ago • 728

fishaudio/openaudio-s1-mini

Text-to-Speech • Updated 7 days ago • 1.25k • 181

Qwen/Qwen3-Embedding-0.6B

Feature Extraction • Updated 3 days ago • 27.1k • 160

deepseek-ai/DeepSeek-R1-0528-Qwen3-8B

Text Generation • Updated 11 days ago • 180k • 714

nvidia/Nemotron-Research-Reasoning-Qwen-1.

Text Generation • Updated 5 days ago • 3.45k • 140

 [Discord](#) [GitHub](#) [Models](#)

Embedding

Vision

Tools

Thinking

Popular

**deepseek-r1**

DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.

thinking

1.5b

7b

8b

14b

32b

70b

671b

47.5M Pulls

35 Tags

Updated 1 week ago

**gemma3**

The current, most capable model that runs on a single GPU.

vision

1b

4b

12b

27b

5.6M Pulls

21 Tags

Updated 1 month ago

**qwen3**

Qwen3 is the latest generation of large language models in Qwen series, offering a comprehensive suite of dense and mixture-of-experts (MoE) models.

tools

thinking

0.6b

1.7b

4b

8b

14b

30b

32b

235b

2.1M Pulls

35 Tags

Updated 1 week ago

**devstral**

# Links

◆ Ollama → <https://ollama.com>

◆ Huggingface → <https://huggingface.co>

# Model Benchmarks

- ◆ Kreatives Schreiben: [https://eqbench.com/creative\\_writing.html](https://eqbench.com/creative_writing.html)
- ◆ AI Analysen diverser Modelle → <https://artificialanalysis.ai>

# Welches Model?

- Wie finde ich ein passendes?

- Testen, Testen, Testen

- Analysieren & Anpassen

- Wenn vorhanden Model-Cards lesen  
(bei Huggingface oft vorhanden)  
Geben Details zum Model & Training

- Reddit Einträge suchen – gibt dort sehr viel Austausch

# Model Tipps

Ein kleiner Auszug aus  
Modellen.

Reddit usw. gibt es unzählige  
Diskussionen und Reviews

■ Llama3.1

■ qwen

■ deepseek

■ gemma3(n)

■ mistral



The background features a dark blue gradient on the left side, transitioning into a lighter blue diagonal band that runs from the top left towards the bottom right. The remaining area on the right is white.

Welche Tools?

# Ollama

- Direkte Nutzung von Modellen in der Konsole/Terminal
- Kann mit OpenWebUI & anderen Tools verknüpft werden

Ollama Website: <https://ollama.com>

# LM Studio

- ◆ Gute Übersicht
- ◆ Einfach Installation
- ◆ RAG Möglichkeit direkt integriert
- ◆ Direkt Zugriff auf Huggingface
- ◆ Modelle können über Benutzeroberfläche direkt von Huggingface heruntergeladen werden

LM Studio Website: <https://lmstudio.ai>

# Anything LLM

- Einfach Installation
- kann etwas „träge“ sein
- RAG Möglichkeit nicht so stabil wie bei LM Studio

Anything LLM Website: <https://anythingllm.com>

# OpenWebUI

- Webserver mit Schnittstelle zu Ollama
- Ollama verwaltet Modelle OpenWebUI macht sie zugänglich
- CustomGPTs ähnlich ChatGPT & Co können angelegt werden
- Benutzerverwaltung und Modellsteuerung pro Benutzer
- Admin Verwaltung
- Installation über Docker möglich (Docker muss extra installiert werden)

OpenWebUI Website: <https://openwebui.com>

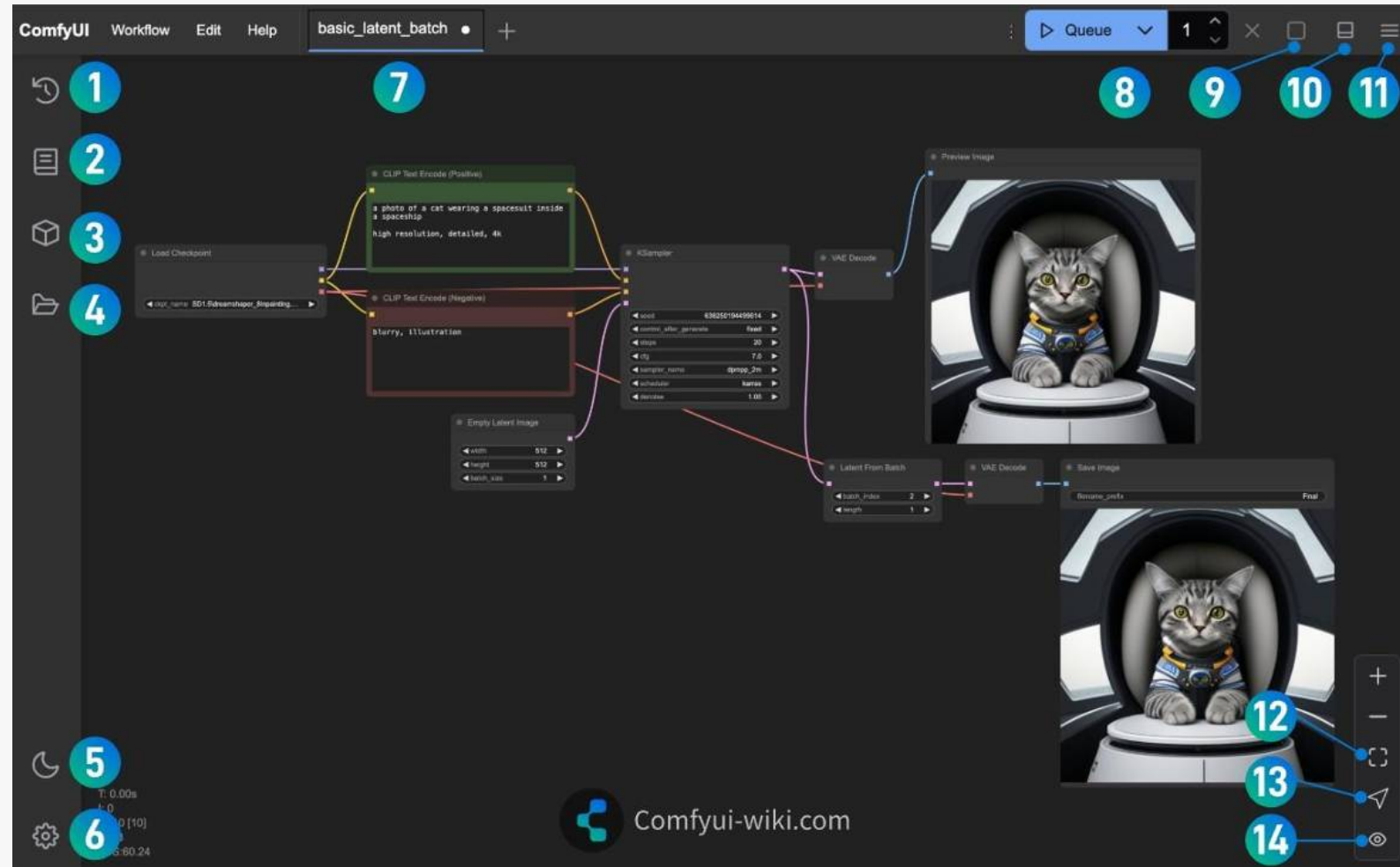
A person wearing a beanie and a scarf is holding a camera up to their eye, taking a photo. The background is blurred, suggesting an outdoor setting. The image has a blue tint.

