

Elisabeth Mödden

Automatische Schlagwortvergabe in der DNB

Inhaltsverzeichnis

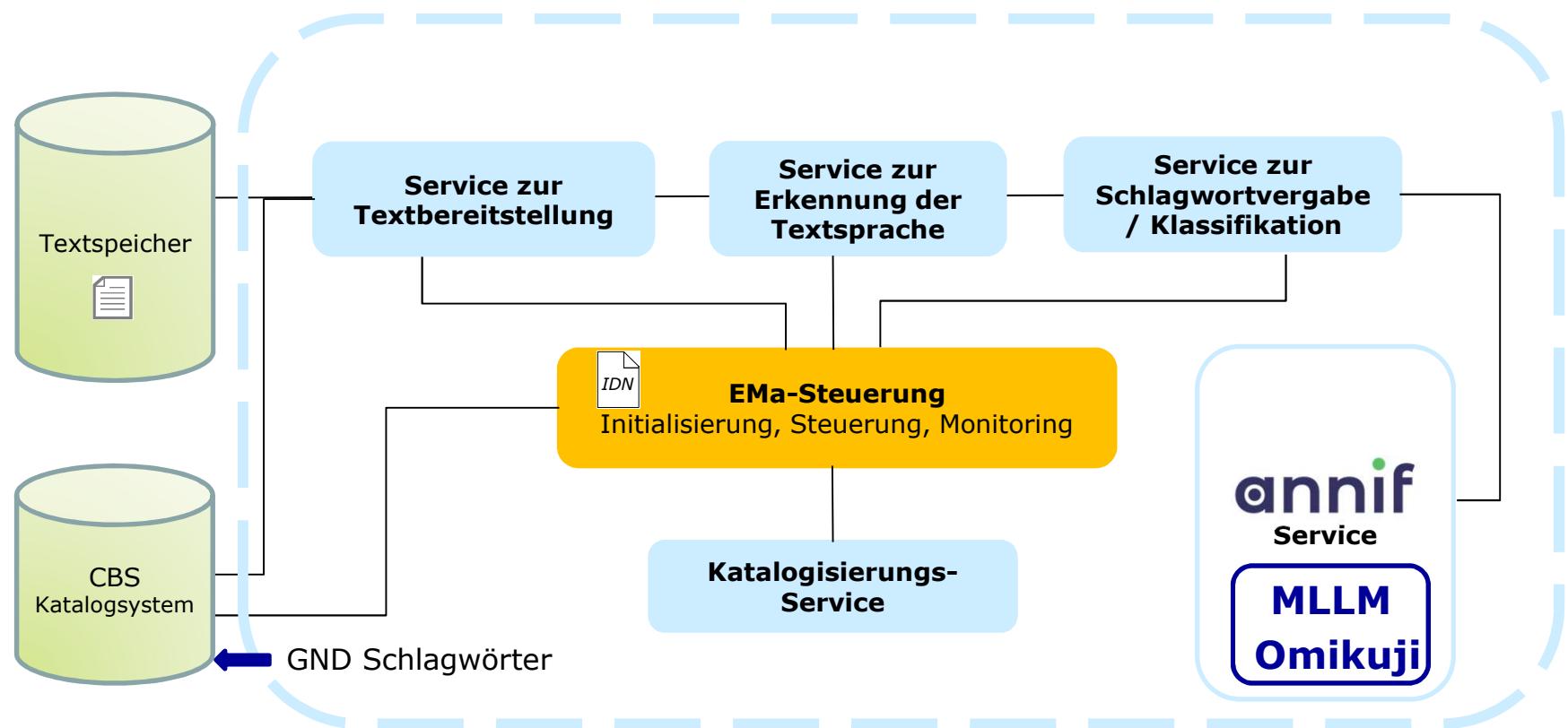
- 1. Kontext: Automatische Schlagwortvergabe – im Regelbetrieb**
- 2. Ergebnisse aus dem KI-Projekt**
- 3. Fazit und Ausblick**

Automatische Schlagwortvergabe im Regelbetrieb

- Seit 2014 produktiv, zunächst als rein lexikalisches Verfahren mit der Averbis-Software
- Seit 2022 komplette Neuentwicklung, Erschließungsmaschine EMa mit modulare Architektur
- Kernkomponente: Annif¹, ein Open-Source-Toolkit der Finnischen Nationalbibliothek mit diversen Erschließungsverfahren und der Möglichkeit, diese zu Ensembles zu kombinieren (z. B. lexikalische und überwachte Machine-Learning-Verfahren)
- Durchsatz: ca. 170.000 Publikationen pro Jahr
- Automatische Schlagwortvergabe auf Basis der GND mit > 1,4 Mio. potenziellen Schlagwörtern

¹Suominen, O. Etal. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing. <https://doi.org/10.4403/jlis.it-12740>

Modulare Erschließungsmaschine EMa



Lexikalische Verfahren

Abgleich zwischen relevanten Termen im Text und Termen im kontrollierten Vokabular

Als **Wasserbau** werden Maßnahmen, technische Eingriffe und Bauten im Bereich des **Grundwassers**, der **Oberflächengewässer** und der **Meeresküsten** bezeichnet. Heute weniger gebräuchlich ist die Bezeichnung **Hydrotechnik** für dieses Fachgebiet..^{*} Text aus Wikipedia

Wasserbau

Sachbegriff

GND-Nummer : 4064700-6

Untergliederung : [Allgemeinbegriff \[saz\]](#)

Grundwasser

Systematik : 31.3b [Bautechnik](#)

Sachbegriff

DDC-Notation : 627

GND-Nummer : 4022369-3

Quelle :

M

Untergliederung : [Allgemeinbegriff \[saz\]](#)

Varianten : Hydrotechnik

Systematik : 19.3 [Hydrologie, Meereskunde](#)

Wasserbauwerk

DDC-Notation : 2-1698

Thematischer Bezug : [Unterwasserbau](#) (verwandt)

551.49

553.79

553.79

628.114

628.114

333.9104

Küste

Sachbegriff

GND-Nummer : 4033469-7

Untergliederung : [Allgemeinbegriff \[saz\]](#)

lle : M

Systematik : 19.1b [Physische Geografie](#)

rbegriffe : [Unterirdisches Wasser](#) (Oberbegriff generisch)

DDC-Notation : 2-146

551.457

577.699

Quelle : M

Verwendungshinweis : Mit einzelnen Gewässern wird, soweit s
gebildet

Varianten : Meeresküste

Meerufer

Meerstrand

MLLM

(Maui-like Lexical Matching)

