

Elisabeth Mödden

Automatische Indexierung mit dem
umfangreichen Vokabular der GND

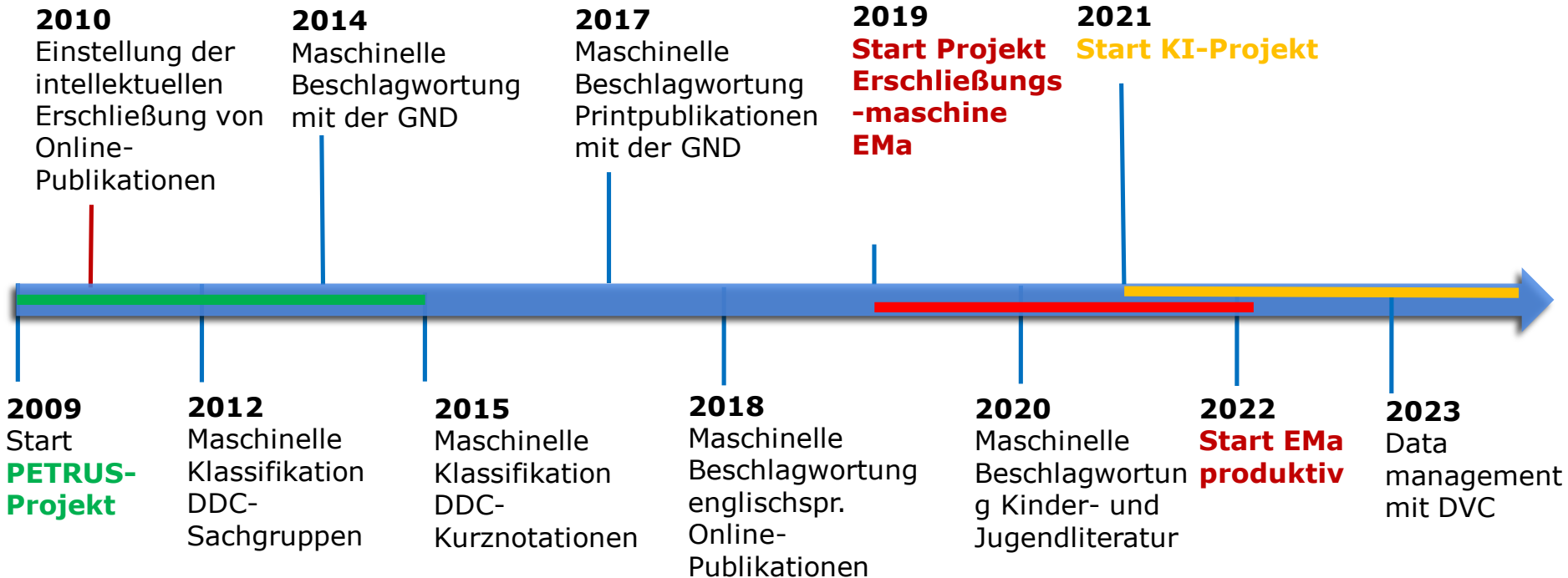
Automatische Schlagwortvergabe als
extremes Multilabel-Lernproblem

