Repositorien und KI

Dr. Gernot Deinzer

Universität Regensburg
UNIVERSITÄTSBIBLIOTHEK

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Vorüberlegungen

KI – LARGE LANGUAGE MODELLS (LLMS)

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Funktionsweise LLM

Zerlegung des Textes in Bausteine (Tokens)
- Wörter oder Silben

Einbettung
- Zuordnundg von Wort zu Vektor
- Neuronalen Netzen (Transformator Modell)

Vorhersage des nächsten Tokens
- Neuronale Netze
- Erkennung von Eingaben
- Berechnung von Wahrscheinlichkeiten
- Vorraussetzung Training

Dekodierung
- Auswahl eines Tokens
- Wahrscheinlichkeitsberechnung

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Zerlegung eines Textes in Bausteine

Repositorien spielen eine wichtig Rolle in der KI-Land
schaft.

Erstellt mit https://tiktokenizer.vercel.app/

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Funktionsweise LLM

Zerlegung des Textes in Bausteine (Tokens)
- Wörter oder Silben

Einbettung
- Zuordnundg von Wort zu Vektor
- Neuronalen Netzen (Transformator Modell)

Vorhersage des nächsten Tokens
- Neuronale Netze
- Erkennung von Eingaben
- Berechnung von Wahrscheinlichkeiten
- Vorraussetzung Training

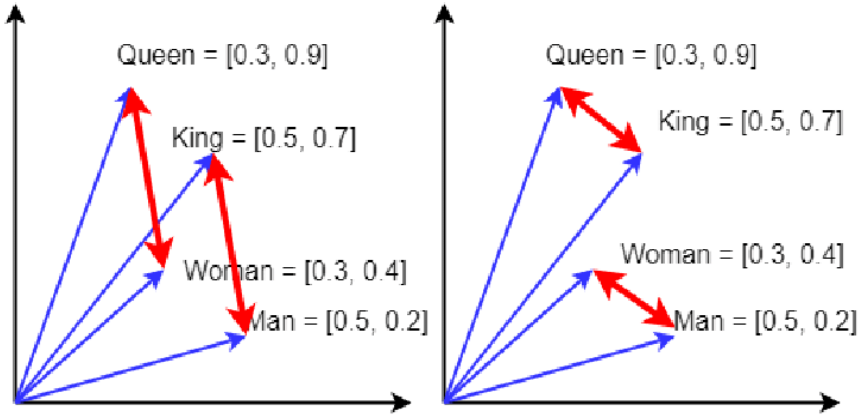
Dekodierung
- Auswahl eines Tokens
- Wahrscheinlichkeitsberechnung

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Funktionsweise LLM

Zuordnung Wort – Vektor
Wichtig Beinhaltung semantischer Strukturen



Quelle: https://pantelis.github.io/cs634/docs/common/lectures/nlp/word2vec

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Funktionsweise LLM

Zerlegung des Textes in Bausteine (Tokens)
- Wörter oder Silben

Einbettung
- Zuordnundg von Wort zu Vektor
- Neuronalen Netzen (Transformator Modell)

Vorhersage des nächsten Tokens
- Neuronale Netze
- Erkennung von Eingaben
- Berechnung von Wahrscheinlichkeiten
- Vorraussetzung Training

Dekodierung
- Auswahl eines Tokens
- Wahrscheinlichkeitsberechnung

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Problematik

Halluzination
Realistischwirkende Information
Abweichung von Input

Fehlende Übereinstimmung (faithfulness)
Mangelnde Richtigkeit (factualness)

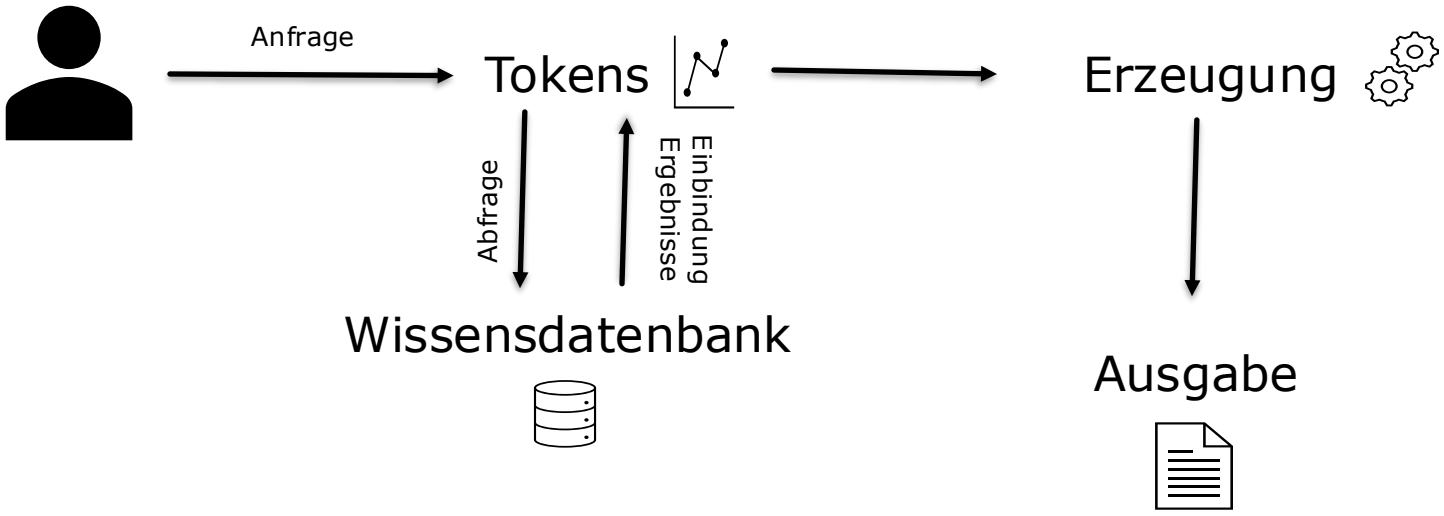
Erkennen:
Unsicherheit messen
Mit Faktendatenbanken vergleichen

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Retrieval Augmented Generation (RAG)

Ergänzung LLMs mit einer Wissensdatenbank



Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Datenquelle für KI

REPOSITORIEN

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Repositorien - LLMs

Repositorien
- Qualitätsgesicherte Inhalte im Internet
- Dauerhafte Texte
- Ergebnisse zu verbessern und richtige Ergebnisse anzuzeigen

Ziel: **Repositorieninhalte in LLMs zu berücksichtigen**

Zusätzlich:
- Möglichkeit einer Wissensdatenbank
- Inhalte besser erzeugen und kontrollieren
- Halluzinationen vermindern

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Policies

Leitlinien für die Nutzung Inhalte und Metadaten

- **Maschinenlesbar**
- Internationalität

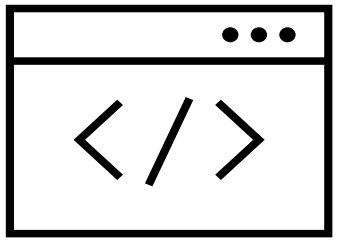
Metadaten frei und nachnutzbar (CC-0)

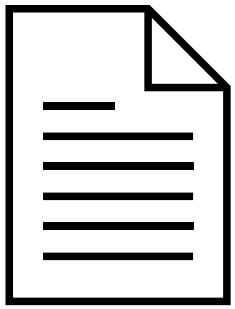
Barcelona Declaration on Open Research Information (https://barcelona-declaration.org/)
(9 Unterzeichner Deutschland, 1 Bayern)

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Präsentation





Landing Page

Alle Information zum Dokumen (Metadaten)

Dokument (pdf)

Volltext des Beitrags

Problem: Verlinkung nur von Landing Page zu Dokument, nicht beidseitig

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Dateiformate- pdf

Standardformat

Gut menschenlesbar
Maschinenlesbarkeit ?

Fehlende explizite Struktur
Komplexe und fehleranfällige Extraktion

Wichtig
- **Text hinterlegt**
- **Metadaten in Eigenschaften ausgefüllt**

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Dateiformate - HTML

Natives Format für Internet

Strukturierte und semantische Information
Semantische Tags, Listen, Überschriften

Trennung von Inhalt und Layout
CSS getrennt von html

Direkte Verarbeitung
HTML-Dokument bekannt aus Pretraining

Effizientere Extraktion relevanter Inhalte
z. B. Entfernung Navigation, Werbung

Bessere Ergebnisse in RAG-Systemen
Aufgrund semantischer Tiefe

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Dateiformate – JATS

Präzise und standardisierte Struktur:
Hohe Interoperabilität und Wiederverwendbarkeit

Metadaten im Format strukturiert

Maschinenlesbarkeit und **Suchfähigkeit**

Erfolgreicher Einsatz mit LLMs

Potenzielle Probleme
- **Komplexität**
- **Tokelimitierungen bei LLMs**

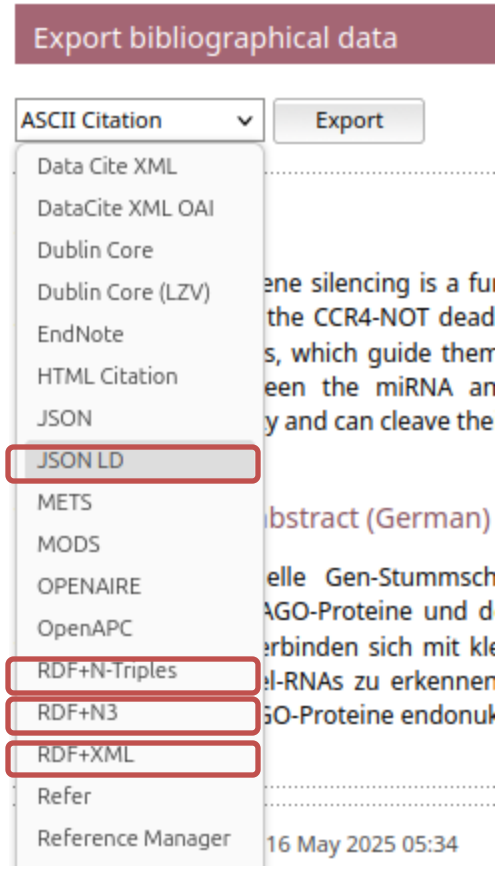
Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Metadaten

