HammerHAI

Die KI-Fabrik für Industrie, Fertigung, Ingenieurwesen und Forschung

Der neue KI-Supercomputer in Stuttgart

Vortragsabend „Simulierte Welten“ · Stadtbibliothek Stuttgart · 24. November 2026

Dennis Hoppe, HLRS

Über mich

Dennis Hoppe, Höchstleistungsrechenzentrum Stuttgart (HLRS)

1 Heute

- Leiter Converged Computing am HLRS (Supercomputing × KI × Cloud)
- Co-Koordinator von HammerHAI

2 Mein Thema

- Wie man Supercomputer so einfach nutzbar macht wie die Cloud

3 Werdegang

- Informatikstudium in Weimar
- Softwareentwicklung in der Industrie, Jena
- ein Jahr in Neuseeland und Australien (Work & Travel)
- seit 2014 am HLRS





Worum es heute geht

Von der ersten Simulation bis zum neuen KI-Supercomputer in Stuttgart

1. **HLRS & Simulation**
   30 Jahre Supercomputing in Stuttgart
2. **KI verstehen**
   Was steckt hinter ChatGPT & Co.?
3. **Warum Europa handelt**
   Digitale Souveränität und KI-Fabriken
4. **HammerHAI**
   Der neue KI-Supercomputer am HLRS
5. **Ausblick**
   Fazit und Brücke zum nächsten Vortrag



Wie alles begann

Stuttgarts erster Supercomputer: 1986 kam die Cray-2, damals der schnellste Rechner der Welt

1,9 GFlops

Spitzenleistung der Cray-2: knapp 2 Milliarden Rechenschritte pro Sekunde; 1986 Weltspitze

2011

Das iPad 2 ist so schnell wie die Cray-2; ein Supercomputer passt 25 Jahre später in die Hand



2026

HammerHAI schafft bei KI-Rechnungen rund 8 Milliarden Mal mehr als die Cray-2



Was ist das HLRS?

Das Höchstleistungsrechenzentrum Stuttgart; seit 30 Jahren Supercomputing an der Universität Stuttgart

- **Gegründet 1996** als erstes Bundeshöchstleistungsrechenzentrum Deutschlands
- **Einrichtung der Universität Stuttgart**, Gründungsmitglied im Gauss Centre for Supercomputing (mit Jülich und Garching)
- **Für Wissenschaft, Industrie und öffentliche Hand**
- Schwerpunkt **Ingenieurwesen und angewandte Wissenschaften**
- Koordination vieler **nationaler und europäischer Forschungsprojekte**, darunter auch HammerHAI
- **Schulungen, 3D-Visualisierung (CAVE)** und Beratung für Unternehmen



HLRS in Stuttgart-Vaihingen · CAVE · Rechnerraum · Veranstaltungsraum

Von Hawk bis Herder

Die Supercomputer am HLRS: immer stärker GPU-beschleunigt für Simulation und KI



2020

Hawk

Flaggschiff für Simulation von 2020 bis 2025



2025

Hunter

Aktuelles Flaggschiff: CPU und GPU auf einem Chip (AMD)



2026

HammerHAI

KI-Supercomputer für Industrie und Ingenieurwesen



2027/28

Herder

Großer Supercomputer für Simulationen im Neubau HLRS III

Visualisierung HLRS III: Benthem Crouwel Architects & Birk Heilmeyer und Frenzel Architekten

Simulierte Welten

Zwei Wege, die Welt zu berechnen: beide brauchen enorme Rechenleistung

1 Simulation

Wir kennen die Naturgesetze und rechnen sie im Computer nach.

Beispiele: Wetter und Klima, Strömungen, Crashtests

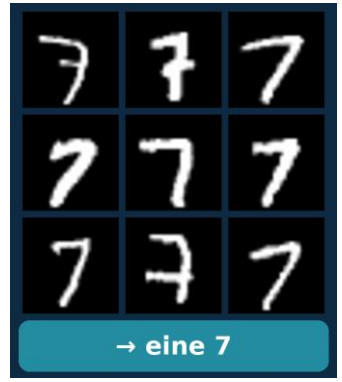


CAVE am HLRS

2 Künstliche Intelligenz

Wir zeigen dem Rechner sehr viele Beispiele und er lernt die Muster selbst.

Beispiele: Handschrift, Sprache, Qualitätsprüfung von Bauteilen



Handschrift-Ziffern (MNIST)

Am HLRS stehen beide Welten nebeneinander: Hunter (bald Herder) für Simulation und HammerHAI für KI.