Inhaltsverzeichnis

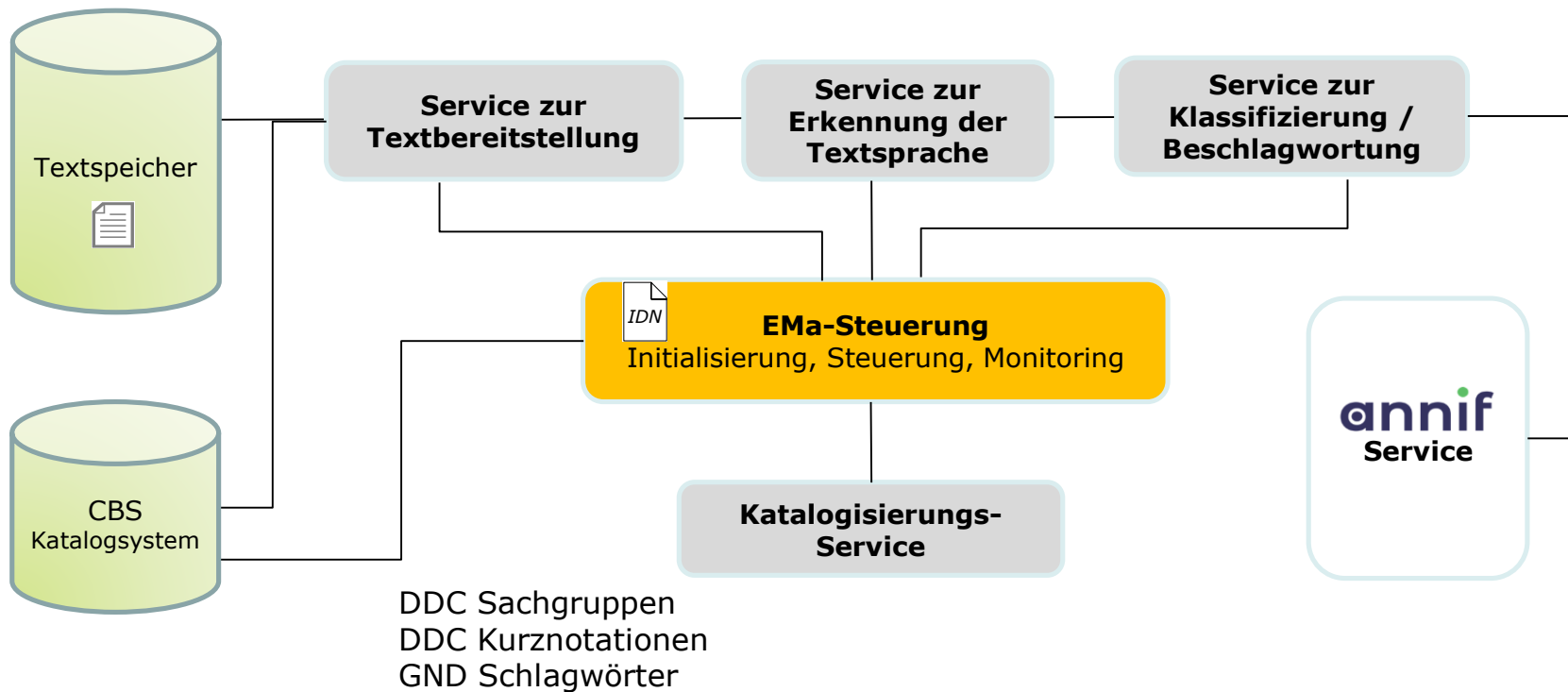
1. Historie und Erschließungsmaschine

2. KI – Projekt

- 1. Projektziel**
- 2. XMLC-Problem – Test XMLC-Verfahren**
- 3. Test Open-Source-LLMs**
- 4. Evaluation und Ausblick**



Modulare Erschließungsmaschine EMa



Automatisierung von Routineaufgaben

Maschinelle Beschlagwortung

Modell	Precision	Recall	F1-Score
mllm FT	0,18989	0,425856	0,24022
omikuji FT	0,392931	0,511763	0,390173
omikuji TITLE	0,442288	0,49413	0,414167
omikuji TOC	0,390357	0,457306	0,357317
Ensemble	0,454969	0,612402	0,473828