# Lokale Video/Bild-Generierung

# ComfyUI

<https://www.comfy.org>

- No/Low-Code Builder
- OpenSource Modelle können integriert werden
- umfangreiches Tool mit enormen Möglichkeiten
- LoRa:  
<https://civitai.com/models>
- Gängige Modelle: Stable Diffusion oder Flux



Quelle: <https://comfyui-wiki.com>

# Was ist LoRA?

LoRA wird auf ein einziges neues Konzept, einen Charakter, einen Stil usw. trainiert und führt dann, wenn es auf ein bestehendes Modell angewendet wird, dazu, dass dieses neue Konzept zum Ausdruck kommt.

- LoRa = **Low-Rank Adaptation of Large Language Models**

<https://arxiv.org/abs/2106.09685>

Quelle Cornell University

- Erklärung LoRa & Co: <https://www.youtube.com/watch?v=dVjMiJsuR5o>





# Model Tricks

RAG

Quantisierung

# Wissen in das Model bringen

## ■ RAG (Retrieval-Augmented Generation)

- Daten in Vektordatenbanken (unterstützt Semantische Suche [Suche nach Bedeutung])
- Verbessert Genauigkeit und Relevanz der KI antworten

## ■ KI trainieren

- hohe Systemanforderungen
- Kleinere Open-Source Modelle können trainiert werden – jedoch weiterhin hoher technische Aufwand
- Zeitintensiv

Beides muss laufend optimiert & aktualisiert werden

# Große Modelle kleine Computer

- Modelle sind groß – sehr groß
- Quantisierung ermöglicht Platz zu sparen auf Kosten von Präzision
- Interessante Lösung für KI auf Geräten mit wenig Leistung (RAM/Grafikspeicher)

# Quantisierung-Benennung

- Q2

- Q4

- Q8

- je kleiner die Zahl, desto geringer die Präzision des Models – umso mehr Speicher wird gespart

- Beispiel:

- Es gibt eine bestimmte Anzahl an Postkästen. Jede Zahl muss sich einen Postkasten suchen, je kleiner die Q Zahl desto weniger präzise muss der Postkasten gewählt werden.

# Quantisierung - Erweiterung

- K-Quantisierung (Q4\_K, Q4\_KM, Q4\_KL)
- KS (klein), KM (mittel) und KL (groß) beziehen sich auf den Detaillierungsgrad der Aufzeichnungen darüber, wie die Zahlen in diesen spezialisierten Bereichen angeordnet sind. Q4\_K oft ein guter Ausgangspunkt
- Anstatt nur wenige, gleich große Postkästen zu haben, erstellen K-Quants **mehrere, spezialisierte "Posträume"**.
- Kleine Zahlen bekommen ihren eigenen, präzisen Bereich.
- Große Zahlen bekommen ihren entsprechend dimensionierten Bereich

# Pros & Cons Quantisierung

## Pros



- Reduzierte Größe
- Schnellere Inferenz
- Weniger Speicher
- Weniger Energie
- Zugänglichkeit

## Cons



- Genauigkeitsverlust
- Komplexität
- Hardwareunterstützung

Inferenz: Das Modell nutzt sein erlerntes Wissen, um neue, ungesehene Daten zu analysieren und Vorhersagen zu treffen oder Entscheidungen zu treffen.

# Los Geht's

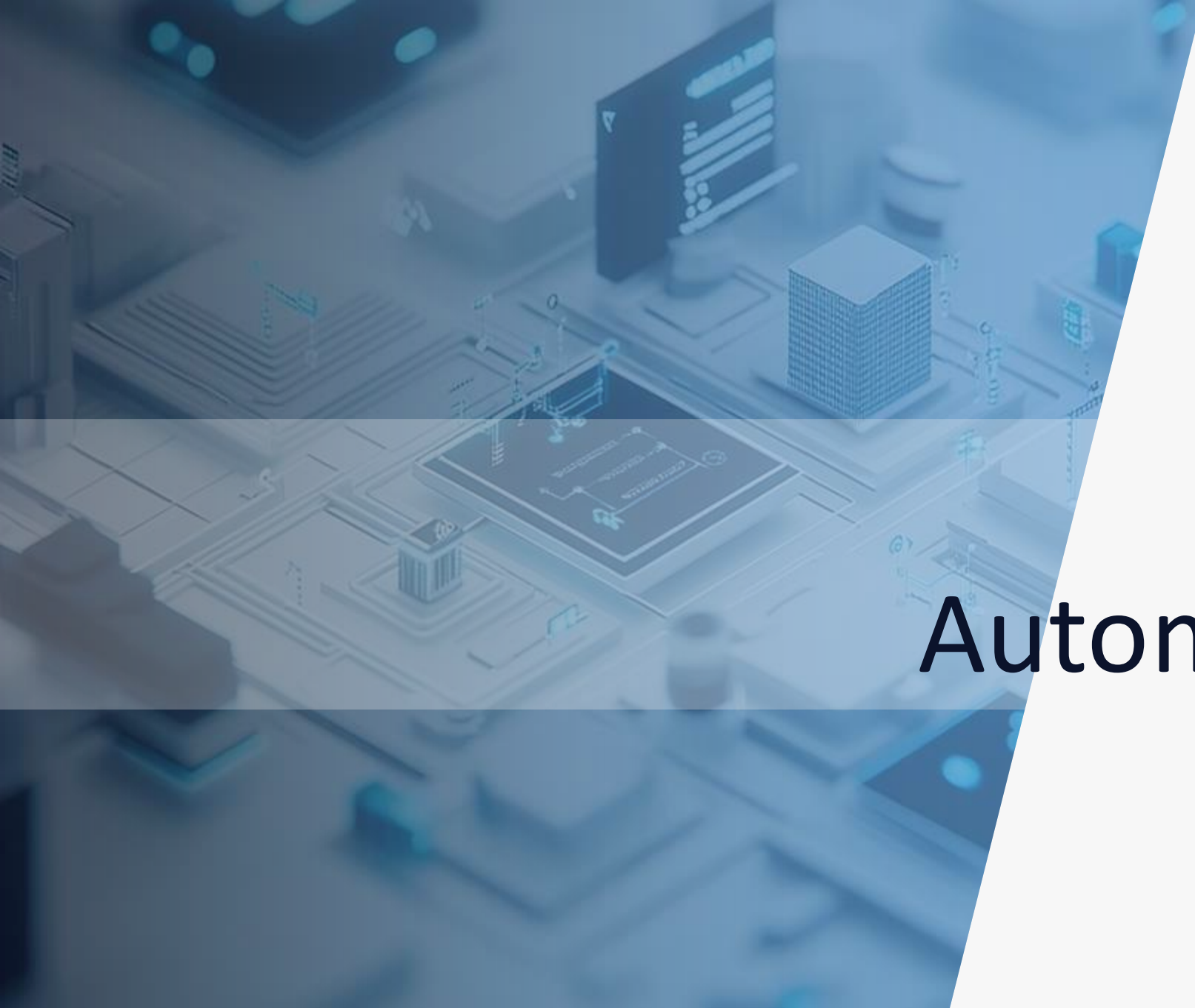
- LM Studio  
(<https://lmstudio.ai>)  
herunterladen & Installieren
- Model aussuchen
- Testen
- (Alternative Software: Jan:  
<https://jan.ai>)

- Modelgrößen empfehlungen:  
1B – 4B
- 7B-8B wenn Leistung vorhanden
- Welches Model?
- Deepseek
- qwen
- gemma3
- llama3.2

# Tablet / Smartphone Apps

- ◆ Private LLM (iOS): <https://privatellm.app/de>
- ◆ Private AI (Android):  
<https://play.google.com/store/apps/details?id=us.valkon.privateai>
- ◆ Pocket Pal (Android & iOS):  
<https://github.com/a-ghorbani/pocketpal-ai> (links zu Stores etwas weiter in der Mitte) – Empfehlung