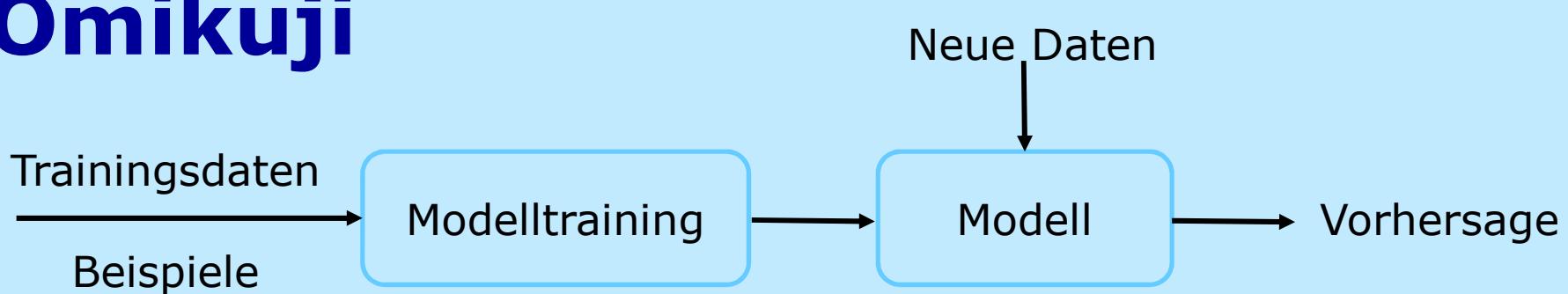
Maui-like Lexical Matching

<https://github.com/NatLibFi/Annif/wiki/Backend%3A-MLLM>

Überwachtes ML – „klassische“ KI

- Überwachtes maschinelles Lernen nutzt Beispieldokumente, die intellektuell mit GND-Schlagwörtern erschlossen wurden.
- Aus diesen Beispielen lernt das Modell, welche sprachlichen Muster zu welchen normierten GND-Schlagwörtern gehören, und kann diese Schlagwörter für neue Dokumente vorschlagen.

Omikuji



Omikuji Rust-Bibliothek: <https://github.com/tomtung/omikuji>

Aktuelle Ergebnisse Automatische Schlagwortvergabe im Regelbetrieb

Modell	Precision	Recall	F1-Score	
mllm	0,216	0,406	0,257	Stichprobenumfang: 10.682 Dokumente
omikuji Volltexte	0,443	0,498	0,413	Pro Publikation max. 6 Schlagwörter
omikuji Titel	0,492	0,549	0,471	
omikuji Inhaltsverz.	0,424	0,472	0,386	Schwellenwert mind. 0.05
ensemble	0,503	0,624	0,510	

Bessere Ergebnisse durch Kombination von KI-Verfahren

Automatische Schlagwortvergabe im Regelbetrieb

- Seit 2014 produktiv, zunächst als rein lexikalisches Verfahren mit der Averbis-Software
- Seit 2022 komplette Neuentwicklung, Erschließungsmaschine EMa mit modulare Architektur
- Kernkomponente: Annif¹, ein Open-Source-Toolkit der Finnischen Nationalbibliothek mit diversen Erschließungsverfahren und der Möglichkeit, diese zu Ensembles zu kombinieren (z. B. lexikalische und überwachte Machine-Learning-Verfahren)
- Durchsatz: ca. 170.000 Publikationen pro Jahr
- Automatische Schlagwortvergabe auf Basis der GND mit > 1,4 Mio. potenziellen Schlagwörtern

¹Suominen, O. Etal. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing. <https://doi.org/10.4403/jlis.it-12740>

KI-Projekt (2021 – 2025)



Der Beauftragte der Bundesregierung
für Kultur und Medien

Förderung durch BKM im Rahmen
der nationalen KI-Strategie



- Qualität der automatischen Schlagwortvergabe von deutschsprachigen wissenschaftlichen Netzpublikationen mit der GND durch passende Verfahren / Algorithmen messbar verbessern
- 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

Woran haben wir gearbeitet?

Infrastruktur

Datenmanagement

HPC-Nutzung

Forschungs-
Kooperation

Testen von
Verfahren

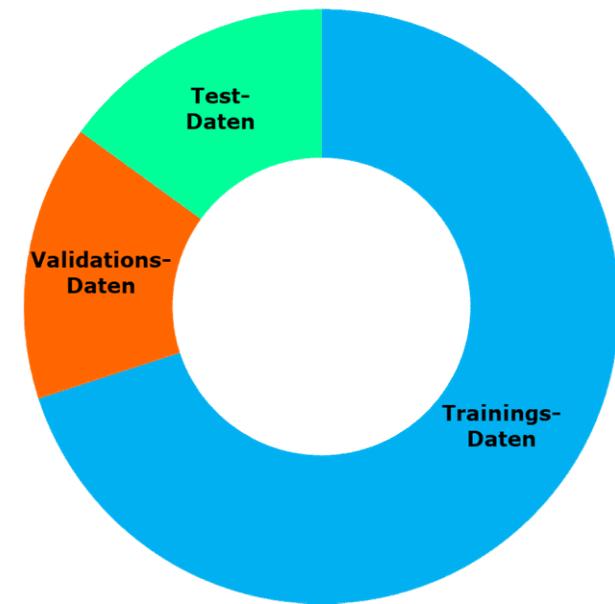
Evaluation

Literaturrecherche

Kommunikation und
Vorträge

Evaluation von Verfahren in zwei Tasks

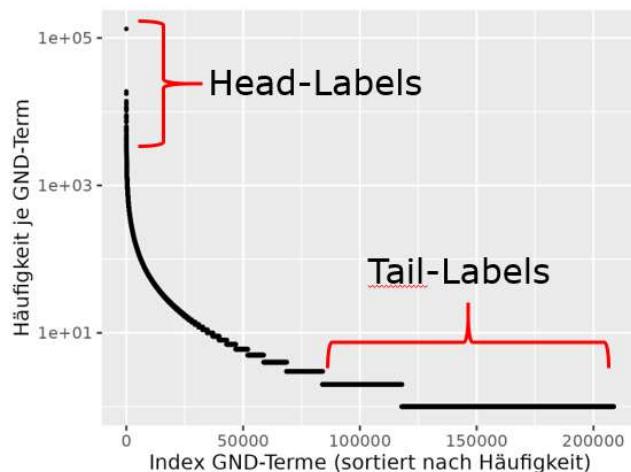
- I. Schlagwortvergabe auf Basis von Titeldaten: ca. 1 Mio. Titeldaten mit intellektueller GND-Erschließung
- II. Schlagwortvergabe auf Basis von Volltexten: ca. 200.000 Volltexte mit intellektueller GND-Erschließung



