



RUB

RWTH AACHEN
UNIVERSITY



UNIVERSITÄT
ZU KÖLN

RUHR-UNIVERSITÄT BOCHUM

VOM LMS-PLUGIN-PROTOTYPEN ZUR HOCHSCHULSOUVERÄNEN KI-INFRASTRUKTUR

Das LMS als Innovationstreiber. Eine Hands-On
Chronologie aus Nordrhein-Westfalen.



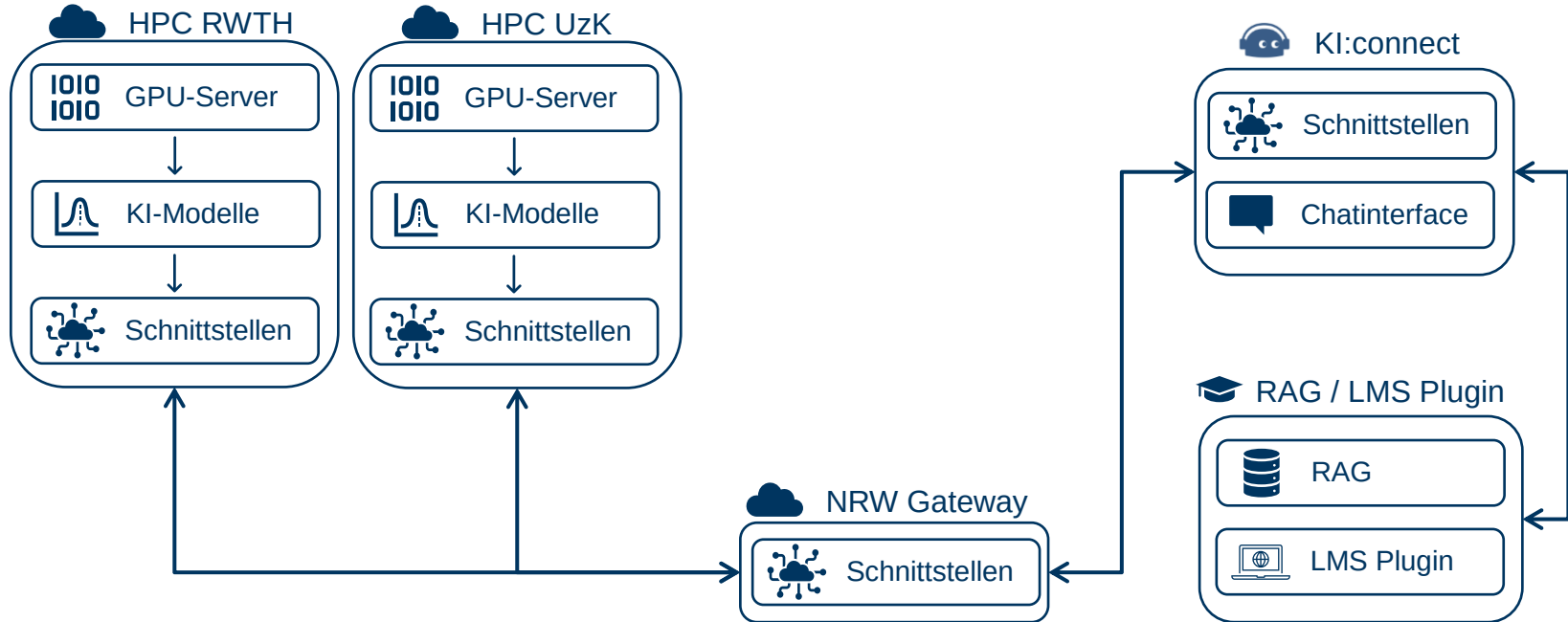
Gefördert durch:

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen



Wo wir (Hochschulen in NRW) heute stehen

- **KI:Expertisezentrum.nrw** im Juli 2026 gestartet; konsolidiert mehrere Landesprojekte



Generative KI in den Kinderschuhen

- **November 2022:** Launch ChatGPT
- **Juli 2023:** Release Llama 2
 - Erstes „relevantes“ Open-Weight LLM
- **Oktober 2023:** Funktionaler Prototyp zum Betrieb eines LLMs an RUB
 - LLM-Betrieb auf ausrangiertem GPU-Server (Llama 2)
 - Kopplung an LMS (Moodle) via RAG-Plugin-Prototyp
- Zentrale Motivation: Generative KI nicht „einfach so“ an die Hochschule bringen, sondern mit einem konkreten Anwendungsfall!