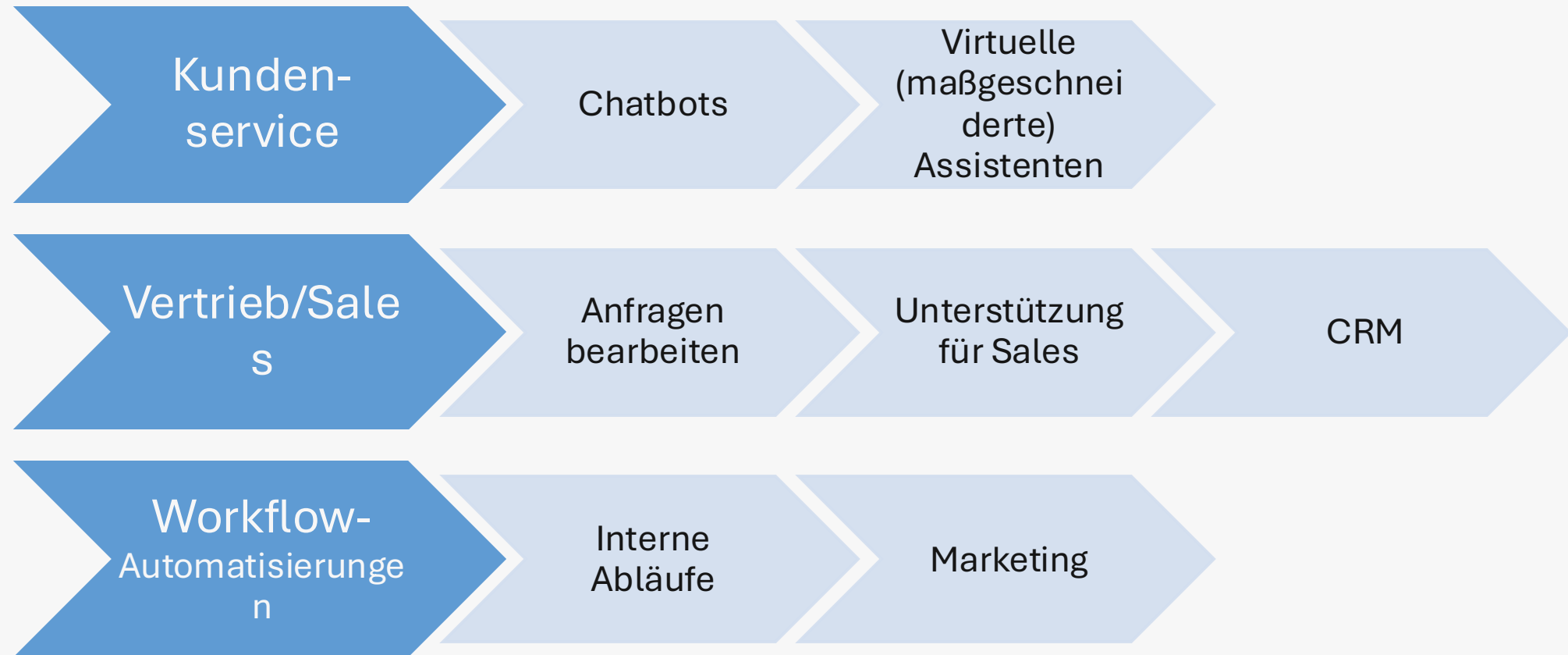


An isometric illustration of a city with various buildings, streets, and a central square. A semi-transparent blue overlay covers the top half of the image, creating a modern, technological feel.

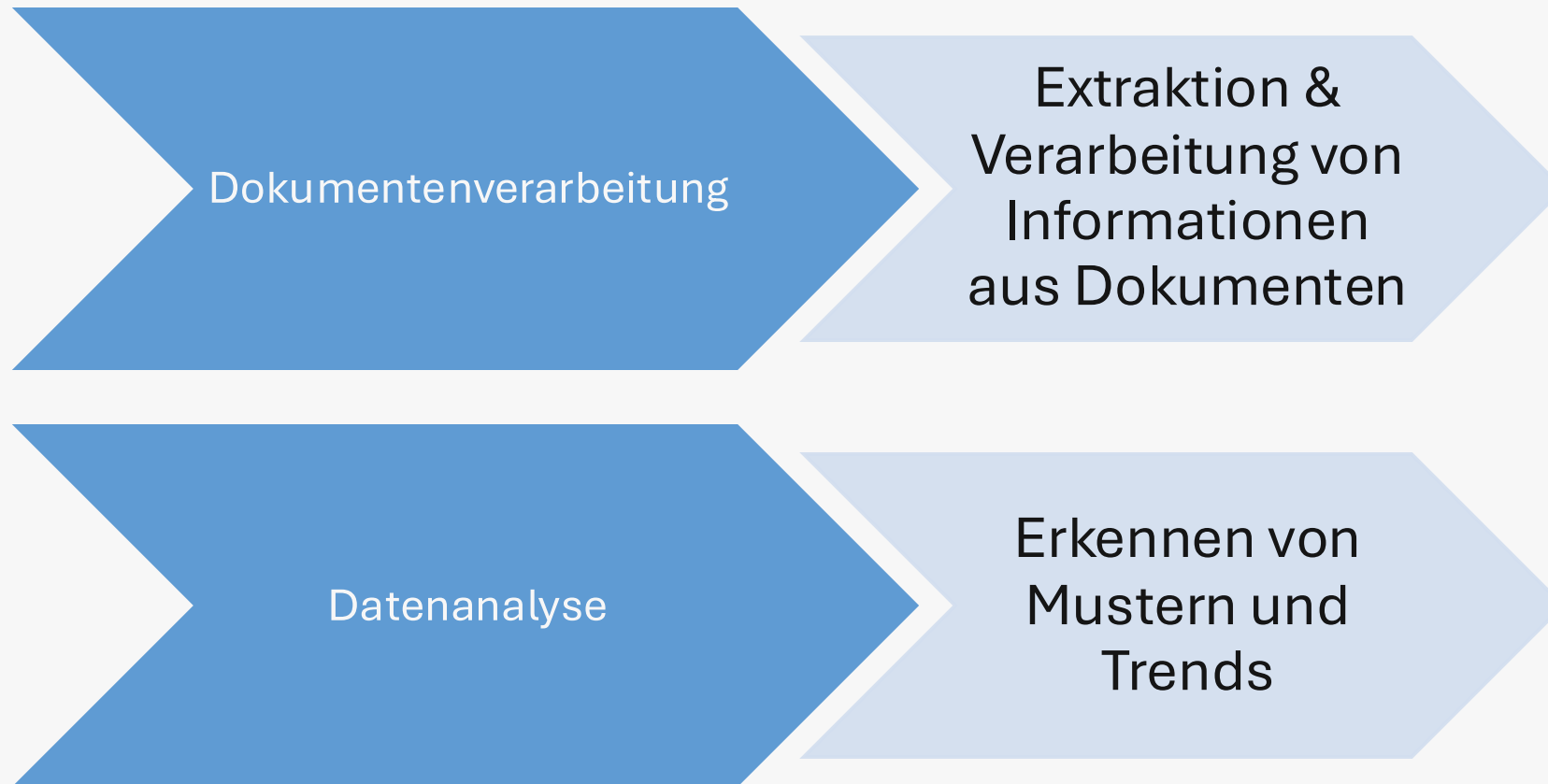
# Automatisierungen

Was ist nach der Installation  
der Lokalen KI noch möglich?