Siehe Vortrag
Frech Andreas, KI &
Metadaten - Generierung, Extraktion und Bereitstellung (Andreas Frech, UB der LMU)

Wichtig:
Maschinelle Repsäsentation



Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Suchmaschinenoptimierung

Verändertes Suchverhalten
Fragen an KI-gestützete Systeme – Antworten ohne Klick (Zero-Click ergebnisse)

Neue Anforderungen an Inhalte
Nicht ausschließlichKeywords
vermehrt semantischer Relevanz und Kontext

-> Hochwertiger, tiefgehender, nutzerzentrierter Content wird wichtiger als reine Keyword-Dichte

LLM-SEO als neue Disziplin
klassische Suchmaschinen als auch für KI-Modelle attraktiv

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Knackpunkte KI

PROBLEME

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

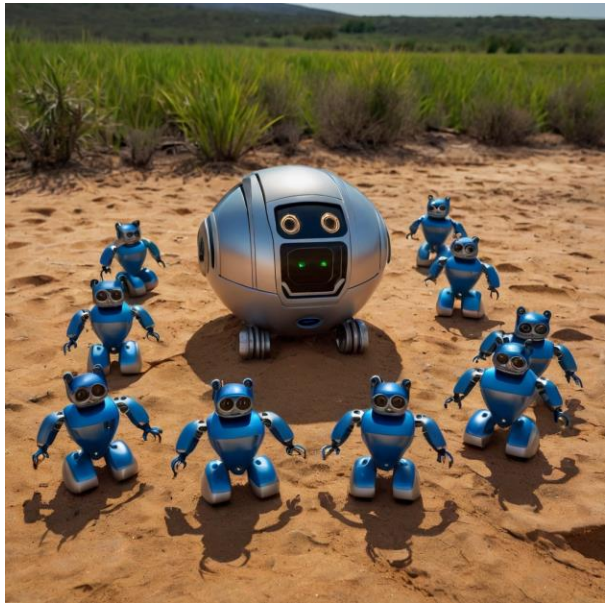
KI-Bots

Training von großen LLMs

Scannen des Internets
analog klassischen Suchmaschinen

Aber:
häufiger
tiefer (mittels KI-Technologie)

Hoher Zugriff auf Repositorien
Sehr viele LLMs
Auch vermehrt in in Suchanfragen



Generated with leonardo.ai

Probleme
Schwer zu filtern
Ignoranz gegenüber herkömmlicher Tools (robots.txt)

[1] Siehe https://coar-repositories.org/news-updates/open-repositories-are-being-profoundly-impacted-by-ai-bots-and-other-crawlers-results-of-a-coar-survey/

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

KI- Bots

Siehe AI bots are destroying Open Access, https://go-to-hellman.blogspot.com/2025/03/ai-bots-are-destroying-open-access.html

Lösungsmöglichkeiten

CAPTCHAS
Sperrung gegenüber Zugriffen spezieller Muster

Problem:

Fehlender Inhalt in LLMs
Bedeutung Repositorien bezüglich LLMs

Open Access
Kein freier Zugriff auf content

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Nur gemeinsam haben wir eine Chance

VERNETZUNG

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

DINI

Arbeitsgruppe
Elektronsiches Publizieren (E-Pub)
https://dini.de/e-pub

Ziel: Unterstützung und konzeptionelle Weiterentwicklung des elektronischen Publizierens an wissenschaftlichen Einrichtungen.

Bsp.: DINI-Zertifiakt
-> Vortrag Daniel Beucke

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Confederation of Open Access Repositories (COAR)

Notwendigkeit eiens starken Zusammenarbeit, um KI und Repositorien bestmöglich zu verknüpfen

Internationales Netzwerk

https://coar-repositories.org

COAR Confederation of Open Access Repositories

13 Mitglieder aus Deutschland
1 Mitglied aus Bayern

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Action Plan for Open Repositories in Europe: IMPACT-REPO

https://coar-repositories.org/wp-content/uploads/2025/03/IMPACT-REPO-2025.pdf

Structure metadata and repositories for AI-driven research, discovery, and automation.

- Adopt machine-readable metadata, linked data practices, and text/data mining capabilities to ensure repository resources remain accessible and valuable in an AI-driven research ecosystem.

Powered by AI-ready infrastructure

Universität Regensburg

Dr. Gernot Deinzer
Leitung Abteilung IT- und Publikationsdienste
Universitätsbibliothek Regensburg

Danke für Ihre Aufmerksamkeit

Fragen

Kontakt:

Dr. Gernot Deinzer
Open Access Beauftragter
Abteilungsleitung IT- und Publikationsdienste
Geschäftsführung UR Data Hub
Fachreferent Mathematik, Physik
93042 Regensburg
email: gernot.deinzer@ur.de
http://www.uni-regensburg.de/bibliothek