Herausforderung GND

Automatische Schlagwortvergabe von Texten mit Konzepten aus der GND lässt sich als sogenanntes Extreme Multi-Label Classification-Problem abstrahieren

Long-Tail Charakteristik



*GND gesamt**
1.428.353 Schlagwörter

*mind. einmal oder häufiger im
Bestand der DNB verknüpft***
208.688 Schlagwörter

nicht im Bestand der DNB verknüpft
1.219.665 Schlagwörter

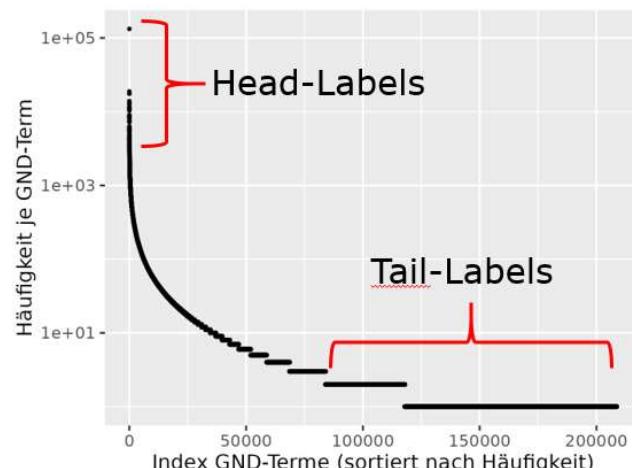
*Anm.: Nur Katalogisierungslevel 1 oder z und aus dem Teilbestand s

**Anm.: Bezieht sich nur auf die Verknüpfung mit *deutschsprachigen* Publikationen in unserem Trainings-Datensatz

Herausforderung GND

Automatische Schlagwortvergabe von Texten mit Konzepten aus der GND lässt sich als sogenanntes Extreme Multi-Label Classification-Problem abstrahieren

Long-Tail Charakteristik



Suche nach geeigneten XMLC-Verfahren

*GND gesamt**
1.428.353 Schlagwörter

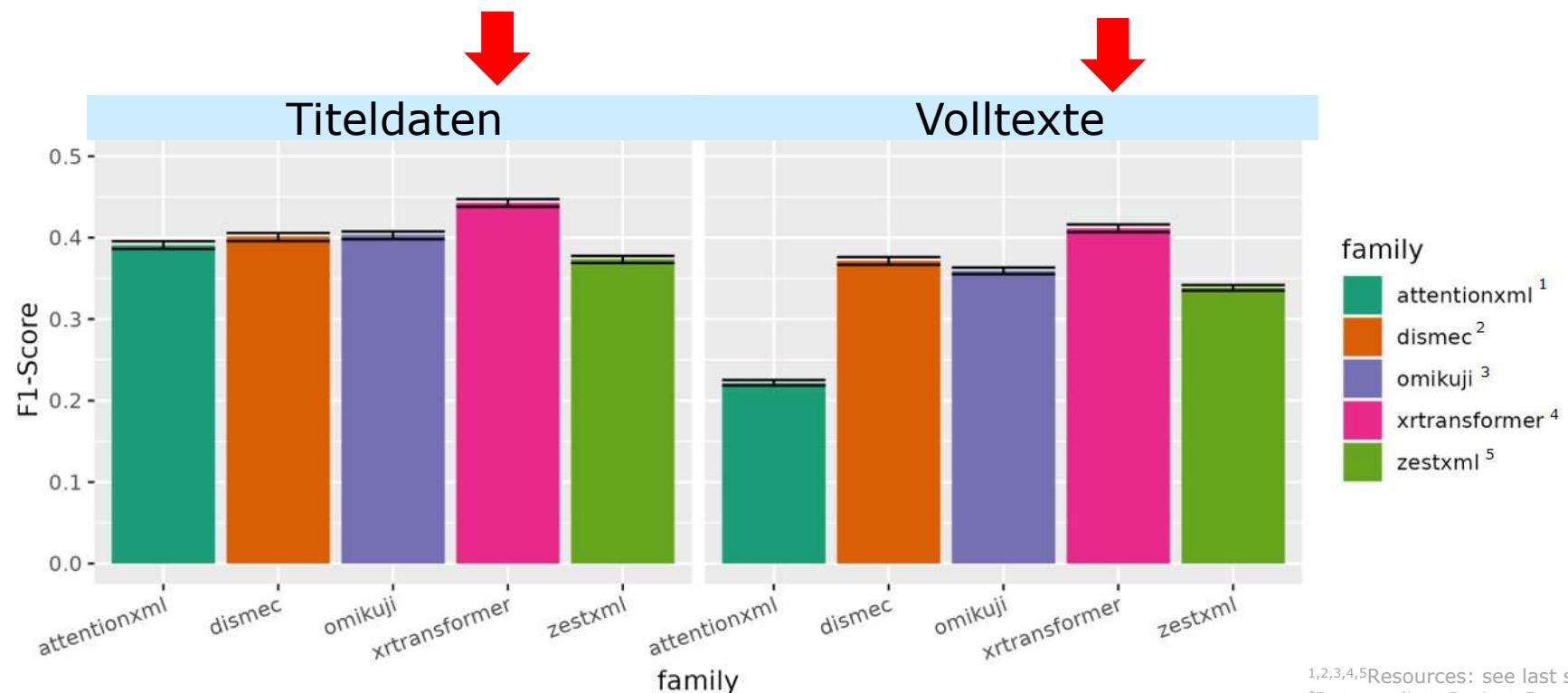
*mind. einmal oder häufiger im Bestand der DNB verknüpft***
208.688 Schlagwörter

nicht im Bestand der DNB verknüpft
1.219.665 Schlagwörter

*Anm.: Nur Katalogisierungslevel 1 oder z und aus dem Teilbestand s

**Anm.: Bezieht sich nur auf die Verknüpfung mit *deutschsprachigen* Publikationen in unserem Trainings-Datensatz

Benchmarking XMLC-Verfahren*



^{1,2,3,4,5}Resources: see last slide
*Intermediate Project-Results as of 09/2024

GND-Schlagwortvergabe mit LLMs

- Problem: LLMs kennen die GND nicht
- Grundprinzip: Few-Shot-Prompting
 - Generiere freie Schlagwortvorschläge mit einem LLM
 - Erzeuge ein Mapping auf die GND durch ein (kleineres) Encoder-Modell



Die Idee des Few-Shot-Prompting

Das Modell mit wenigen Beispiel-Eingaben und -Ausgaben auf eine Aufgabe „einstimmen“, damit es das gewünschte Muster lernt.