Simulierte Welten im Kino

„Die Biene Maja“ (2018): die Bilder des Kinofilms wurden am HLRS berechnet

- **Mehr als 230.000 3D-Bilder** hat das Stuttgarter Animationsstudio M.A.R.K.13 auf dem HLRS-Supercomputer "Hazel Hen" berechnen lassen
- **Rendering ist Simulation:** Für jedes Bild berechnet der Rechner, wie Licht in der virtuellen Welt fällt, sich spiegelt und Schatten wirft. Für 3D sogar zweimal, fürs linke und rechte Auge.
- **Hochaufgelöste Animation braucht enorme Rechenleistung:** mit normaler Hardware wäre das zu langsam gewesen
- Fazit: **Auch die Kreativbranche nutzt Supercomputer**



Biene Maja fliegt mit Rechenpower aus Stuttgart

06.03.2018 · Redakteur: Juliana Pfeiffer ·

Die Biene Maja schwirrt mit ihren Freunden seit dem 1. März erneut über die deutschen Kinoleinwände. Für die Berechnung der mehr als 230.000 3D-Bilder des Films hat das Animationsstudio M.A.R.K.13 mit Sitz in Stuttgart die HazelHen, den größten Supercomputer des Höchstleistungsrechenzentrums (HLRS) in Stuttgart genutzt.

Für die Berechnung der mehr als 230.000 3D-Bilder des Films nutzte das Animationsstudio M.A.R.K.13 mit Sitz in Stuttgart die HazelHen, den größten Supercomputer des Höchstleistungsrechenzentrums (HLRS) in Stuttgart.
(Bild: Studio B)

Der Film „Die Biene Maja – Die Honigspiele“ wurde von der in München ansässigen Studio 100 Media und Studio B Animation, in Koproduktion mit dem ZDF, produziert. „Wir haben bereits für den ersten Kinofilm mit der Biene Maja 2014 auf die Rechenkapazitäten des HLRS zurückgegriffen, weil unsere internen Kapazitäten auch aus Zeitgründen nicht ausreichten“, sagt Holger Weiss, CEO von M.A.R.K.13. „Beim zweiten Kinoabenteuer haben wir in Sachen Rendering am HLRS stark

Quelle: industrial-production.de, 06.03.2018

media solution center baden württemberg

Heute bündelt das Media Solution Center Baden-Württemberg solche Projekte mit der Kreativbranche (msc-bw.com)

Crashtest im Rechner

Simulation in der Industrie: Porsche entwickelt Fahrzeuge mit Rechenzeit am HLRS

- Porsche simuliert am HLRS **Crashtests und Aerodynamik**
- **Tausende virtuelle Prototypen** ersetzen teure echte Testfahrzeuge
- **Kürzere Entwicklungszyklen:** neue Modelle kommen schneller auf den Markt
- Vom Ausprobieren in der Werkhalle **zur digitalen Entwicklung**





Und morgen? KI lernt aus solchen Simulationen → HammerHAI

Simulation trifft KI

Beispiel aus dem Ingenieurwesen: KI lernt aus Simulationen und beschleunigt die Entwicklung

1 Simulieren

Supercomputer berechnen viele Varianten eines Bauteils; jede dauert Stunden.

Stunden pro Variante

2 Lernen

Auf HammerHAI lernt ein KI-Modell (einmalig) aus diesen Ergebnissen.

einmal trainieren

3 Anwenden

Das Modell ersetzt die rechenintensive Simulation und schätzt neue Varianten in Sekunden ab. Genau simuliert werden nur noch die besten.