vLLM

vLLM v0.1.0

WoosukKwon released this Jun 20, 2023

v0.1.0

67d96c2

SGLang

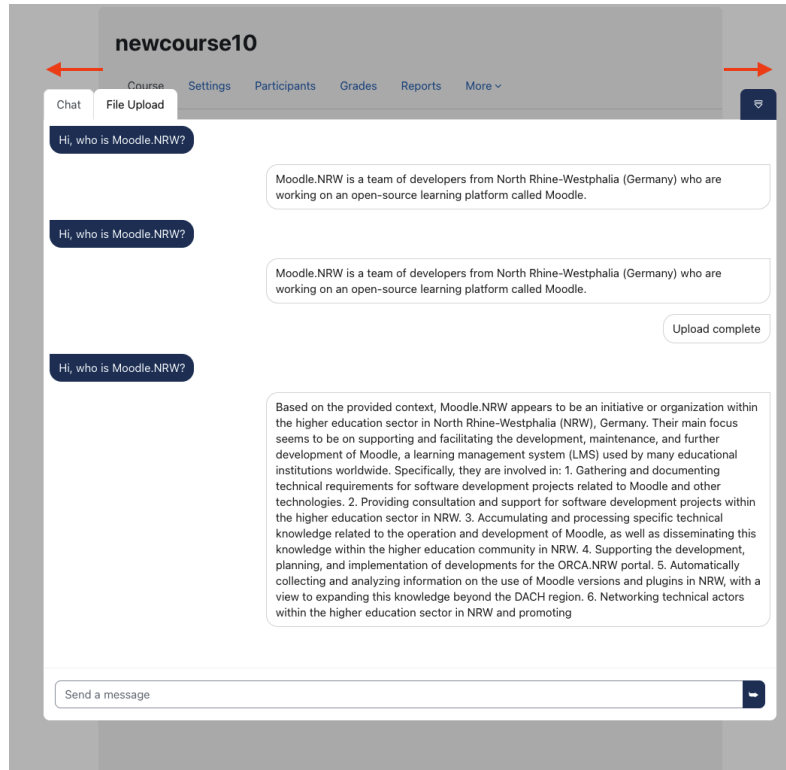
Release v0.1.5

merrymercy released this Jan 18, 2024

v0.1.5

22ec7bc

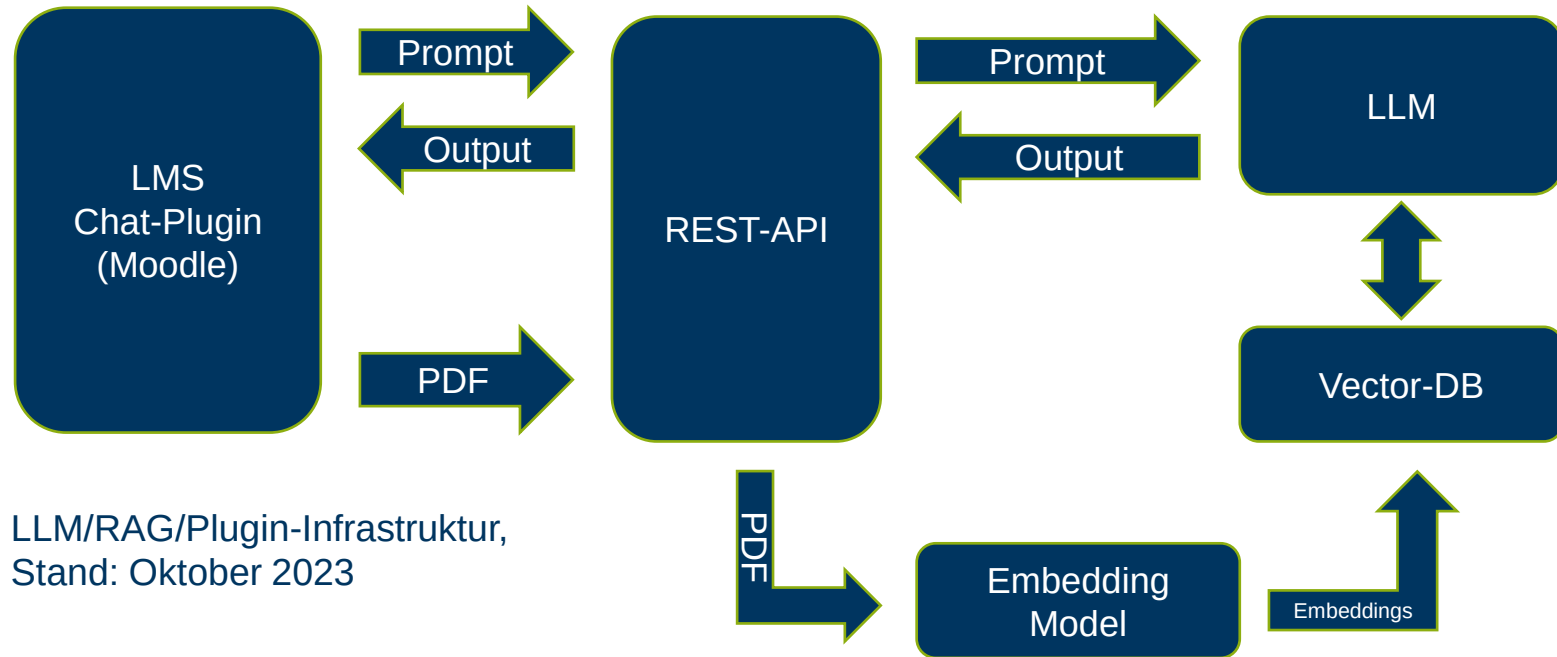
Moodle RAG Prototyp – V1



Moodle Plugin, Stand: Oktober 2023



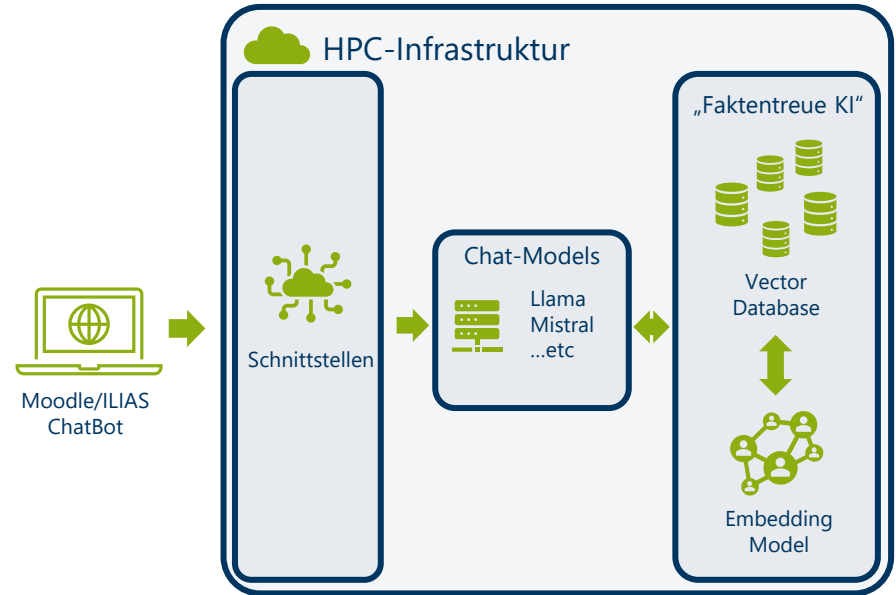
Moodle RAG Prototyp – V1



LLM/RAG/Plugin-Infrastruktur,
Stand: Oktober 2023

Vom LMS-Plugin-Prototypen zum (Vor-)Projekt

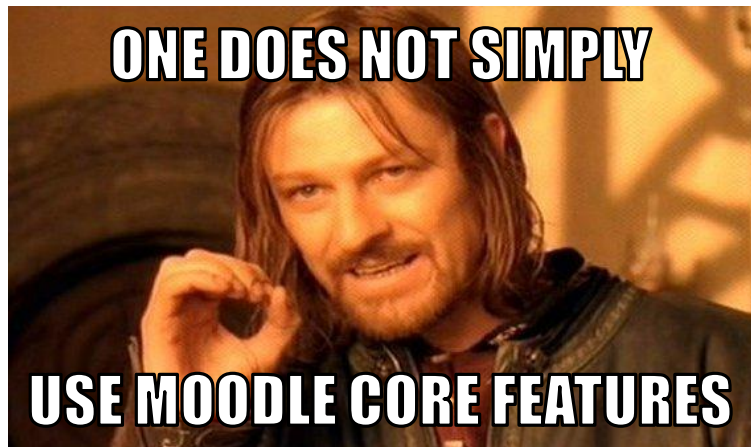
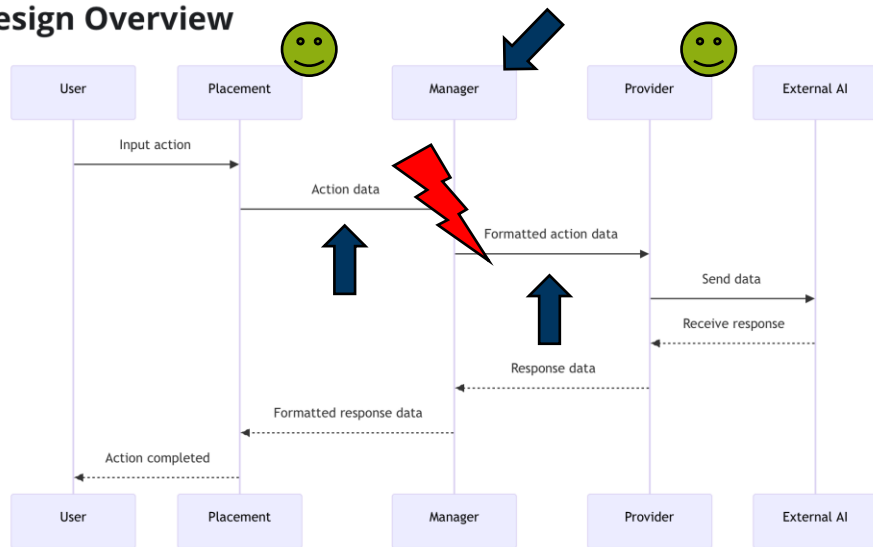
- **Dezember 2023:** Ausarbeitung Vorprojekt-Antrag [Open Source-KI.nrw](#)
 - Basierend auf LMS-Prototypen
- **Juli 2024:** Projektstart [Open Source-KI.nrw](#)
 - Konsortium aus RUB & Uzk
 - Aufbau prototypischer KI-Infrastruktur für Inferenzbetrieb im HPC der Uzk