Stichprobenumfang:
10.682 Dokumente

Pro Publikation max.
7 Schlagwörter,
Schwellenwert mind.
0.04

Bessere Ergebnisse durch Kombination von KI-Verfahren

KI-Projekt

Automatisches Erschließungssystem

<https://www.dnb.de/ki-projekt:>

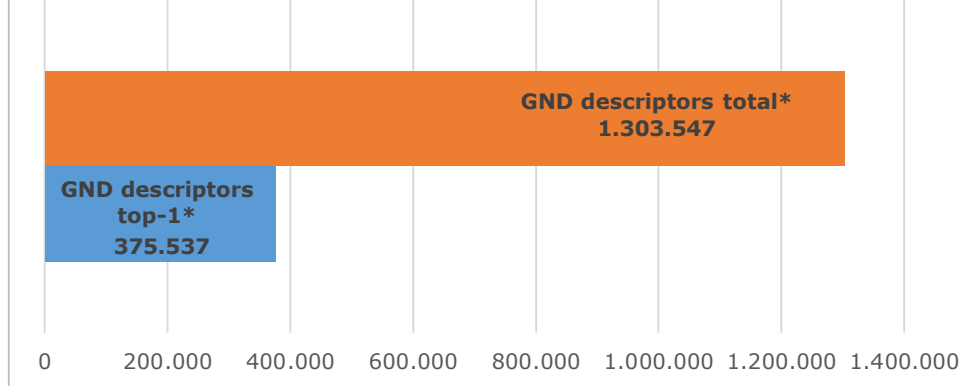
- Förderung durch BKM im Rahmen der nationalen KI-Strategie
- Laufzeit: 4 1/4 Jahre (Oktober 2021 – Dezember 2025)
- Ausstattung
 - Personal (ca. 3 VZÄ)
 - Hardware für Datenlabor (Leistungsstarke GPU etc.)
 - Mittel für Forschungsk Kooperationen

7 Projektziel KI-Projekt

- Qualität der maschinellen Beschlagwortung von deutschsprachigen wissenschaftlichen Netzpublikationen mit der GND durch passende Verfahren / Algorithmen messbar verbessern: F-Score = 0,4 oder besser
- Technologie- und Wissenstransfer in die bibliothekarische Praxis

Das KI-Projekt ist unser Labor für die Erforschung neuer Lösungen – die modular erweiterbare Erschließungsmaschine das System für die Produktion

Trainingsdaten



- Ø 5 intellektuell vergebene GND-Konzepte pro Publikation
- Unausgewogene Verteilung der Trainingsdaten für die GND-Konzepte
 - nur 375.537 GND-Konzepte haben mindestens ein Trainingsobjekt
 - einige GND-Konzepte sind sehr häufig, viele selten vergeben
 - 928.010 GND-Konzepte haben kein Trainingsobjekt
- Mit 1,3 Mio. ist die Anzahl möglicher GND-Konzepte sehr hoch

Maschinelle Beschlagwortung als XMLC-Problem

- Maschinelle Beschlagwortung von Texten mit Konzepten aus der GND lässt sich als sogenanntes Extreme Multi-Label Classification-Problem abstrahieren
 - Eingehende Text-Dokumente werden mit a-priori feststehenden Labels, hier GND-Konzepten, verknüpft
 - Die Menge der zutreffenden Labels pro Dokument ist nicht beschränkt
- Charakteristisch für XMLC-Probleme sind¹
 - Große „Label-Menge“ $\sim 10^5 - 10^6$ Labels
 - Long-Tail-Charakteristik: Ein Großteil der möglichen Labels kommt in Trainingsdaten selten oder nie vor

¹vgl. u.a. Jain et al. 2016 "Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications"

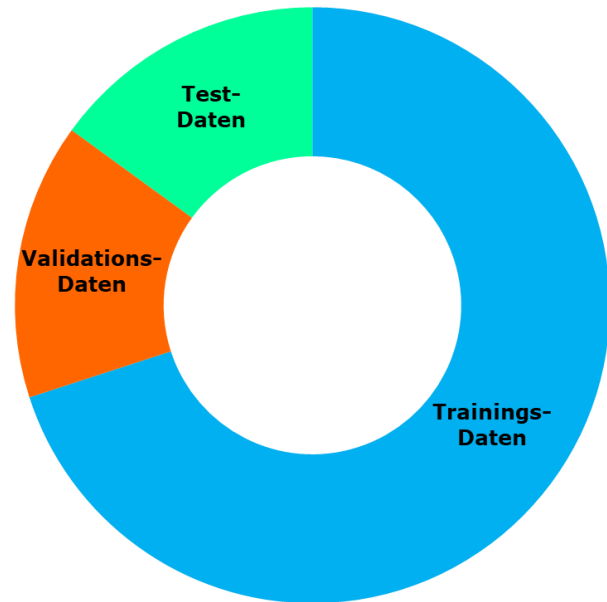
Schwerpunkt des KI-Projektes

- Evaluation verschiedener XMLC-Verfahren
- Evaluation verschiedener Large Language Models (LLMs)

Entwicklung eines Ensembles sich ergänzender Verfahren zur Verbesserung der Beschlagwortung deutschsprachiger wissenschaftlicher Publikationen mit der GND