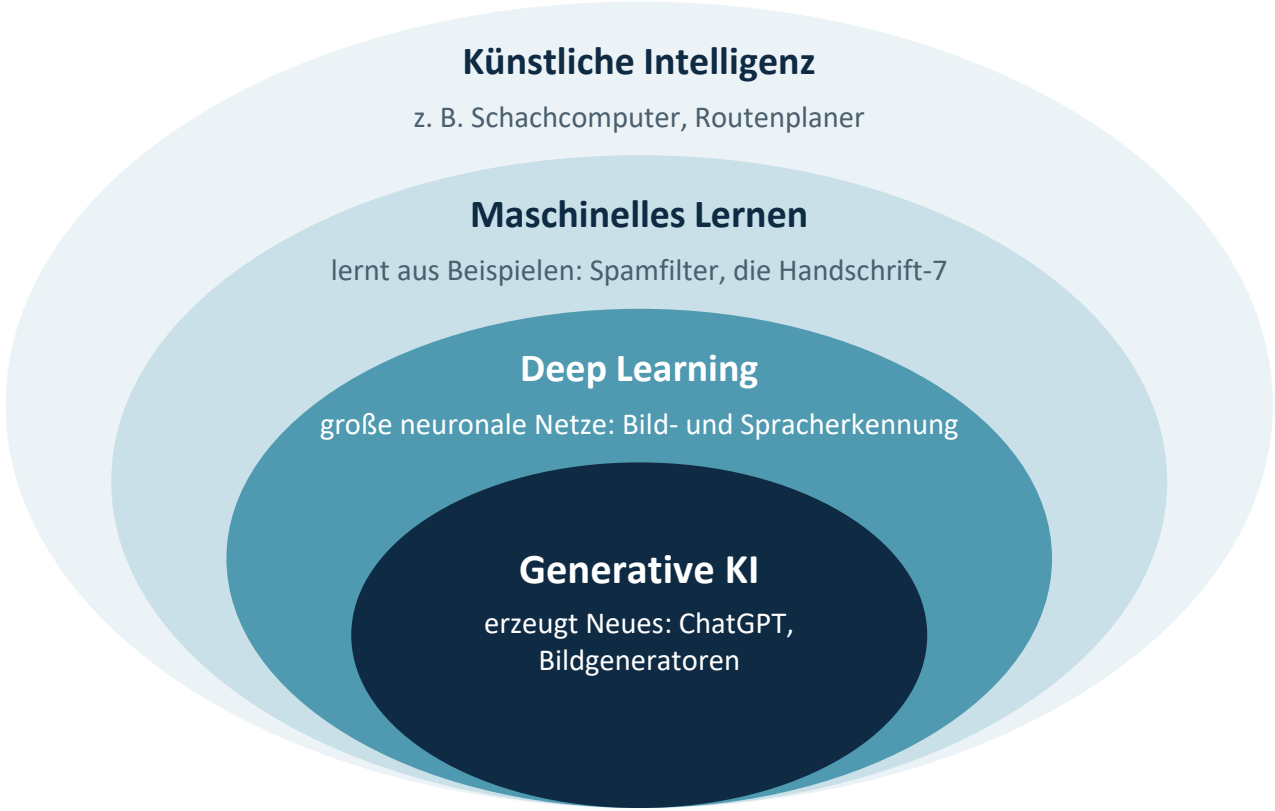
Sekunden pro Variante

Alles an einem Ort:

Simulation auf der HLRS-Infrastruktur, KI-Training auf HammerHAI. Daten müssen den Standort nicht verlassen.

KI im Überblick

Large-Language Models (LLMs) wie ChatGPT bilden nur den innersten Kreis: KI ist jedoch viel mehr



DREI ZUTATEN FÜR JEDE KI

Daten (Big Data)

Ohne viele Beispiele kein Lernen. Daten sind der Rohstoff.

Algorithmen

Die Lernverfahren, z. B. neuronale Netze. Viele davon sind offen verfügbar.

Rechenleistung

Je weiter innen im Kreis, desto mehr wird gebraucht. Dafür gibt es HammerHAI.

Hinter ChatGPT & Co.

Chatbots wie ChatGPT, Claude, Gemini, Le Chat oder DeepSeek

Ein Sprachmodell

Hinter jedem Chatbot steckt ein Modell, das aus riesigen Textmengen gelernt hat, welche Wörter sinnvoll aufeinander folgen.

Gelernt, nicht programmiert

Niemand schreibt die Antworten vor. Das Modell erkennt Muster in Milliarden Beispielen, ähnlich wie wir aus Erfahrung lernen.

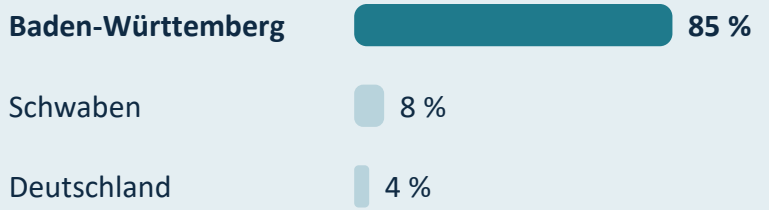
Gebaut auf Supercomputern

Trainiert wird über Wochen und Monate auf tausenden Grafikprozessoren. Genau solche Rechner betreibt auch das HLRS.