 - Erprobung durch Entwicklung von RAG-System und LMS-Integration
- Trotz spezifischem Usecase von Beginn an „größer“ gedacht
 - Infrastruktur skalierbar umgesetzt
- **Juli 2025:** Test-Inferenzbetrieb für RUB & Uzk



Die LMS-Aufholjagd

- Moodle AI Subsystem ab Moodle 4.5 (Oktober 2024)
 - Erstellen von Texten im Editor / Zusammenfassen von Texten im Kurs-UI
- Aber: von der Architektur ungenlenk, für RAG-Usecase ungeeignet

Design Overview



Die LMS-Aufholjagd / Moodle RAG Prototyp – V2

The screenshot shows a Moodle interface with a chat window titled "Speech and Language Processing". The chat window has a header "Course chat" and a question: "What is, according to Jurafsky and Martin, the difference between word type and word instance?". The answer provided is:

According to Jurafsky and Martin, the difference between word type and word instance is as follows:

- **Word type:** This refers to the number of distinct words in a corpus, often denoted as $|V|$. For example, in the sentence "They picnicked by the pool, then lay back on the grass and looked at the stars," the word types are "They," "picnicked," "by," "the," "pool," "then," "lay," "back," "on," "grass," "and," "looked," "at," "stars." Even though some words are repeated, each distinct word is counted as one type.
- **Word instance:** This refers to the total number of running words in a corpus, often denoted as N . In the same sentence, the word instances would be counted as 16, considering each occurrence of a word, including repetitions.