Evaluation der Verfahren in zwei Tasks

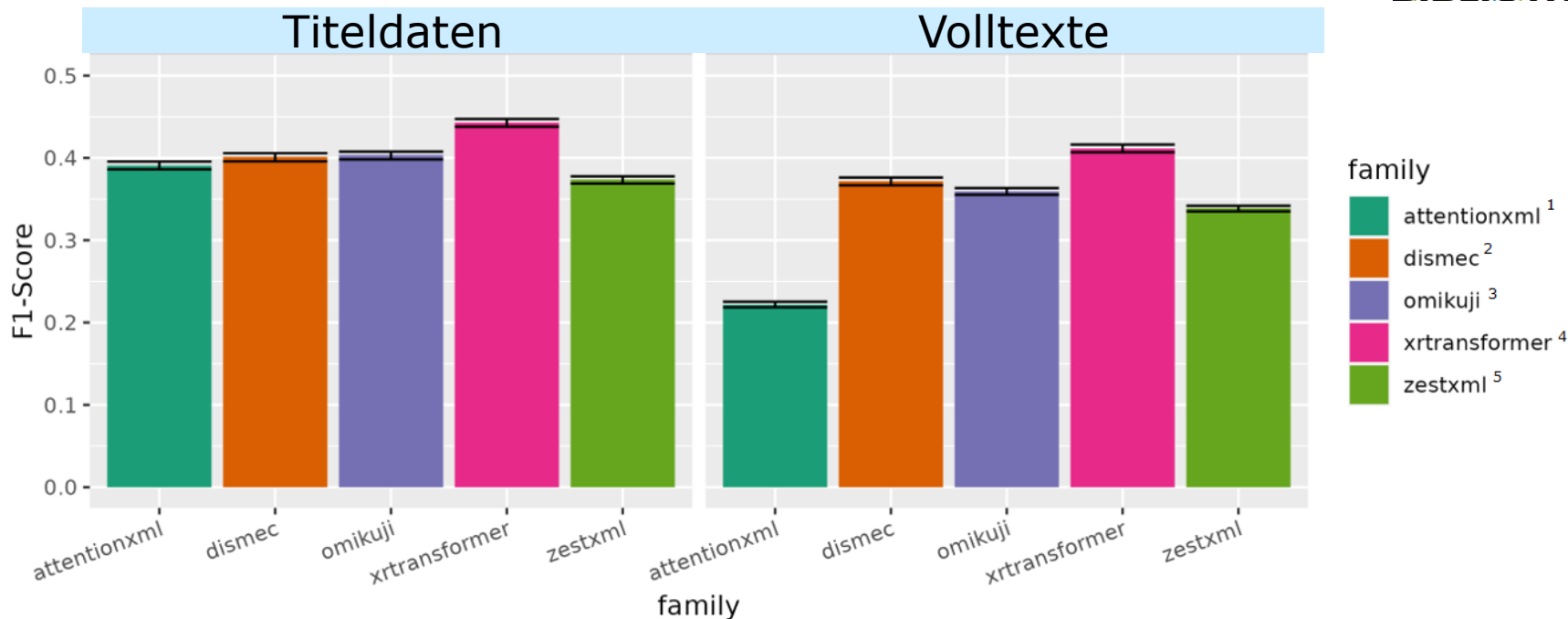
- I. Beschlagwortung von Titeldaten: ca. 1 Mio. Titeldaten mit intellektueller GND-Erschließung
- II. Beschlagwortung von Volltexten: ca. 200.000 Volltexte mit intellektueller GND-Erschließung



Wissenschaftliche Experimente

- Auswahl ML-Verfahren z.B. durch Literaturstudie
- Implementierung des Verfahren im Datenlabor als Test-Pipeline
- Durchführung von Experimenten, Optimierung der Parameter
- Evaluation der Ergebnisse