SO ARBEITET EIN SPRACHMODELL

Stuttgart ist die Hauptstadt von …

Welches Wort passt als Nächstes?*



Wort für Wort entsteht so die ganze Antwort. *Werte zur Veranschaulichung.

5x pro Jahr: So stark wächst seit 2020 die Rechenleistung, mit der die stärksten KI-Modelle trainiert werden. (Quelle: Epoch AI).

Training, RAG & Co.

Vier Begriffe, die man oft hört, erklärt wie eine Berufsausbildung

Training
Die Ausbildung

Das Modell lernt von Grund auf, aus sehr vielen Beispielen.

Beispiel: Eine Kamera am Fließband lernt aus tausenden Fotos, Kratzer zu erkennen.

Fine-Tuning
Die Spezialisierung

Ein fertiges Modell lernt die Besonderheiten eines Betriebs dazu.

Beispiel: Ein Sprachmodell lernt die Fachbegriffe eines Maschinenbauers.

RAG
Das Nachschlagen

Die KI sucht erst in Ihren Unterlagen und antwortet dann mit Quelle. Wie eine Klausur mit offenem Buch.

Beispiel: Ein Service-Assistent zeigt die passende Seite im Maschinenhandbuch.

Inferenz
Das Anwenden

Das fertige Modell arbeitet: jede Antwort, jede Prüfung.

Beispiel: Jedes Mal, wenn ein Chatbot antwortet oder eine Kamera ein Bauteil prüft.

Training und Fine-Tuning brauchen die meiste Rechenleistung. Genau dafür ist HammerHAI gebaut. Echte Beispiele: hammerhai.eu/de/erfolgsgeschichten

Wer baut die KI-Modelle?

Die meisten Modelle kommen aus den USA und China, offene Modelle holen schnell auf

USA

365 Modelle

ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google)

meist geschlossen: nur über den Anbieter nutzbar

China

248 Modelle

DeepSeek, Qwen (Alibaba), Kimi (Moonshot AI)

meist offen: Modell kann man herunterladen und selbst betreiben

Europa

deutlich weniger

Mistral (Frankreich), SOOFI (Deutschland)

meist offen: gute Basis für eigene Anpassungen

Nur rund 4 Monate liegen die besten offenen Modelle hinter der Spitze. Wer sie auf eigener Infrastruktur betreibt, behält die Kontrolle.

Quelle: Epoch AI. Anzahl dokumentierter KI-Modelle Januar 2025 bis September 2026; Abstand offene zu geschlossenen Modellen seit Januar 2026.

Digitale Souveränität

Souveränität heißt nicht, alles selbst zu machen, sondern selbst entscheiden zu können

Kontrolle

Wo liegen meine Daten, wer kann darauf zugreifen, welches Recht gilt? Bei uns: in Stuttgart, nach europäischem Recht.

Auswahl

Offene Modelle, verschiedene Anbieter, jederzeit wechseln können. Statt Abhängigkeit von einem einzigen Konzern.

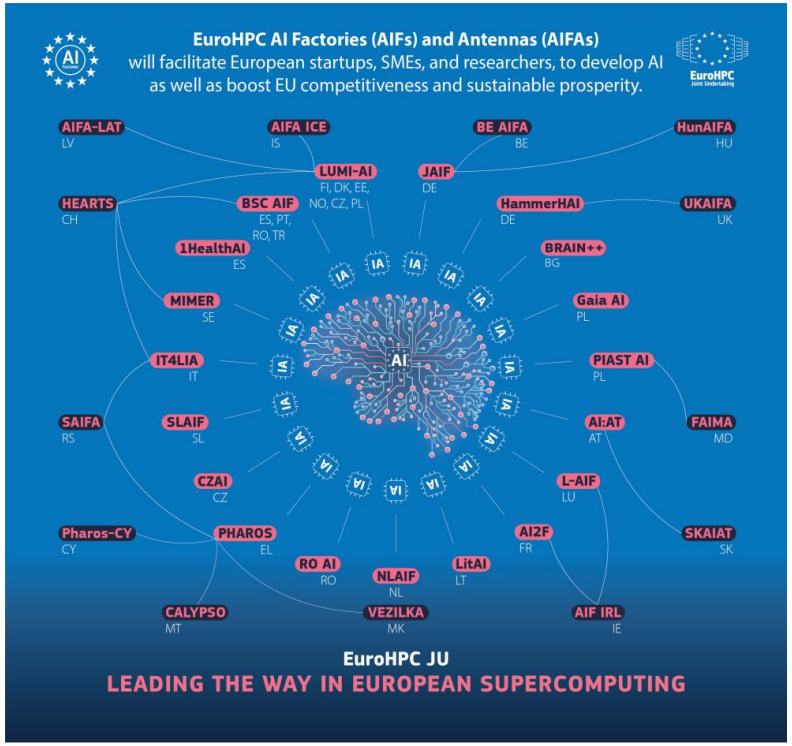
Einfachheit

Souveränität nützt nur, wenn man sie auch bedienen kann: so einfach wie die Cloud, mit Beratung von Anfang an.