# Mögliche Anwendungsfälle I



# Mögliche Anwendungsfälle II



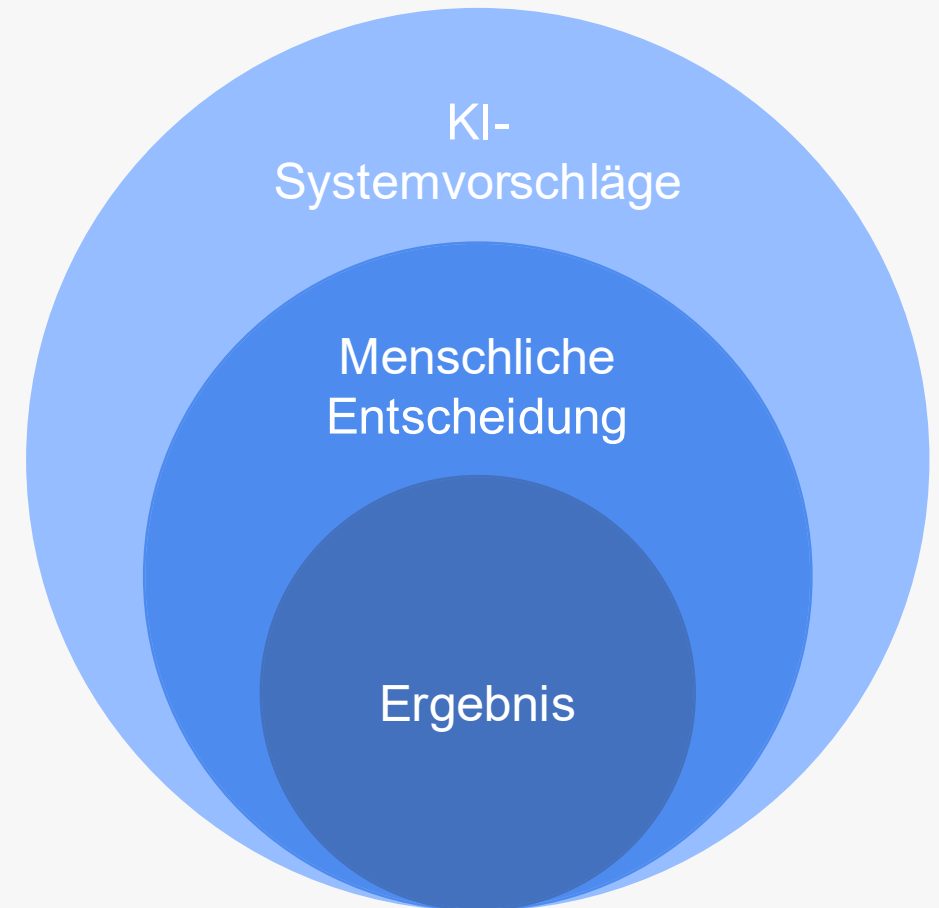
# Human in the Loop

Mensch und KI arbeiten als Team: Menschen behalten Kontrolle und Entscheidungshoheit in automatisierten Prozessen.

KI-generierte Optionen  
und Einblicke

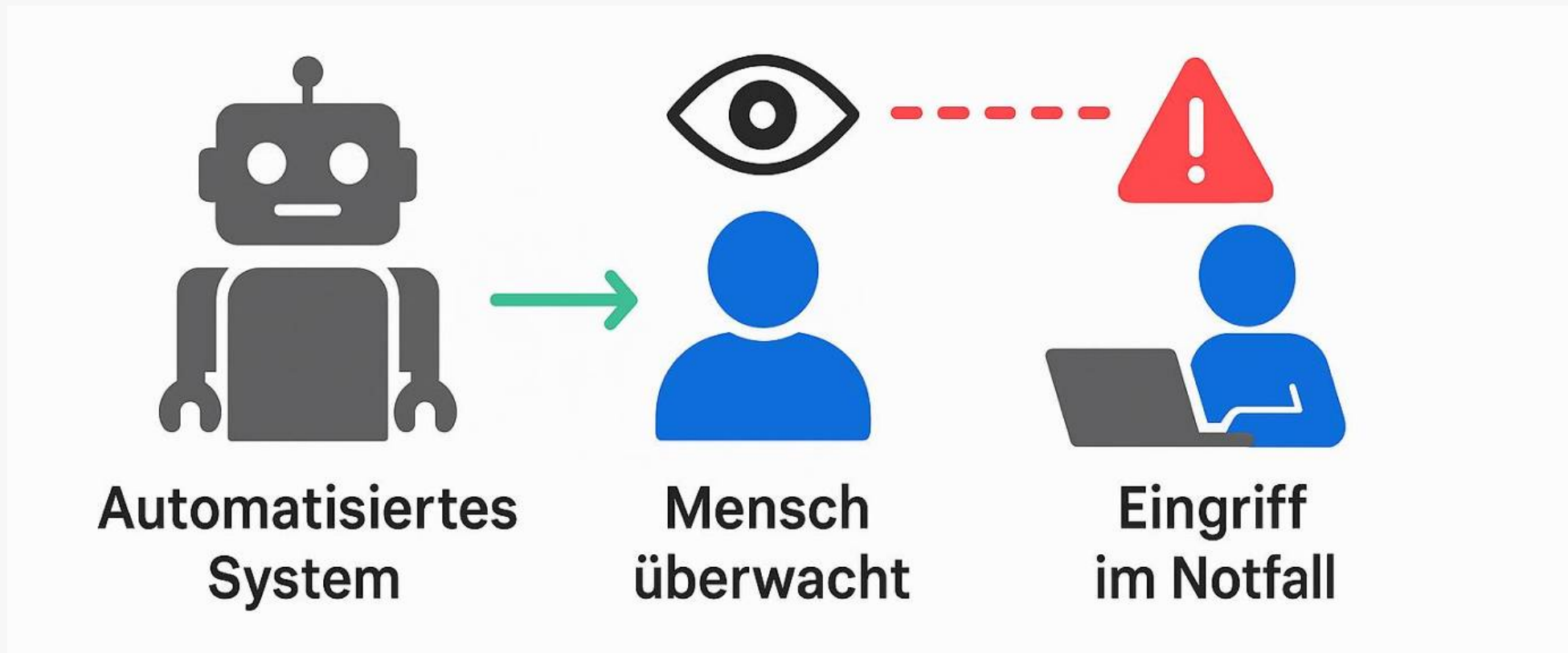
Menschliche Urteilskraft  
und Auswahl

Das endgültige Ergebnis  
des Prozesses



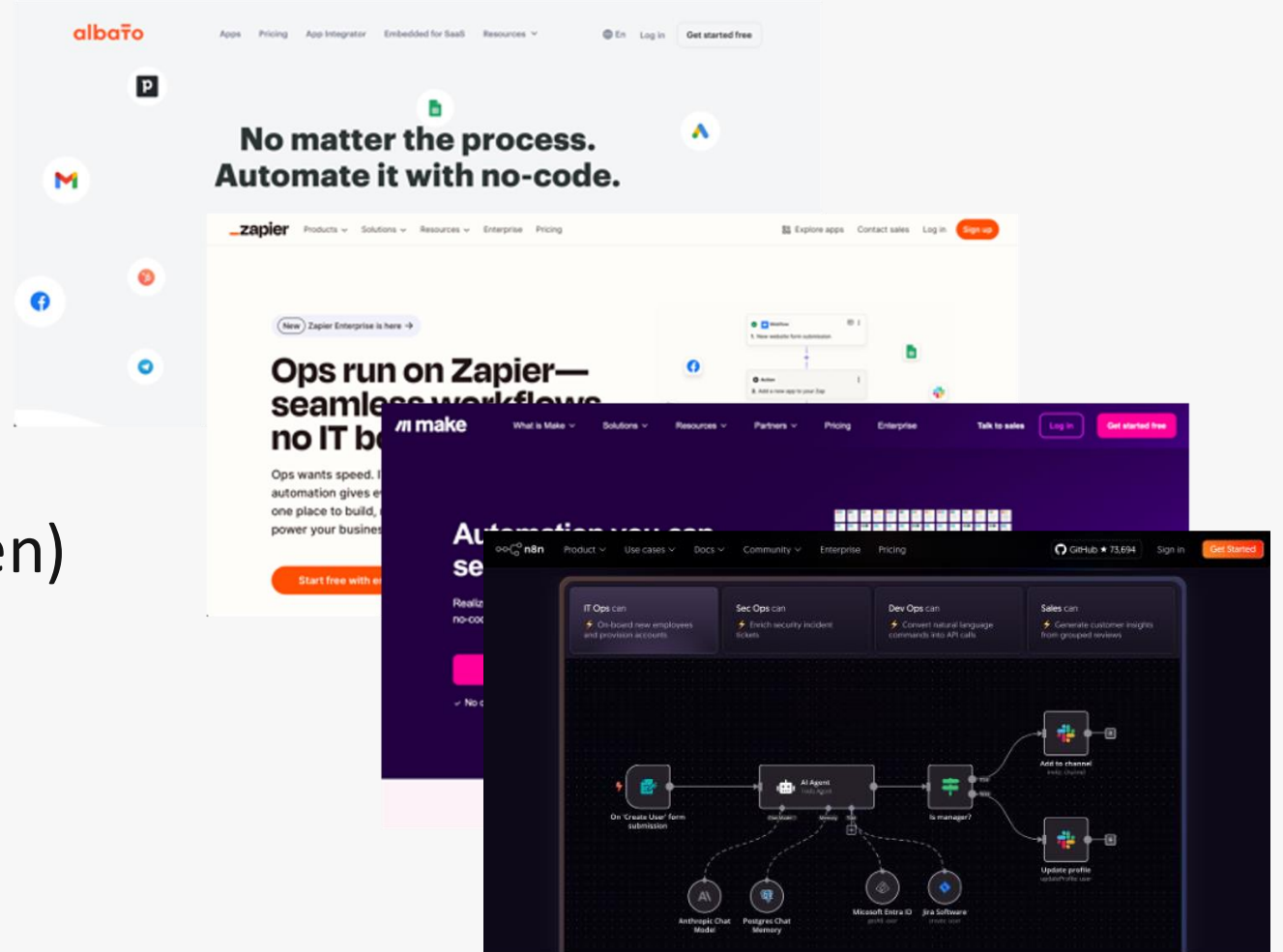
# Human on the Loop

KI arbeitet eigenständig. Mensch überwacht das System und greift nur im Ausnahmefall ein.



# Tools

- ◆ Zapier
- ◆ Make.com
- ◆ Albato (Portugal EU)
- ◆ n8n (kann selbst gehostet werden)
- ◆ Flowise AI (on-premise)



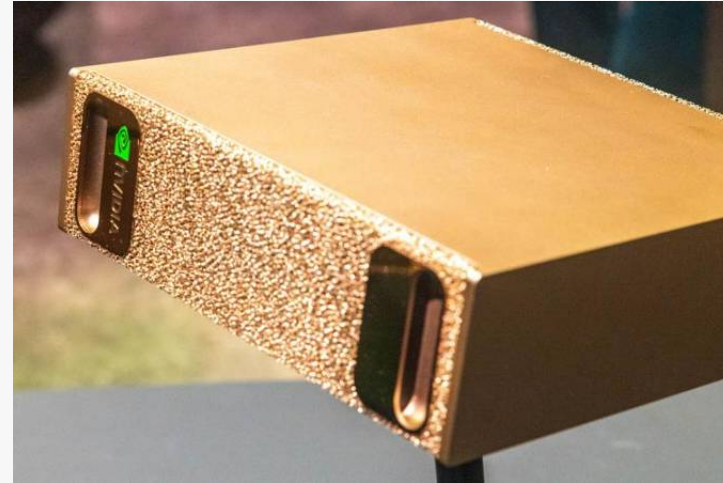
An isometric illustration of a city skyline, where the buildings are represented by various computer hardware components. In the center, a large square chip with internal circuitry and labels like 'MICROPROCESSOR', 'SYSTEM', and 'MEMORY' serves as the focal point. Surrounding it are other components like RAM modules, a tower case, and various peripheral devices. The entire scene is rendered in shades of blue and white, with a soft glow effect.