Prompt

Extrahiere Schlagwörter aus dem Text.

Text: *Beispieltext*

Schlagwörter: *Beispielschlagwörter*

###

...

###

Text: **Testtext**

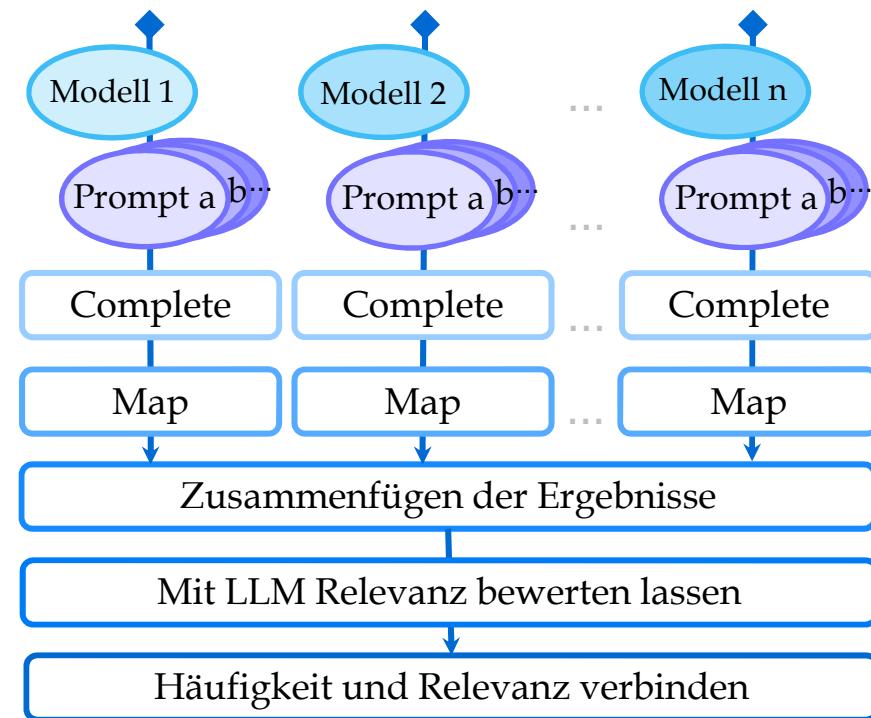
Schlagwörter: **Generierte Schlagwörter**

Complete

LLM

LLM-Ensemble: Warum mehrere Modelle?

- Mehrere Modelle ergänzen sich z. B. Prototyp aus dem KI-Projekt mit den LLMs:
 - Mistral-7B,
 - Mixtral-8x7B,
 - Llama-3.1-70B
- Dadurch entstehen mehr korrekte und präzisere Schlagwortvorschläge



¹Kluge, L., & Kähler, M. (2024).
Few-Shot Prompting for Subject Indexing of German Medical Book Titles.
<https://aclanthology.org/2024.konvens-main.16>

²Prototyp aus dem DNB-KI-Projekt:
https://github.com/deutsche-nationalbibliothek/semeval25_llmensemble

Verfeinerungen des Few-Shot-Prompting Ansatzes

- **Ensemblierung:** Anwendung von Complete + Map durch mehrere Modelle und Prompts (→ mehr Breite in den Vorschlägen)
- **Reranking:** Überprüfen des Mappings durch LLM-basierte Relevanzbewertung
- **Knowledge-Injection:** Suche von passenden Few-Shot-Beispielen durch RAG-Ansatz

LLM-Ensemble

Kluge, L. and Kähler, M. (2025) “DNB-AI-Project at SemEval-2025 Task 5: An LLM-Ensemble Approach for Automated Subject Indexing,”
<https://aclanthology.org/2025.semeval-1.148/>.

KIFSPrompt

Kähler, M., Kluge, L. and Konermann, K. (2025) “DNB-AI-Project at the GermEval-2025 LLMs4Subjects Task: KIFSPrompt - Knowledge-Injected Few-Shot Prompting,”
<https://aclanthology.org/2025.konvens-2.42/>.

Word Embedding - Weiterentwicklung der „klassischen“ Ansätze durch Transformer

Embedding-basiertes Matching (EBM)¹:

Das String-basierte Matching-Verfahren im lexikalischen Ansatz wird durch einen Abgleich von transformer-basierten Text-Embeddings ersetzt oder ergänzt.



XR-Transformer²:

Die Bag-of-Words basierte TFIDF-Vektorisierung der Texte wird durch Embeddings eines feinabgestimmten Transformermodell es ersetzt oder ergänzt.

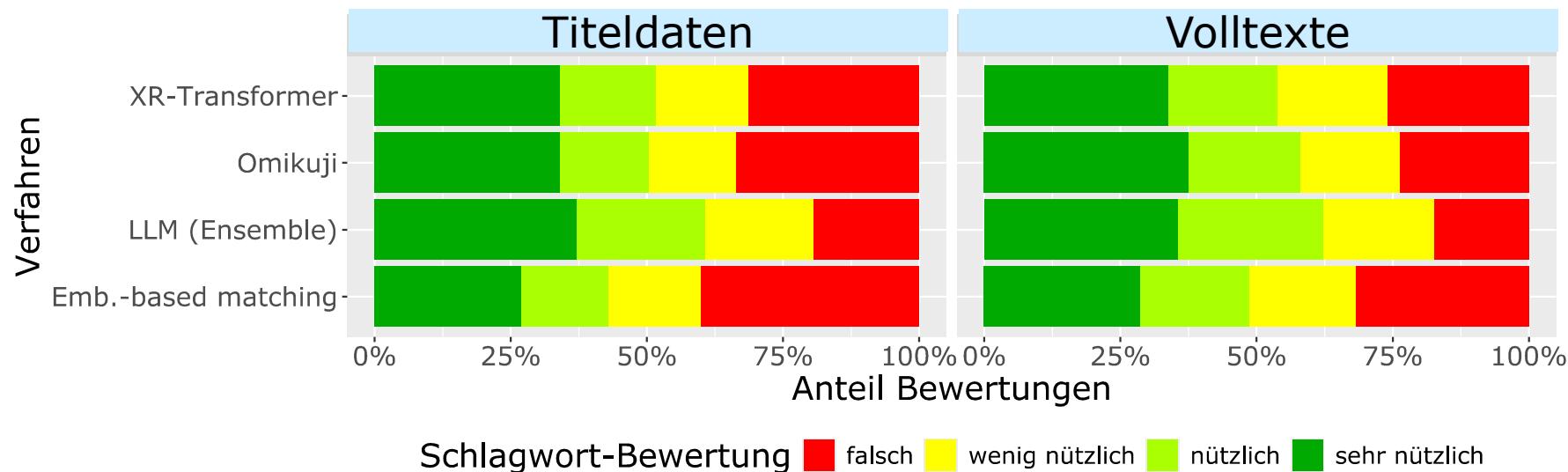
¹Prototyp aus dem DNB-KI-Projekt:

<https://github.com/deutsche-nationalbibliothek/ebm4subjects>

²Zhang et al. (2021). *Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification*.