Deshalb baut Europa eigene KI-Fabriken: öffentlich, offen für alle und mit den Daten in Europa.

Die Antwort: KI-Fabriken

Koordiniert vom EuroHPC Joint Undertaking: öffentliche KI-Rechenleistung für ganz Europa



Grafik: EuroHPC JU

19

KI-Fabriken, z.B. HammerHAI

13

Antennas, z.B. UKAIFA in Edinburgh

2,6 Mrd. €

von EU und Mitgliedstaaten

2

in Deutschland: Jülich und Stuttgart

Die Idee: KI-Supercomputer mit Daten und Fachleuten, offen für alle

Noch größer: KI-Gigafabriken

HammerHAI im Überblick

Deutschlands KI-Fabrik: souveräne KI-Rechenleistung und Service für Industrie und Forschung

860
NVIDIA-GB200-GPUs in 215 Rechenknoten

15 ExaFlops
KI-Spitzenleistung (INT4)

42 Mio. €
Investition in Hardware

36
Monate Laufzeit, Start 04/2025

5
Projektpartner aus ganz Deutschland

Warum

- Start-ups und KMU fehlen KI-Rechenleistung und Know-how
- Ingenieursdaten sind sensibel und brauchen eine sichere Umgebung in Deutschland
- Klassische Rechenzentren passen nicht zur Arbeitsweise von KI-Teams

Was

- KI-optimierter Supercomputer am HLRS, so einfach nutzbar wie die Cloud
- Alles aus einer Hand: Beratung, Onboarding, Schulung, Daten, Sicherheit

Für wen

- Start-ups und KMU bis hin zu Industrie
- Entwicklungs- und Forschungsteams in Deutschland und der EU
- Öffentliche Hand

Das System

860
NVIDIA-GPUs (GB200)

15 ExaFlops
Spitzenleistung (INT4)

10 PB
Speicher (VAST Data)

Aufbau

Hersteller	HPE
GPU-Knoten	215 Knoten mit je 4 NVIDIA-GPUs (860 GPUs)
Europäische KI-Chips	9 Knoten mit Chips von Axelera AI (ab 2027)
Speicher und Netz	10 PB schneller Speicher (VAST Data) Hochgeschwindigkeitsnetz von NVIDIA

Rund zwei Jahre bis zum Start

04/2025
Öffentliche, europaweite Ausschreibung

02/2026
Vertrag mit Systemlieferanten

Herbst/Winter 2026
Lieferung und Aufbau läuft gerade

Anfang 2027
Erste Nutzer

1. Halbjahr 2027
Regelbetrieb

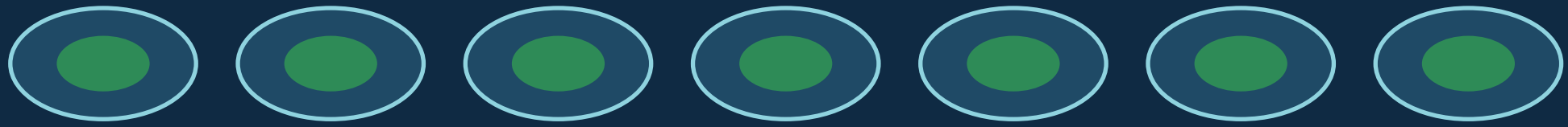
Bestellen und loslegen? Von der Ausschreibung bis zum Regelbetrieb vergehen bei einem System dieser Größe aktuell rund zwei Jahre.

Wie schnell ist das?

15 ExaFlops KI-Spitzenleistung, übersetzt in Smartphones

So viel wie 430.000 Smartphones

Das sind 7 voll besetzte Stadien wie die MHPArena, in denen jeder Fan ein aktuelles Handy in der Hand hält.