# Hardware

# Hardware



ASUS Ascent GX10  
ca 2.999 USD



NVIDIA DGX Spark  
ca 3.999 USD

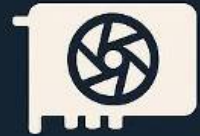
Kleine AI PCs  
Clusterbildung möglich  
pro Einheit bis zu 200B Modelle

Quellen: <https://www.servethehome.com/>  
<https://www.asus.com/>



# Anforderungen Kurz und Knapp

**7B**



Läuft auch unterwegs



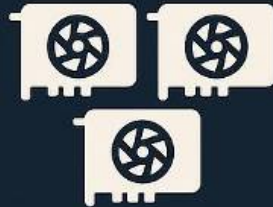
**32B**



Vielseitiger Einsatz



**70B**



Flexible Möglichkeiten



# Hardware Beispiel

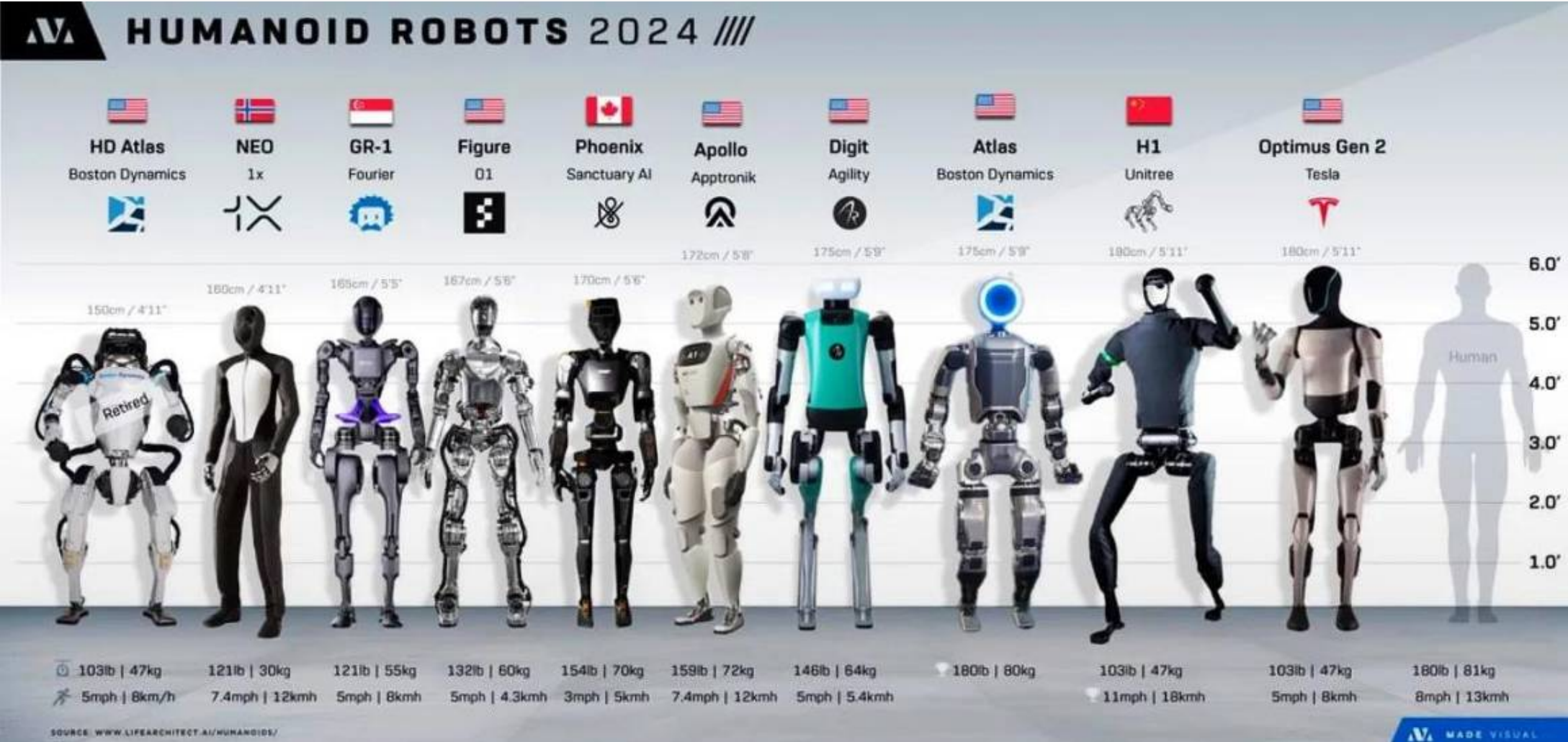
- Abhängig des gewählten Modells und der entsprechenden Anforderungen.
- Jedes Modell hat andere Anforderungen – daher zuerst Anforderungen definieren und dann Hardware organisieren.
- Beispiele:
  - Llama 3.1 405B & Deepseek R1 671B
    - GPU: ca. 16x A100/H100, RAM: 2-4TB
    - 1x GPU A100 8.000-12.000 Dollar....
  - Llama 3.2 3B
    - 16GB Ram, 8GB GPU  
ca. 3GB Speicher



# Ausblick

Wo geht die Reise hin?