Benchmarking XMLC-Methoden*



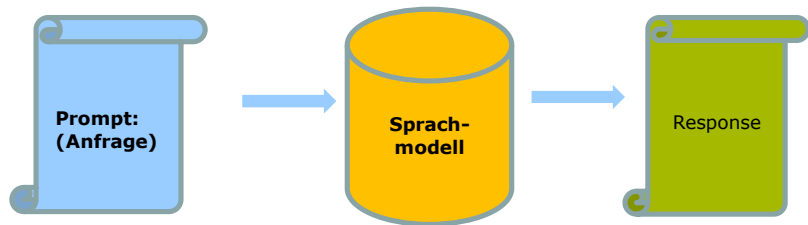
Testmenge: Die Methoden werden auf einem Test-Set von 8.415 deutschen wissenschaftlichen Büchern evaluiert, die gleichmäßig auf 18 verschiedene Fachkategorien verteilt sind

^{1,2,3,4,5}Resources: see last slide

*Intermediate Project-Results as of 09/2024

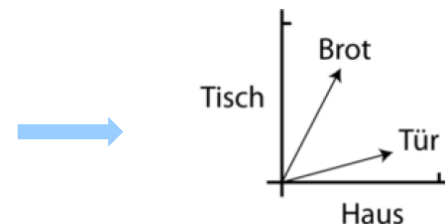
Zweitstufiges Verfahren

1. Vergabe freier Schlagwörter



2. Mapping mit der GND

$$\text{Brot} = \begin{bmatrix} 0.76 \\ 0.92 \\ 0.1 \\ \dots \end{bmatrix}$$



GND-Beschlagwortung mit LLMs

Titel: Hensslers
Schnelle Nummer
Schlagwörter:
Kochen, Alltagsküche
Titel: Kaffeewelten
Schlagwörter:
Kaffeewirtschaft,
Sozioökonomisches
System
###

Titel: Tanzende
Wörter?
Schlagwörter:?

Vervollständige
das Dokument.

Sprachmodell

[Tanz,
Tanztheater,
Choreograf]

1. Vergabe freier
Schlagwörter

```
(alephalpha) florian@valaenki0:~/alephalpha/src$ python AlephAlpha.py
```

	idn	gnd_idn	term	similarity	freeterm
0	102707684X	2072855866	österreicher	0.982579	österreich
1	102707684X	4704476102	Tanz	0.998007	Tanz
2	102707684X	4553493352	Leistung	0.980913	Performance
3	102707684X	1950799549	Tanztheater	0.998635	Tanztheater
0	881227595	6745213837	Kore	0.975992	Korea
1	881227595	7994151874	Aktienrecht	0.998212	Aktienrecht
2	881227595	5226021900	Recht	0.997620	Recht

2. Mapping mit
der GND

16 Test Open-Source-LLMs

Titeldaten

Open-/Closed-Source	Modell	Bester F1
Open-Source	openchat/openchat-3.5-0106	0.315
	meta-llama/Meta-Llama-3-8B-Instruct	0.319
	teknium/OpenHermes-13B	0.332
	LeoLM/leo-hessianai-13b-chat	0.313
	google/gemma-1.1-7b-it	0.282
	Open-Orca/Mistral-7B-OpenOrca	0.317
	DiscoResearch/DiscoLM_German_7b_v1	0.312
	VAGO solutions/Llama-3-SauerkrautLM-8b-Instruct	0.309
	DiscoResearch/Llama3-DiscoLeo-Instruct-8B-v0.1	0.286
	teknium/OpenHermes-2.5-Mistral-7B	0.334
	LeoLM/leo-hessianai-7b-chat	0.261
Closed-Source (Aleph Alpha)	luminous-base (13B)	0.299
	luminous-extended (30B)	0.301
	luminous-supreme (70B)	0.318

Anmerkung:
Durch
Kombination
verschiedener
Modelle und
Prompts können
höhere Werte
erreicht werden.

17 Test Open-Source-LLMs

Volltexte

Open- / Closed-Source	Modell	Bester F1
Open-Source	meta-llama/Meta-Llama-3.1-8B-Instruct	0.315
	meta-llama/Meta-Llama-3.1-70B-Instruct	0.326
	mistralai/Mistral-7B-Instruct-v0.3	0.248
	mistralai/Mistral-Nemo-Instruct-2407	0.194
Closed-Source (Aleph Alpha)	luminous-base (13B)	0.202*

*nur Open-Access-Texte

Anmerkung:
Durch
Kombination
verschiedener
Modelle und
Prompts können
höhere Werte
erreicht werden.