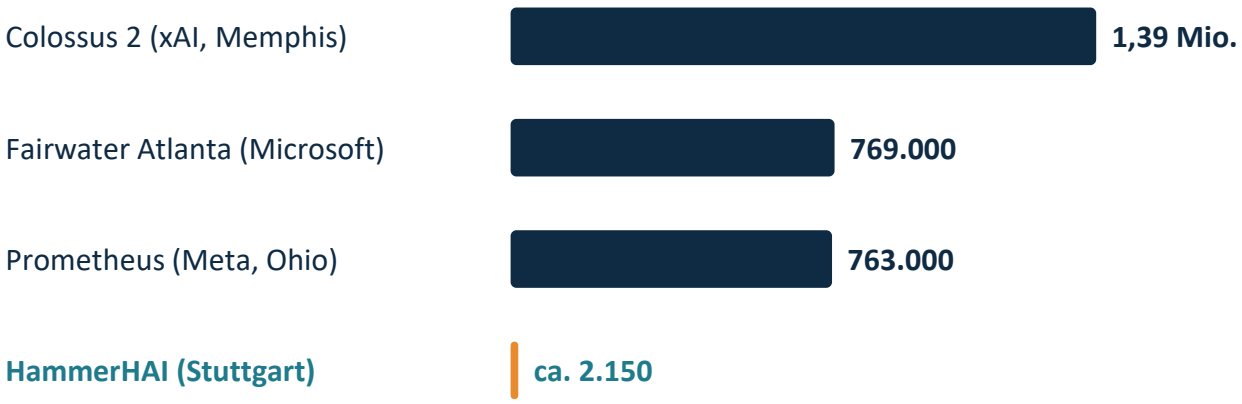


Rechnung: $15 \times 10^{18} \div$ rund 35×10^{12} KI-Rechenschritte pro Sekunde (aktuelles iPhone) ≈ 430.000 · MHPArena: 60.449 Plätze · Spitzenwerte, nur als Größenordnung

HammerHAI im Vergleich

Die größten KI-Rechenzentren der Welt stehen in den USA, und sie sind um ein Vielfaches größer

Rechenleistung in „H100-Äquivalenten“ (Grafikprozessoren)



Das größte KI-Rechenzentrum hat rund 650-mal so viel Rechenleistung wie HammerHAI.

Wie lange dauert ein Training?

- **GPT-4 (2023):** rund 25.000 Grafikprozessoren, etwa 90 bis 100 Tage
- **Größtes Rechenzentrum heute:** über 20 solcher Trainings pro Monat
- **HammerHAI:** knapp ein Jahr für ein Modell in GPT-4-Größe
- **ChatGPT, Claude, Gemini, Mistral heute:** Dauer und Aufwand werden nicht veröffentlicht, Schätzungen liegen deutlich höher

HammerHAI muss nicht mit den Größten mithalten, sondern setzt den Fokus auf spezialisierte Modelle für Industrie und Forschung. Für noch größere Trainings plant Europa KI-Gigafabriken.

Quellen: Epoch AI (Nov. 2026). HammerHAI: eigene Schätzung, je GPU ca. 2,5 H100-Äquivalente; Trainingsdauer als Größenordnung.

Was HammerHAI kann

Groß genug für ein eigenes deutsches Sprachmodell und für hunderte Anpassungen gleichzeitig

2 Wochen

bräuchte HammerHAI für SOOFI S, trainiert von Grund auf, mit der ganzen Maschine.

- **Was:** offenes Sprachmodell aus Deutschland, entwickelt von einem Konsortium aus Forschung und Start-ups, gefördert vom Bund
- **Größe:** 30 Milliarden Parameter
- **Training:** 27 Billionen Wortbausteine, vor allem Deutsch und Englisch; bis zu 512 Grafikprozessoren, rund 253.000 GPU-Stunden
- **Ergebnis:** im Deutschen das stärkste offene Modell seiner Größe; als Basis zum Anpassen gedacht

Hunderte

Fine-Tunings gleichzeitig

Rechnerisch passt auf jeden der 860 Grafikprozessoren ein Fine-Tuning: Ein großes offenes Modell lässt sich so in rund einem Tag an eigene Daten anpassen.

RAG

braucht kein Training

Nachschlagen in eigenen Dokumenten läuft auf wenigen Grafikprozessoren, nur für die Antworten.

Unternehmen müssen nicht bei null anfangen: Sie passen offene Modelle an ihre Daten an. Dafür ist HammerHAI genau richtig dimensioniert.

Quellen: SOOFI-Konsortium / TU Darmstadt (Juni 2026); GPU-Stunden laut Fachpresse; Dettmers et al., QLoRA (2023). HammerHAI-Werte: eigene Schätzungen.

Die Anlieferung

November 2026: Die GPU-Knoten, das Herzstück von HammerHAI, kommen am HLRS an











Wie geht es weiter?

Verkabeln, Netzwerk anschließen, testen; Einweihung im Februar 2027 und Regelbetrieb erwartet im ersten Halbjahr 2027.

Zugang zu HammerHAI

Offen für Wissenschaft und Wirtschaft: von freier Forschung bis zum kommerziellen Einsatz

vertraulichere Daten + höhere Sicherheit + bezahlter Zugang

	Forschung *Hochschulen, Institute und Unternehmen in Forschungsprojekten*	**Pilot und Machbarkeitsstudie** *Wo Start-ups und KMU meist starten*
Zweck	Wissenschaftliche Forschung und vorwettbewerbliche Entwicklung	KI mit eigenen Daten testen, bevor investiert wird
Ergebnisse	Werden veröffentlicht	Bleiben im Unternehmen
Daten	Meist teilbar	Unternehmensdaten, oft sensibel
Zugang und Kosten	Kostenfrei, da Ergebnisse veröffentlicht werden	Für KMU kostenfrei möglich, nach Beihilferecht
Beispiel	Ein Forschungsinstitut trainiert ein digitales Modell einer Fabrikhalle	Ein Start-up testet, ob KI Arbeitsschritte in Produktionsvideos erkennt

Ihr Weg zu HammerHAI

Der Concierge-Service: ein fester Ansprechpartner, von der ersten Idee bis zur fertigen Lösung

1

Anfragen

Kurze E-Mail: wer Sie sind, was Sie vorhaben, was Sie brauchen.

2

Gespräch

Persönliches Gespräch mit dem Concierge: Ziel, Daten, Rechenbedarf und mögliche Kosten klären.

3

Freischalten

Nutzungsbedingungen unterschreiben, Zugang erhalten, Einführung und Schulung bei Bedarf.

4

Loslegen

Mit fertigen Modellen und Software-Vorlagen. Schon jetzt auf Partnersystemen (u.a. KISSKI) testen.

access@hammerhai.eu

hammerhai.eu/de/erste-schritte

Rechenzeit auf HammerHAI ab Anfang 2027; Beratung und Tests schon heute

Kostenfrei für Forschung und Pilotprojekte von KMU und Start-ups (nach EU-Regeln); kommerzielle Nutzung gegen Entgelt

Rolle in Deutschland

HammerHAI im deutschen KI-Ökosystem: Unsere Rolle ist das Skalieren

Kennenlernen

KI-Servicezentren, Hochschulen und Netzwerke machen KI bekannt

Erproben

Erste Machbarkeitsstudien mit eigenen Daten

Skalieren

HammerHAI: große Modelle, große Datenmengen, sichere Umgebung

Betreiben

Dauerbetrieb in der Cloud oder auf eigener Hardware

Fazit und Ausblick

1. **Das HLRS rechnet seit 30 Jahren vorne mit**

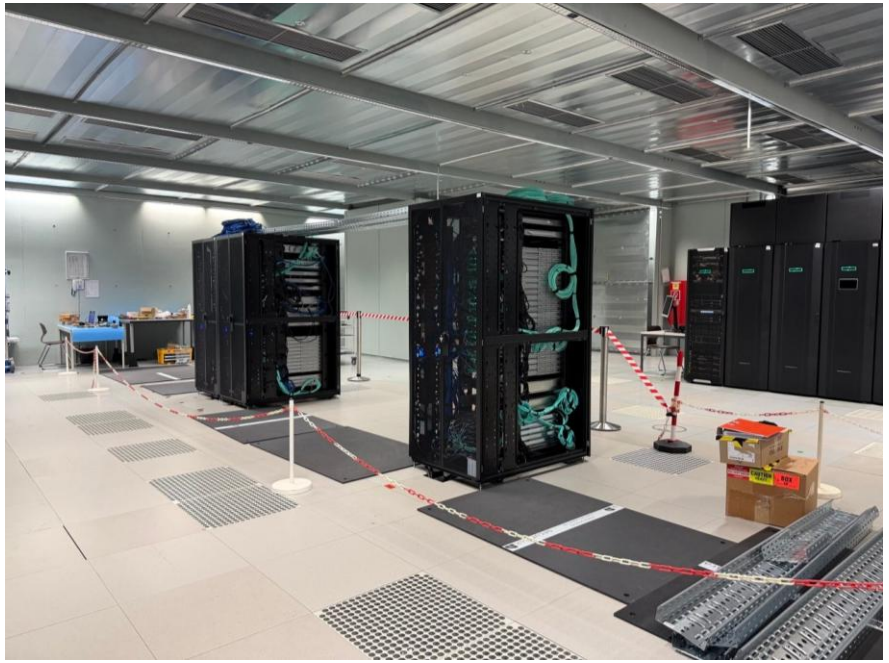
   Von der Cray-2 über Hunter bis zu HammerHAI und Herder: Supercomputing für Wissenschaft und Industrie