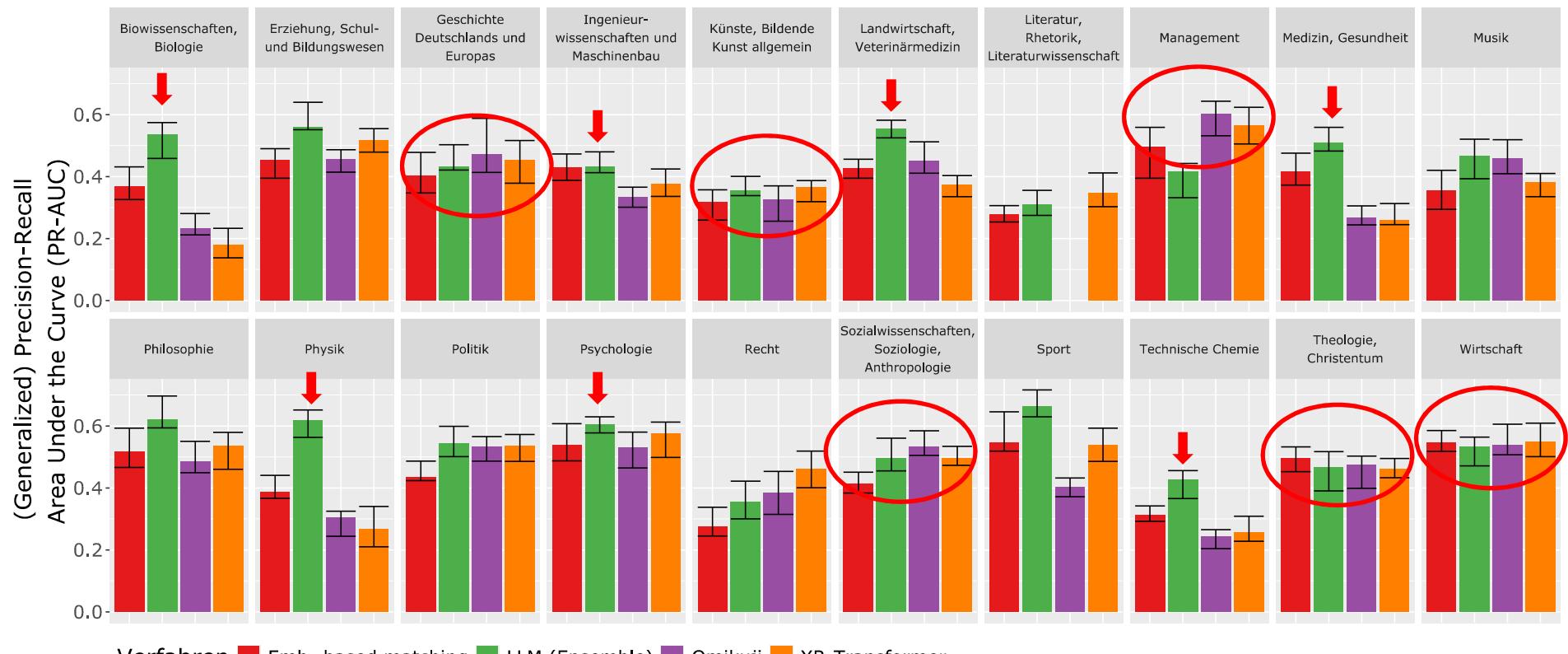
<https://arxiv.org/abs/2110.00685v2>

Vergleich qualitative Bewertung

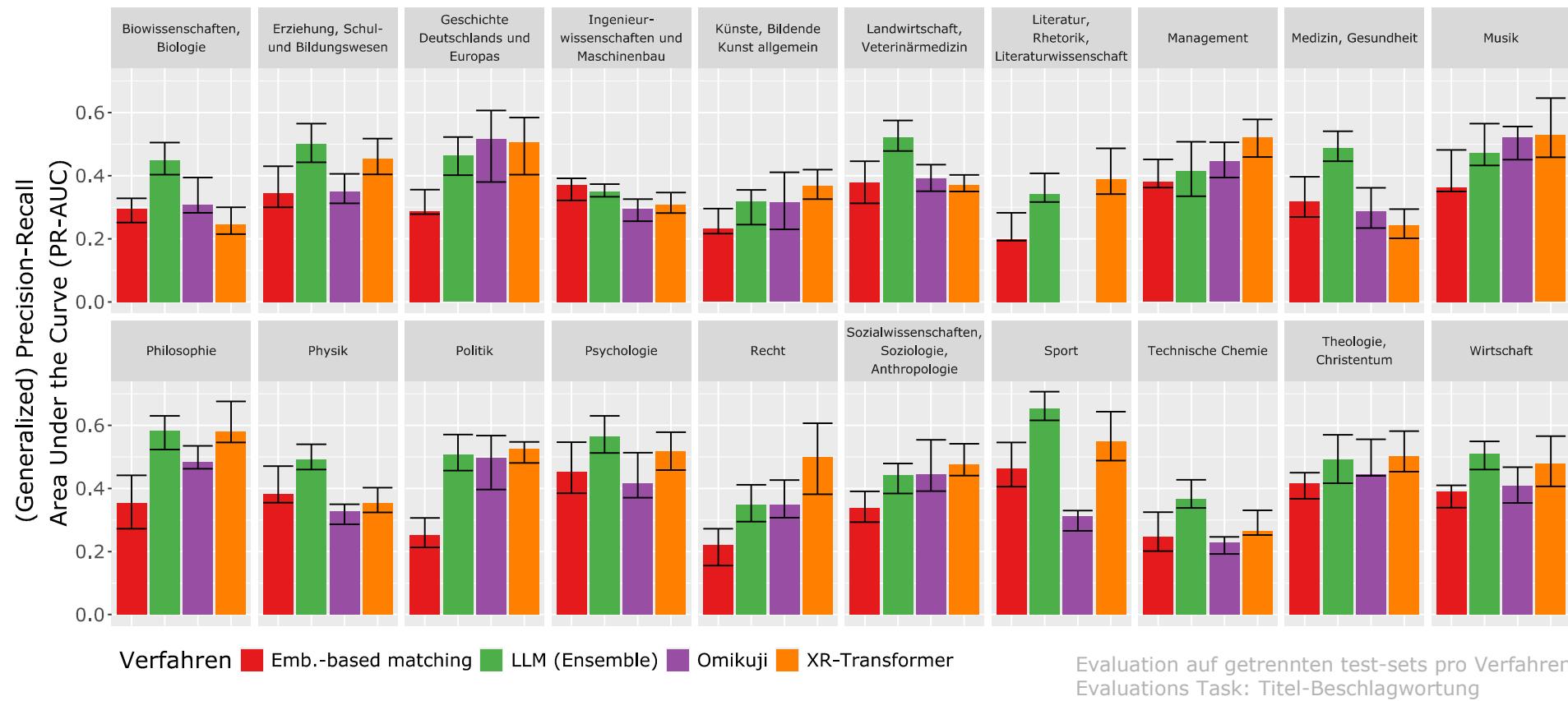


Jedes Verfahren wurde auf einem Testset von ca. 1.100 deutschsprachigen wissenschaftlichen Texten intellektuell durch Fachreferent*innen evaluiert. Bewertet