Herausforderungen beim Umgang mit LLMs

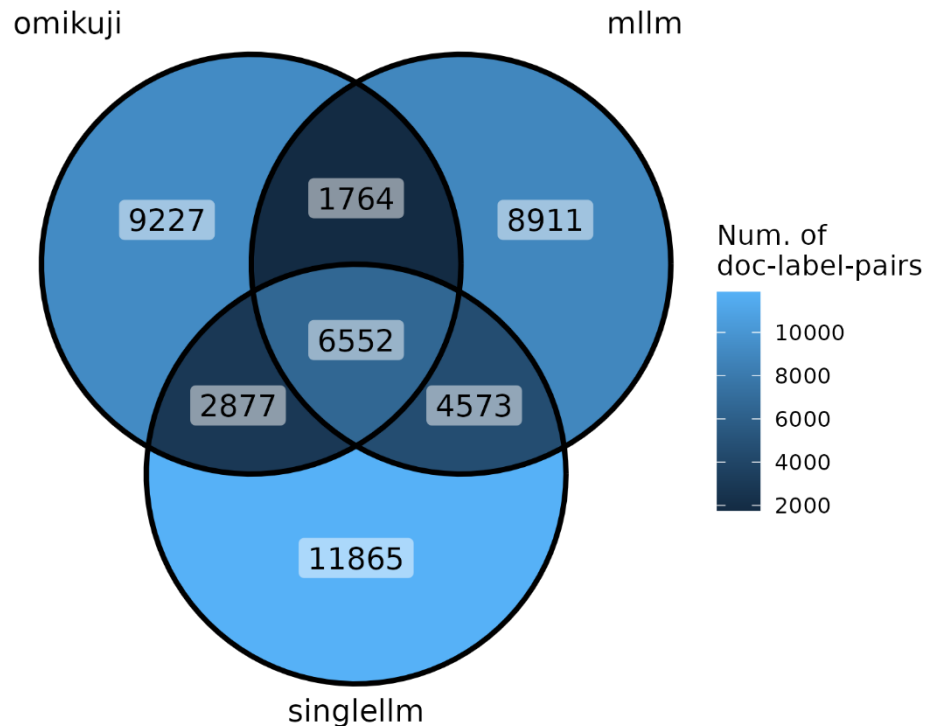
- Verarbeitung langer Texte: Sprachmodelle haben begrenzte Kontextlängen
- Textqualität: die PDF-Extraktion unserer Volltexte ist für LLMs ungeeignet
- Hardware: größere Modelle (über 70B) können nicht von uns lokal betrieben werden

Optimierungsstrategien für LLMs

- **Prompt Engineering**
 - Optimierung der Eingaben durch präzise Anweisungen oder Beispiele
- **Retrieval Augmented Generation (RAG)**
 - Methode: Integration interner oder externer Datenquellen in die Prompts
 - Ziel: Bereitstellung von Kontext und zusätzlichen Informationen
 - Vorteile: Zugriff auf aktuelle, dynamische, spezielle o. interne Informationen, bessere Beantwortung spezifischer Fragen, Reaktion auf aktuelle Ereignisse
- **Fine-Tuning**
 - Nachträgliches Training des LLM mit ausgewählten Parametern und spezialisierten Datensätzen

Vergleich verschiedener Verfahren*

Venn-Diagramm für die Label-Vorschläge von Omikuji, MLLM und dem besten Open-Source-LLM-Experiment singlellm für Titeldaten



Vergleich verschiedener Verfahren*



Eigenschaften und Herausforderungen

- Es gibt nicht DIE EINE universelle Lösung für alle Aufgaben ... und es gibt sie nicht umsonst („no free lunch“)
- KI-Lösungen sind meistens eine Kombination aus Methoden
- Mit zunehmendem Anspruch wächst auch der Bedarf an Ressourcen

Trainings-
daten

Testdaten

Methoden-
auswahl und
Modellierung

Rechen-
leistungen

Ausblick

- Systematische Evaluierung mit deutschsprachigen wissenschaftlichen Netzpublikationen
- Auswertung des Anteils einzelner Methoden für ein leistungsstarkes heterogenes Ensemble von Methoden
- (Open Source-)Services entwickeln, bereitstellen und nutzen

Vielen Dank!

Fragen?