2. **Souveräne KI-Rechenleistung für Europa**

   Als eine von 19 europäischen KI-Fabriken: ausgewählte Pilotnutzer ab Anfang 2027, Regelbetrieb im ersten Halbjahr 2027 erwartet

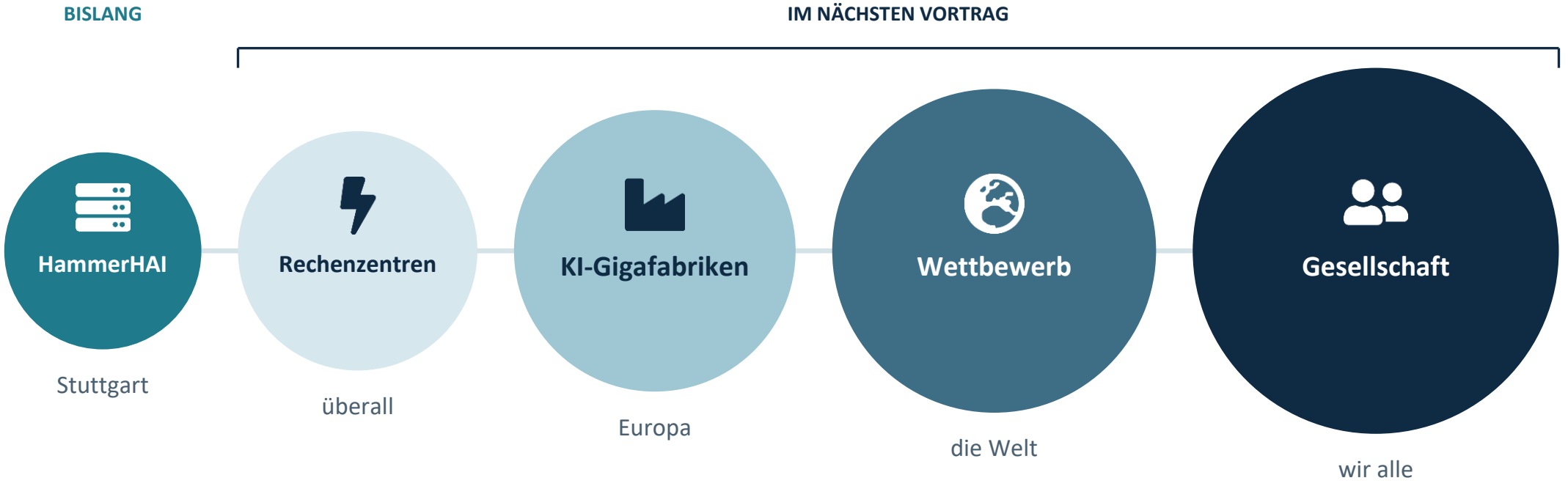
3. **Für Start-ups, Mittelstand und Industrie**

   Mit Beratung, Schulung und sicherer Datenhaltung: die Daten bleiben in Deutschland



Aufbau von HammerHAI am HLRS

Und jetzt: das große Bild



Wir zoomen heraus: von einer Maschine in Stuttgart bis zur ganzen Gesellschaft

Vielen Dank!

Interesse? access@hammerhai.eu

hammerhai.eu/de/erste-schritte

Beschaffung und Betrieb des KI-optimierten EuroHPC-Supercomputers werden gemeinsam vom EuroHPC Joint Undertaking über das Programm „Digitales Europa“ der Europäischen Union sowie vom Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) und vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg finanziert.

Die geäußerten Ansichten sind ausschließlich die der Autoren und spiegeln nicht unbedingt die der Europäischen Union oder des EuroHPC JU wider. Dieses Projekt wird vom European High Performance Computing Joint Undertaking unter der Finanzhilfevereinbarung Nr. 101234027 gefördert. Es wird kofinanziert von der Europäischen Kommission, dem Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR), dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg, dem Bayerischen Staatsministerium für Wissenschaft und Kunst sowie dem Niedersächsischen Ministerium für Wissenschaft und Kultur.