wurden die Top-5-Schlagwortvorschläge pro Dokument und Verfahren mit einer 4-stufigen Bewertungsskala.

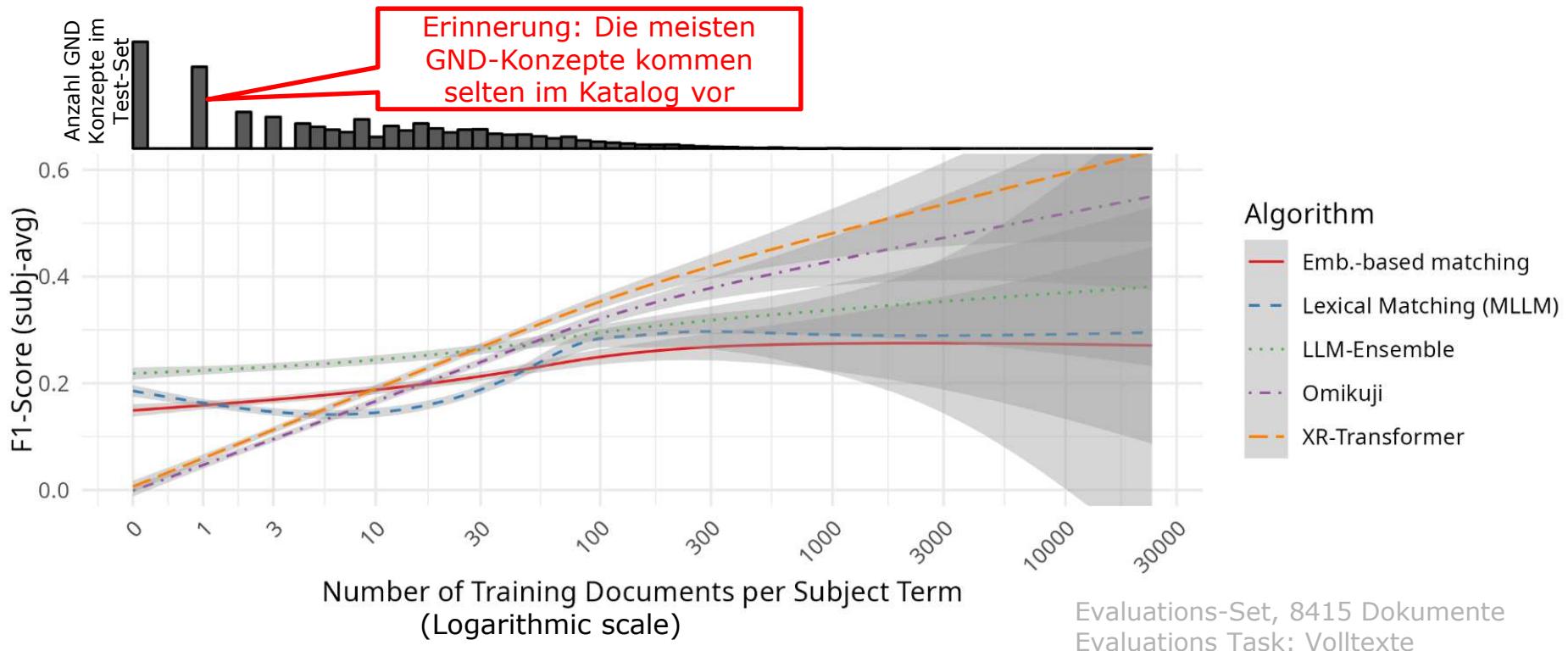
Vergleich Ergebnisse nach Sachgruppe - qualitative Bewertung



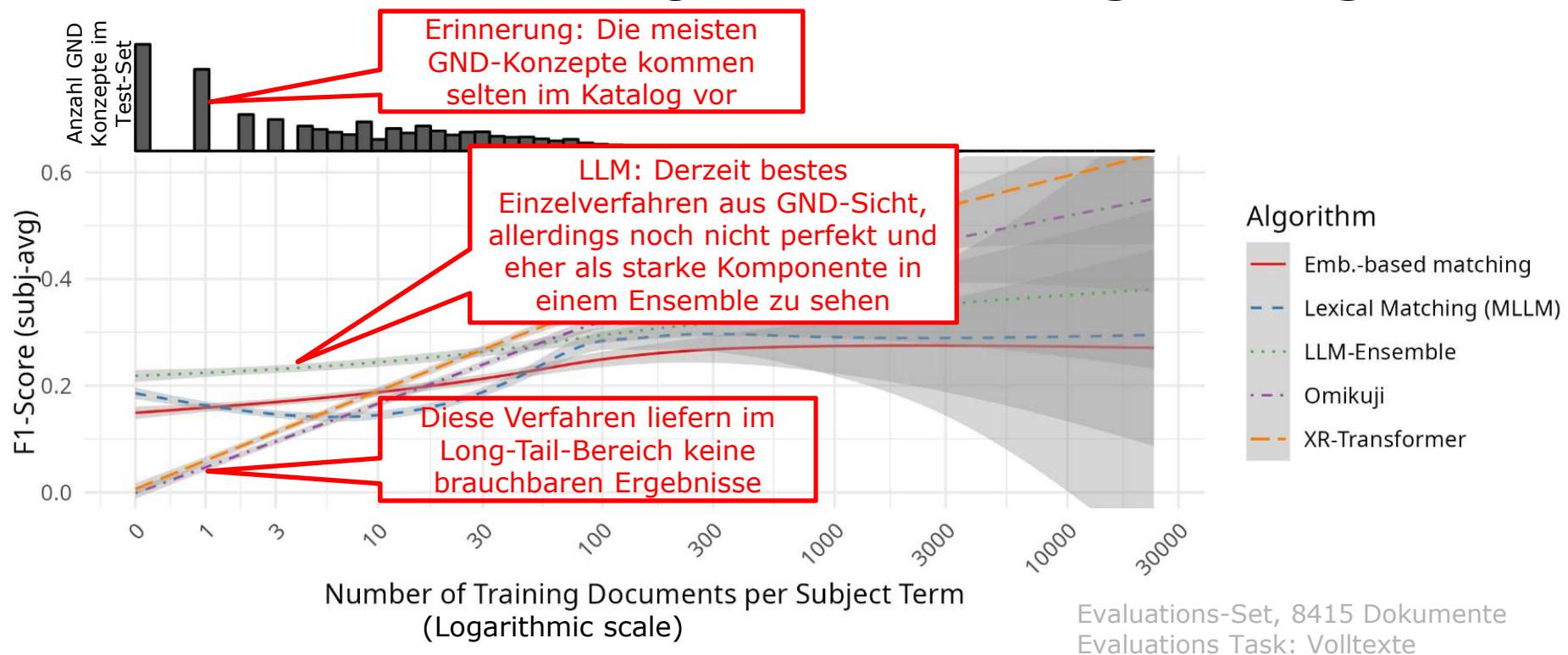
Vergleich Ergebnisse nach Sachgruppe - qualitative Bewertung