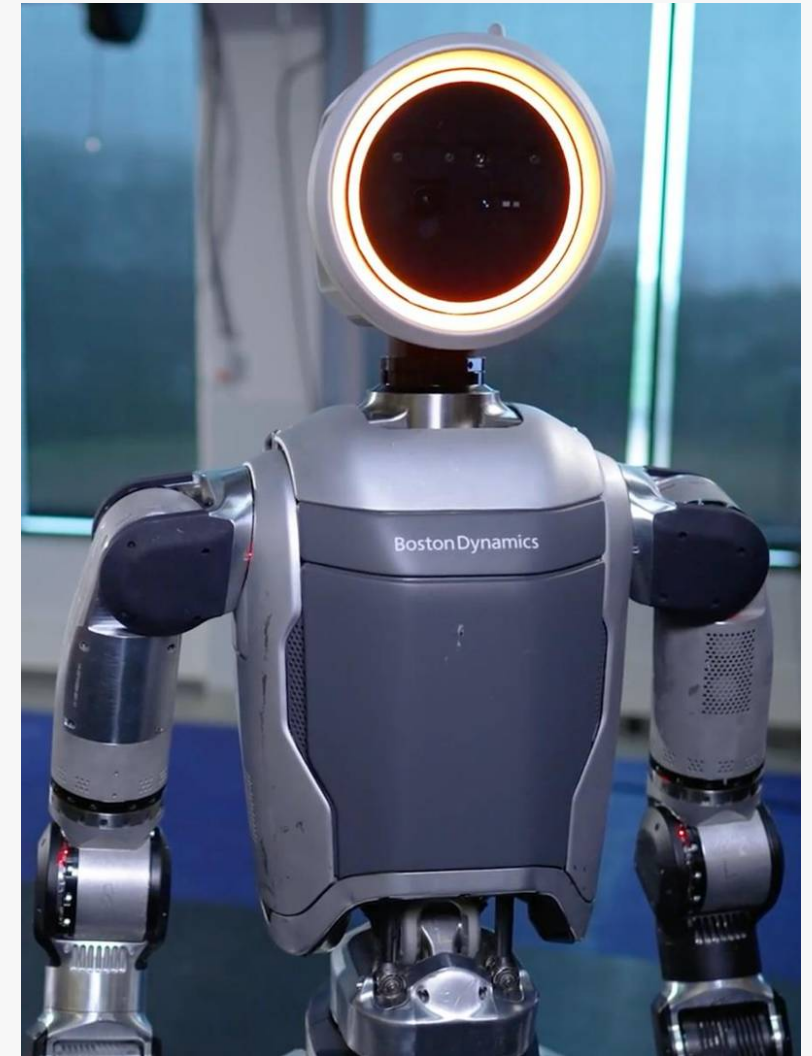
# Humanoid Robots



Quelle: <https://kimbellard.medium.com/ready-for-robots-4560bf51cd4d>

GR-1 Fourier wird von China hergestellt





Quelle: <https://bostondynamics.com/atlas/>

# Unitree – Erster Robot-Fight

[https://youtu.be/WKCDP41\\_4m8?si=an5hRqb5pcnt0FyL](https://youtu.be/WKCDP41_4m8?si=an5hRqb5pcnt0FyL)



# EU – Robots – kleiner Überblick



**Circus Group**  
Hamburg

Roboter Küche  
Aktuell in Tübingen in Klinik im Einsatz

DW Deutschland Doku:  
<https://youtu.be/HirLZbo7fO8>

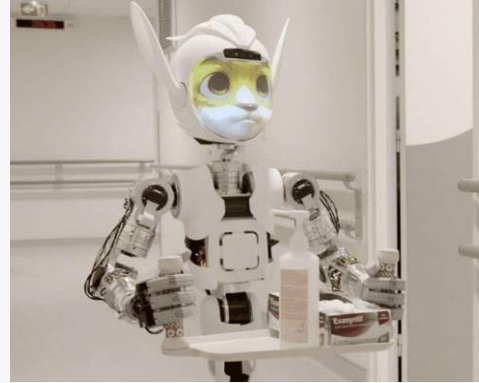
Circus Group:  
<https://www.circus-group.com/ca-1>



**E-Terry**  
Erfurt

Autonome Agrar-Roboter

E-Terry:  
<https://e-terry.de/>



**Enchanted Tools**  
Paris

vielseitiger Roboter für  
Gesundheitswesen,  
Gastgewerbe und  
Einzelhandel

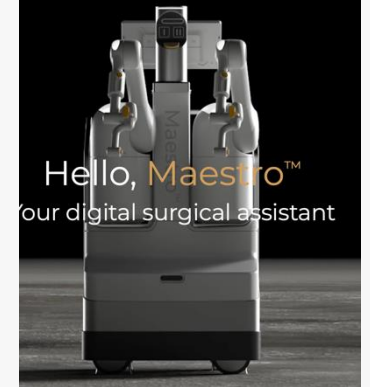
Enchanted Tools:  
<https://enchanted.tools>



**Keyper**  
Barcelona

autonomer Roboter,  
für industrielle  
Inspektionen

Keyper:  
<https://keybotic.com>



**Maestro**  
Paris

chirurgischer Roboter  
zur Unterstützung der  
Laparoskopie.

Moon Surgical:  
<https://moonsurgical.com>

Quelle – EU-Startups: <https://www.eu-startups.com/2024/09/from-fiction-to-function-10-european-startups-pushing-the-boundaries-of-robotics/>



# Shenzhen

<https://youtu.be/VltBlmlWE4U?si=ZFNuDfmYzDR5JyaF>





Let's rock 🤘

Philipp Bellant

<https://techbello.at>

[office@techbello.at](mailto:office@techbello.at)

Tel: 0660 65 144 01



# Bonus I – Musik Generierung

Lokal / Cloud

■ Magenta-RT (Streaming – am Lokalen System)

Suno Alternative:

■ Udio

# Bonus II – Text to Speech / Stimmklonung

Huggingface/Cloud/Lokal:

- TorToiSe-TTS

- F5-TTS

- XTTS-2

- Zyphra – Zonos

- Fish Audio

# Links

## AI-Überischt EU-Tools/Modelle:

[AI-Atlas](#)

## Automatisierungs-Tools (Cloud):

[n8n](#) (DE)

[Albato](#) (Portugal)

[Zapier](#) (USA)

[Make.com](#) (USA)

[FlowiseAI](#) (USA)

## Automatisierungs-Tools (On-Premise):

[n8n](#)

[FlowiseAI](#)

## Modell-Übersicht & Testingplatform:

[Huggingface](#) (EU)

[Ollama](#) (USA) (On-Premise)

## Online-Kurse:

[AIGP](#) (Governance Professional) ([Udemy](#))

[AI: Ethic & Societal Challenges](#) (Coursera)