So, in summary:

- Word type: distinct words (vocabulary size, $|V|$)
- Word instance: total number of words (N)

At the bottom of the chat window, there is a text input field "Ask about course content..." and a "New section" button.

word type

word instance

Perhaps most important, in thinking about what is a word, we need to distinguish two ways of talking about words that will be useful throughout the book. **Word types** are the number of distinct words in a corpus; if the set of words in the vocabulary is V , the number of types is the **vocabulary size** $|V|$. **Word instances** are the total number N of running words.¹ If we ignore punctuation, the picnic sentence has 14 types and 16 instances:

They picnicked by the pool, then lay back on the grass and looked at the stars.

We still have decisions to make! For example, should we consider a capitalized string (like *They*) and one that is uncapitalized (like *they*) to be the same **word type**? The answer is that it depends on the task! *They* and *they* might be lumped together as the same type in some tasks where we care less about the formatting, while for other tasks, capitalization is a useful feature and is retained. Sometimes we keep around two versions of a particular NLP model, one with capitalization and one without capitalization.

So far we have been talking about **orthographic words**: words based on our English writing system. But there are many other possible ways to define words. For example, while orthographically *I 'm* is one word, grammatically it functions as two words: the subject pronoun *I* and the verb *'m*, short for *am*.

¹ In earlier tradition, and occasionally still, you might see word instances referred to as word *tokens*, but we now try to reserve the word *token* instead to mean the output of subword tokenization algorithms.

Speech and Language Processing. Daniel Jurafsky & James H. Martin.
Draft of August 24, 2025.

Moodle-Plugin, Stand: August 2025

Dezentrale hochschulsouveräne KI-Infrastruktur

- **Juni / Juli / August 2025:** Austausch mit WestAI (KI-Servicecenter für Wirtschaft & Forschung, RWTH) zu möglicher dezentraler hochschulsouveräner KI-Infrastruktur
- **Februar 2026:** Beginn Entwicklung eines Gateways für zentralen Zugriff auf dezentrale KI-Infrastruktur
 - Fokus auf leichte Anbindung weiterer Rechenzentren, Durchsatz, Datenschutz
- **April 2026:** Start „best effort“ dezentrales souveränes KI-Inferenzangebot für NRW-Hochschulen
 - Enge Kopplung an KI:connect als Interface / Middleware
- **Juli 2026:** Start KI:Expertisezentrum.nrw
 - Infrastruktur-Bemühungen von RUB, RWTH und UzK gehen in das Handlungsfeld Inferenz über
- **September 2026:** Erste Konsolidierung der Infrastruktur(en), neue Modelle, neue Modalitäten

...und das LMS?

- Weiterentwicklung des RAG-Systems & LMS Plugins im KI:Expertisezentrum.nrw
 - RAG-System ist Anwendungsagnostisch; kann für andere Anwendungsszenarien adaptiert werden
- Produktiver Rollout an der RUB im WS 26/27
 - Inferenz vollständig auf hochschulsouveräner Infrastruktur
- RAG-System & Moodle Plugin sind Open Source: <https://gitlab.git.nrw/ki-expertisezentrum.nrw>

Dokumenten Chat ⓘ Löschen Mistral Small 4 119B KI-Inferenz.nrw

Indexierte Dokumente: 8

verwendet werden.

- Dürfen **nicht** für Kausalzusammenhänge interpretiert werden (Confounder möglich!). Aus Korrelationen dürfen keine Kausalzusammenhänge abgeleitet werden^[1].

Zusammenfassung der Entscheidungshilfe

Frage	Pearson ((r))	Spearman ((r_s))
Ist der Zusammenhang linear?	✓ Ja	✗ Nein (aber monoton)
Sind die Daten metrisch skaliert?	✓ Ja	✓ Ja (auch ordinal)
Gibt es Ausreißer?	✗ Nein (verzerrt)	✓ Ja (robust)
Soll ein nicht-linearer Trend geprüft werden?	✗ Nein	✓ Ja

Falls du ein konkretes Beispiel oder eine Berechnung besprechen möchtest, lass es mich wissen!

[Zitierte Quellen anzeigen](#)

Stellen Sie eine Frage zu den hochgeladenen Dokumenten... ➤

RUHR
UNIVERSITÄT
BOCHUM

RUB