Die “GND-Sicht” auf die Qualität der Schlagwortvergabe



Die "GND-Sicht" auf die Qualität der Schlagwortvergabe



Technische Vergleichskriterien

	Trainings-aufwand mit eigenen Daten	Inferenz-kosten (im Live-Einsatz)	Vorhersage seltener und bislang nicht verwendeter Schlagwörter	Verarbeitung langer Texte	Disambiguierung	Generalisierung
Omikuji	-	++	--	++	o (für häufige Konzepte erlernbar)	o (für häufige Konzepte erlernbar)
Lexical Matching (MLLM)	+	++	o	++	-	--
Embedding bas. Match. (EBM)	o 	o 	o	++	o	--
XR-Transformer	--	o 	o 	--	o (Nur TFIDF-Features)	o (für häufige Konzepte erlernbar)
LLM-Ensemble	++	--		++	++	+



GPU-Einsatz vorteilhaft

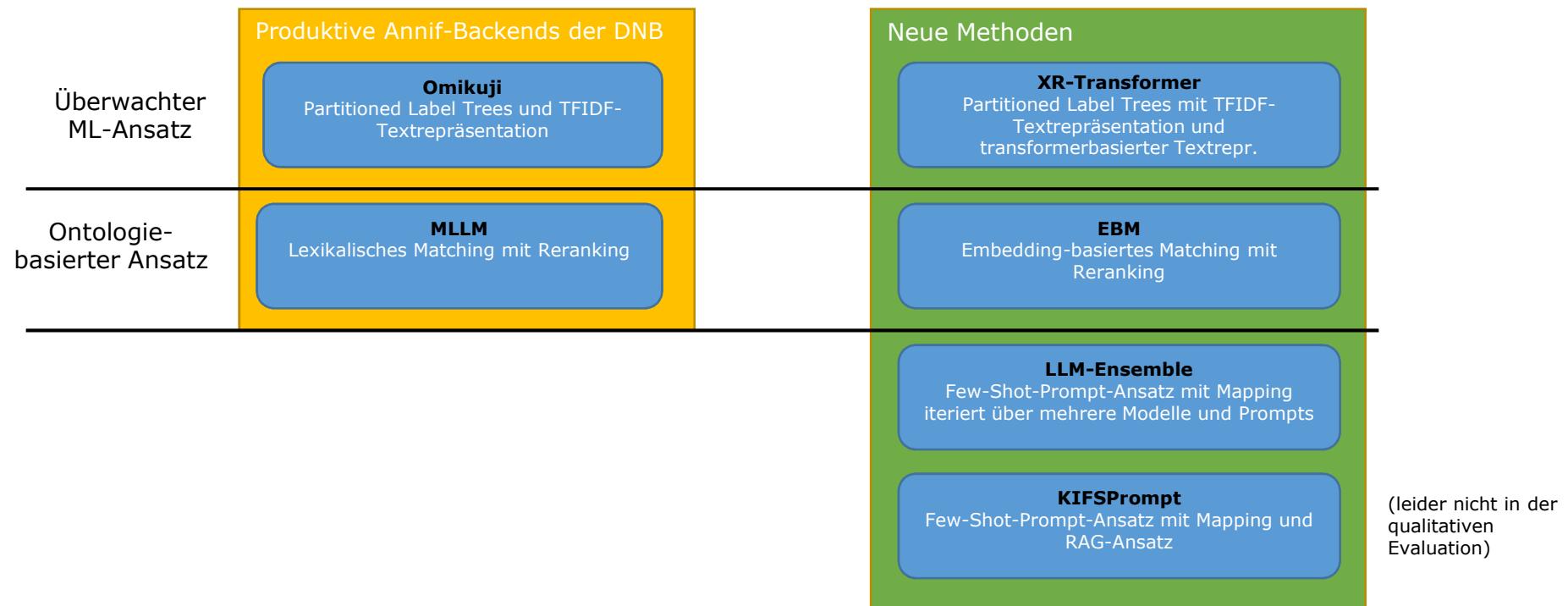


GPU-Einsatz zwingend
Genau Angaben hängen vom Anwendungsfall ab

*größerer RAM bzw. GPU-RAM nötig

GPU-image: Flaticon.com

Zusammenfassung der Methoden



Fazit

- Transformer-Modelle können existierende Erschließungsverfahren verbessern.
- Der Einsatz von LLMs bringt einen qualitativen Sprung.

ABER

- Automatische Schlagwortvergabe bleibt ein **komplexes Problem**. Es ist wichtig, dass sich auch die DNB weiter aktiv in die Forschung einbringt
- Wichtig ist, dass die intellektuelle Erschließung im Zentrum bleibt: Sie liefert das fachliche Know-how, bewertet die Ergebnisse, kuratiert die Trainingsdaten und entscheidet in schwierigen Fällen.
- Kein Verfahren ist allein die beste Lösung. Stärken und Schwächen verschiedener Ansätze lassen sich in einem **Ensemble** geschickt kombinieren.
- Kosten-Nutzen Trade-Off kann für „schlankere“ Verfahren (MLLM, EBM) günstiger sein als für LLMs.

Ausblick

Technische Arbeiten:

- Embedding-basiertes Matching (EBM) entwickeln und in Annif integrieren
- XR-Transformer in Annif integrieren
- LLM-Verfahren für die DNB optimieren
- Neues Ensemble aus diesen Verfahren für unterschiedliche Anwendungsfälle entwickeln, z. B. für englischsprachige Publikationen
- Übertragung der Erkenntnisse auf Prozesse der Automatisierung von Formalerschließung

Beschaffung notwendiger Hardware für KI:

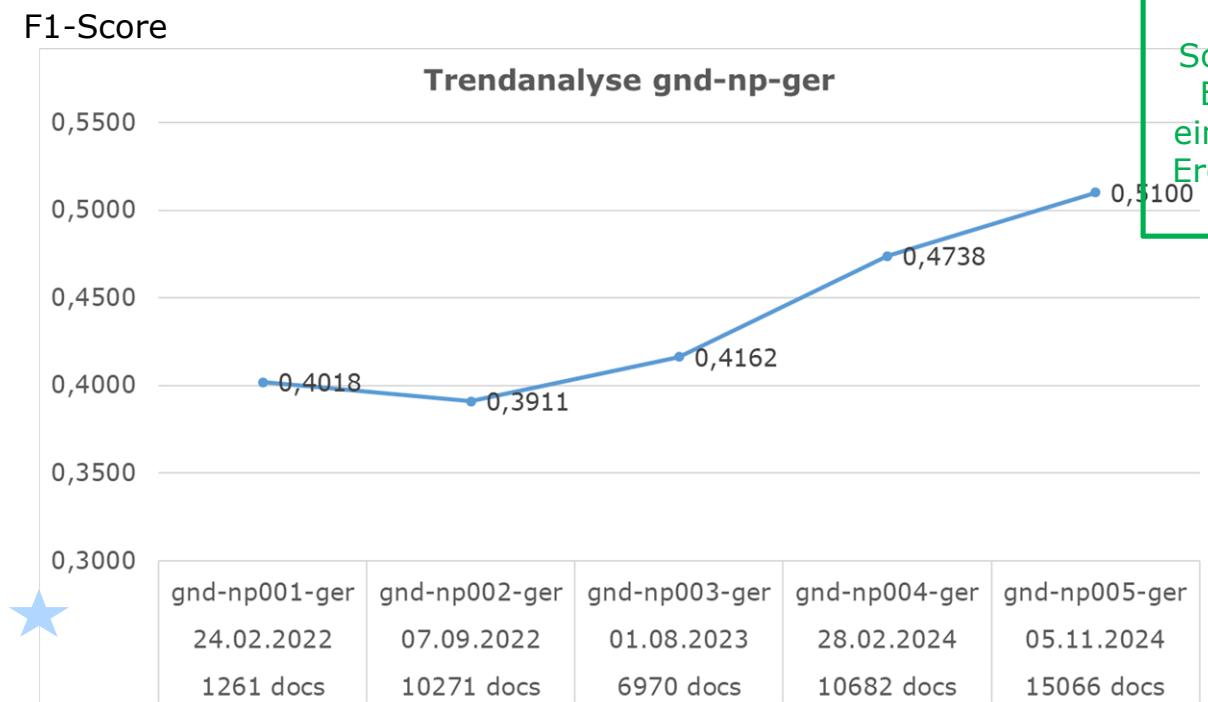
- Einsatz von XR-Transformer und LLMs
- Verarbeitung höherer Publikationsmengen

Die Ergebnisse werden im Abschlussbericht, in Vorträgen und durch die Organisation einer Fachtagung am 29. und 30. Januar 2026 kommuniziert

Verbesserung der Automatische Schlagwortvergabe im Regelbetrieb

vgl. Averbis-

Software: 0,28



Vielen Dank!



Projekt-Team: Maximilian Kähler, Lisa Kluge, Katja Konermann,

DNB: Christoph Poley, Nico Wagner

Mehr Infos zum KI-Projekt:

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

<https://blog.dnb.de/ki-projekt-gewinnt-best-paper-award/>

Fachtagung am 29. und 30. Januar 2026 in der DNB in Frankfurt: [KI in Bibliotheken weiterdenken – Datenqualität, Infrastruktur und Anwendungen für kleine und große Sprachmodelle](#)

Zwei parallele Workshops am 28. Januar:

1. [Workshop EMa Inside - Daten, Prozesse, Modelle und Infrastruktur für die automatische Inhaltserschließung](#)
2. [Workshop Evaluierung automatischer Verfahren zur Inhaltserschließung](#)

Informationen

- Poley, Christoph, et al. "Automatic Subject Cataloguing at the German National Library." *LIBER Quarterly: The Journal of the Association of European Research Libraries*, vol. 35, no. 1, 2025, pp. 1–29, <https://doi.org/10.53377/lq.19422>
- Kluge, Lisa, and Maximilian Kähler. 2025. "DNB-AI-Project at SemEval-2025 Task 5: An LLM-Ensemble Approach for Automated Subject Indexing." arXiv. <https://arxiv.org/abs/2504.21589>
- Kähler, M., Kluge, L., and Konermann, K. (2025). DNB-AI-Project at the GermEval-2025 LLMs4Subjects Task: KIFSPrompt – Knowledge-Injected Few-Shot Prompting. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops* (pp. 455–464). Hsh Applied Academics. <https://aclanthology.org/2025.konvens-2.42/>
- Kähler, M. (2025, September 16). *Benchmarking automatic indexing methods on German scientific literature* [Webinar]. Network Group "AI in Libraries" Webinars 2025, Conference of European National Librarians (CENL). https://c18004-vod.l.core.cdn.streamfarm.net/18004initag/ondemand/app2080931841/cenl/networkgroups/20250916_AI-webinar_AI-for-automated-Indexing.mp4
- Kluge, L. (2024, October 17). *LLM-few-shot-prompting for automated indexing* [Webinar]. Network Group "AI in Libraries" Webinars 2024, Conference of European National Librarians (CENL). https://c18004-vod.l.core.cdn.streamfarm.net/18004initag/ondemand/app2080931841/cenl/networkgroups/20241017_AI-webinar2024-3_LLM-few-shot-prompting_automated_indexing.mp4
- Fachtagung 29. Januar und 30. Januar 2026 in Frankfurt [KI in Bibliotheken weiterdenken – Datenqualität, Infrastruktur und Anwendungen für kleine und große Sprachmodelle](#)