„Inferenz NRW“ ist gestartet: Souveräne KI-Modelle für die Hochschulen in Nordrhein-Westfalen

Um dem stark wachsenden Bedarf an Sprachmodellen seitens der Hochschulen in NRW gerecht zu werden, stehen seit dem 1.4.2026 verschiedene Modelle, insbesondere zur Chatbotnutzung, für alle Hochschulen unter dem Projekttitel „Inferenz NRW“ als Alternative zu anderen kommerziellen und nicht kommerziellen Angeboten bereit. Die Bereitstellung erfolgt über KI:connect. Bis auf weiteres sind die Modelle aufgrund der zentralen Finanzierung durch das MKW ohne weitere Kosten für die Hochschulen in NRW nutzbar.

Open Weight Modelle für NRW

Derzeit stehen folgende Modelle bereit:

- Mistral Small 4 (Mistral AI)
- GPT OSS (Open AI)
- Qwen3 Embedding (Alibaba Cloud)
- E5 Mistral 7B Instruct (intfloat)

Die Modellangebote und deren Verfügbarkeit werden sukzessive erweitert, sobald weitere Hardware-Ressourcen zur Verfügung stehen.

Verbund aus 2 Projekten und 3 Universitäten

Vorbereitet, umgesetzt und bereitgestellt wird Inferenz NRW von einer gemeinsamen Arbeitsgruppe bestehend aus Vertreter:innen des Projekts Open Source-KI.nrw, (gemeinsam betrieben von Universität zu Köln (UzK) und Ruhr-Universität Bochum) sowie dem Projekt KI:connect.nrw (betrieben von der RWTH Aachen). Unterstützt wird das Vorhaben von WestAI.

Inferenz NRW zusammengefasst:

- **Gebunden an KI:connect:** Angeboten werden KI-Modelle für Hochschulen, die KI:connect bereits einsetzen oder einführen.
- **Chatbot und API:** neben der unmittelbaren Chatbot Nutzung stehen die Modelle per API für lokale Anwendungen in den Hochschulen zur Verfügung.
- **Europäische Modelle:** Angeboten werden nicht-kommerziell bereitstellbare Modelle mit dem Fokus auf europäische Anbieter.
- **Angebot aktuell nach dem Best effort-Prinzip:** Inferenz NRW wird sukzessiv zu einem vollständigen Service mit entsprechender Servicequalität ausgebaut.
- **Limits:** Die (derzeit noch) knappen Ressourcen erfordern eine entsprechende Limitierung, die für aktuelle Nutzung insbesondere für Chatbots ausreicht.
- **Modellpalette und Verfügbarkeit:** Beide Punkte hängen von weiteren Hardwareressourcen ab. Die Erweiterung der Infrastruktur wird derzeit von der

RWTH und der UzK vorbereitet und die erste Erweiterung ist für die 2. Jahreshälfte 2026 vorgesehen. Derzeitige Limits werden dann angehoben.

- **Service:** Die Kommunikation zur grundsätzlichen Nutzung und Bereitstellung des Angebots erfolgt über KI:connect. Rückfragen bzgl. der verwendeten KI-Modelle beantwortet OSKI.nrw.
- **KI:Expertisezentrum**: Das Angebot Inferenz NRW wird ab dem Sommer 2026 in das Projekt KI:Expertisezentrum.nrw überführt und von dort betrieben und fachlich begleitet.
- **Keine Bereitstellungskosten:** Inferenz NRW ist bis auf Weiteres für die Hochschulen in NRW zentral finanziert.

Weitere Informationen & Kontakt:

Chatbot / KI-Bereitstellung für NRW

Web: kiconnect.nrw
E-Mail: kiconnect@cls.rwth-aachen.de

Informationen zu aktuell bereitgestellten KI-Modellen und Clusterbetrieb

Web: oski.nrw
E-Mail: kontakt@oski.nrw

Inferenz NRW ist eine Initiative von:

OSKI.nrw · WEST AI KI-Servicezentrum · Ki:connect.nrw · DIGITALE HOCHSCHULE NRW

RUHR UNIVERSITÄT BOCHUM · RUB · UNIVERSITÄT ZU KÖLN · RWTH AACHEN UNIVERSITY

gefördert durch:
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen