Maximilian Kähler, German National Library (DNB)

Benchmarking Automatic Indexing Methods on German Scientific Literature

Network Group “AI in Libraries”
Webinars 2025

DEUTSCHE NATIONAL BIBLIOTHEK

Our Project: AI for automated indexing

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

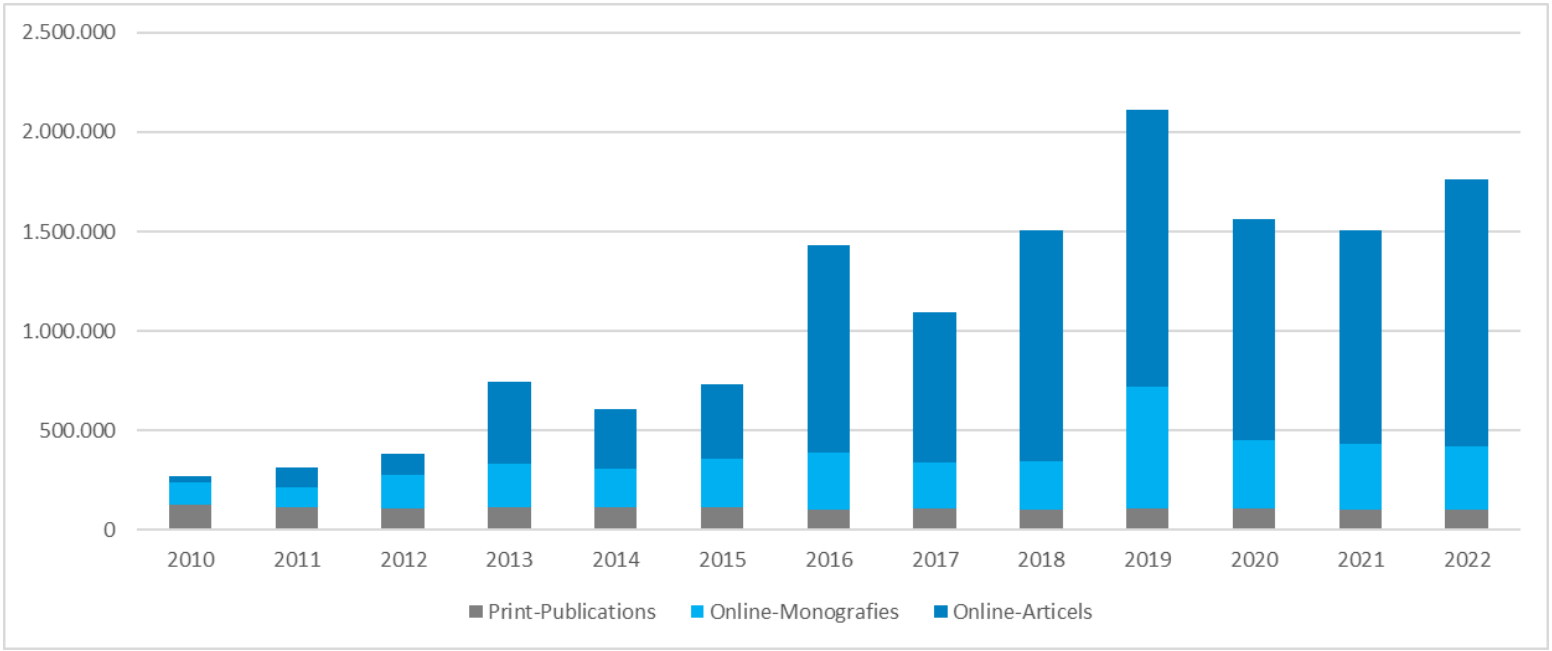
- Funded by German national AI-Strategy through BKM*
- Duration: 4 years (October 2021 – December 2025)

Project Goals:

- Accelerate knowledge transfer of NLP, Data Science and Machine Learning-Methods into the Organization
- Provide new methodological directions to machine-based subject indexing

*BKM: Federal Government Commissioner for Culture and the Media https://www.kulturstaatsminister.de/

Annual numbers of print- and online-publications collected by the DNB since 2010



Machine-based subject indexing@DNB

- In production since 2014
- Complete refactorization of software architecture in 04/2022[1]
- Core component is the open source library annif[2]
- Annual throughput of ~170.000 publications per year
- Our target vocabulary is the **Integrated Authority File**[3] (GND) containing >1.4M potential concepts

[1]Poley, C etal. (2025). Automatic Subject Cataloguing at the German National Library. https://doi.org/10.53377/lq.19422

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

[3]https://gnd.network/#

DEUTSCHE NATIONAL BIBLIOTHEK

Benchmarking Automated Indexing Methods – Why?

- Much progress has been made in NLP in recent years: deep learning, transformers, LLMs…
- Given the complexity of the GND, there is still room for improvement in achieving high-quality automatic indexing with our current system
- Don‘t reinvent the wheel

How did we evaluate?

Gold Standard comparison (Binary Relevance)

- Comparing machine based subject suggestions with previously annotated gold standard
- Gold standard follows the German Rules for Subject Cataloguing (RSWK)*

Expert-Rating (Graded Relevance)

- Qualitative Evaluation on all automated subject suggestions: How helpful are subject suggestions for retrieval?
- Evaluation on data that was not seen by the annotators before

Hardware Resources

- Training time
- Inference time
- GPU-usage

Two **Tasks**:

- Subject Indexing of Document Titles
- Subject Indexing of Long Documents (shortened at 30K characters)

Dataset: German Scientific Literature across 20 Subject groups

*RSWK: https://d-nb.info/1126513032/34

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

XMLC - Algorithms

annif
- ZestXML[1]
- Omikuji[2]
- Dismec++[3]

(annif)
- XR-Transformer[4]
- Attention-XML[5]
- NGAME[6]

DNB-Developed Methods

(annif)
- LLM-Ensemble[7]
- EBM[8]

Other Baselines

annif
- MLLM[9]
- Elastic-Search[10]

Gold-Standard comparison | Expert-Rating | Hardware Resources

[1,..,10]References on last slide

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

Omikuji[2]

- highly efficient re-implementation in rust of the partitioned label tree approach

- purely statistical approach (Label names and label features irrelevant)

- Text representation based on TFIDF-Features* (BoW**)

- Currently available in Annif

*TFIDF – Text Frequency (vs.) Inverse Document Frequency
**BoW – Bag-of-Words

[2]Omikuji (2018) Available at: https://github.com/tomtung/omikuji

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

Maui Like Lexical Matching (MLLM[9])

- Lexical Matching: controlled vocabulary is matched to the text lexically, sentence by sentence
- Matched candidates are ordered by ranker model
- Only ranker model needs training
- Currently available in Annif

[9]Suominen, O. (2021)

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

XR-Transformer[4]

- Fusion of partitioned label tree approach with recursive fine-tuning of transformer model
- XMLC Statistical Approach (Label names and label features irrelevant)
- Text representation both TFIDF and Transformer-based
- Integration into Annif in Testing-Phase (thank you ZBW!)

[4]Zhang, J. et al. (2021)

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

Embedding Based Matching (EBM)[8]

- Similar to MLLM: text is processed sentence-wise
- Matches are generated with vector search based on embedding similarity (in exchange for lexical matching)
- Matched candidates are ordered by ranker model
- Only ranker model needs training
- Annif integration under development by DNB

[8] Kähler, M. and Rietdorf, C. (2025)

DEUTSCHE NATIONAL BIBLIOTHEK

What methods did we evaluate?

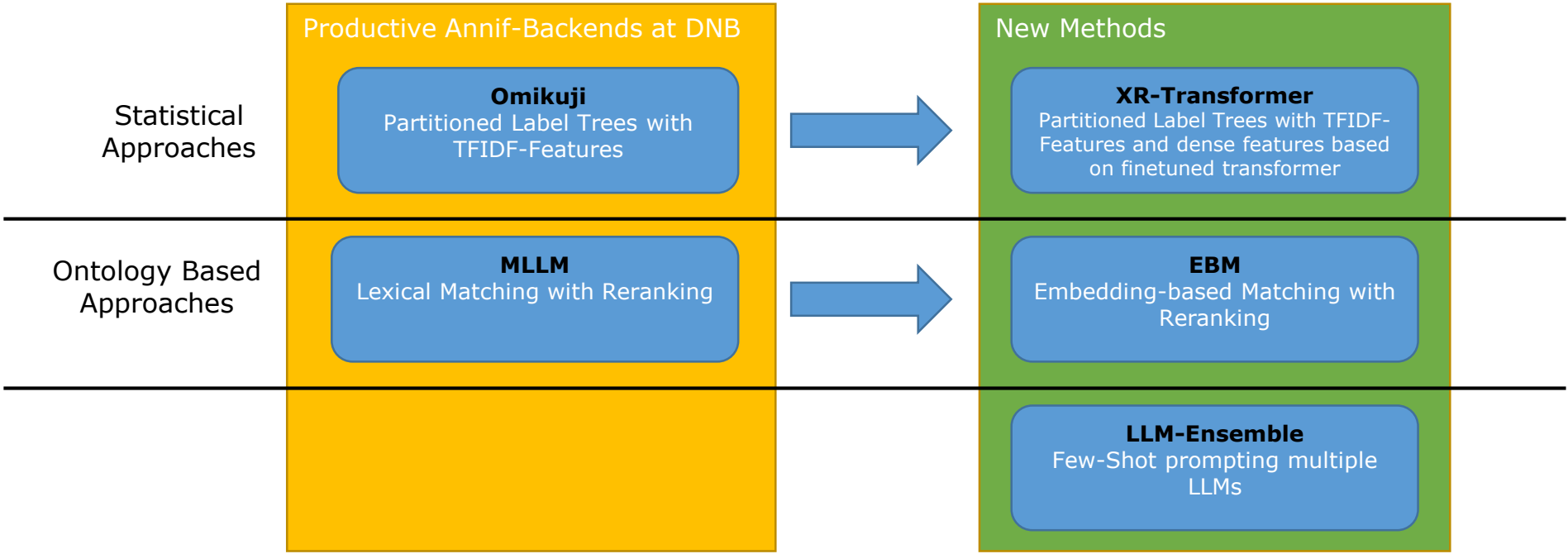
LLM-Ensemble[7]

- Few-Shot prompting of several LLMs with examples + input documents
- LLM-generated keywords are mapped with Encoder-Model to normed vocabulary
- No training required (works with off-the-shelf LLMs)
- no Annif integration

[7] Kluge, L. and Kähler, M. (2025)

DEUTSCHE NATIONAL BIBLIOTHEK

Summary of Methods



DEUTSCHE NATIONAL BIBLIOTHEK

Results

Comparing Subject Suggestions with an Example

DEUTSCHE NATIONAL BIBLIOTHEK

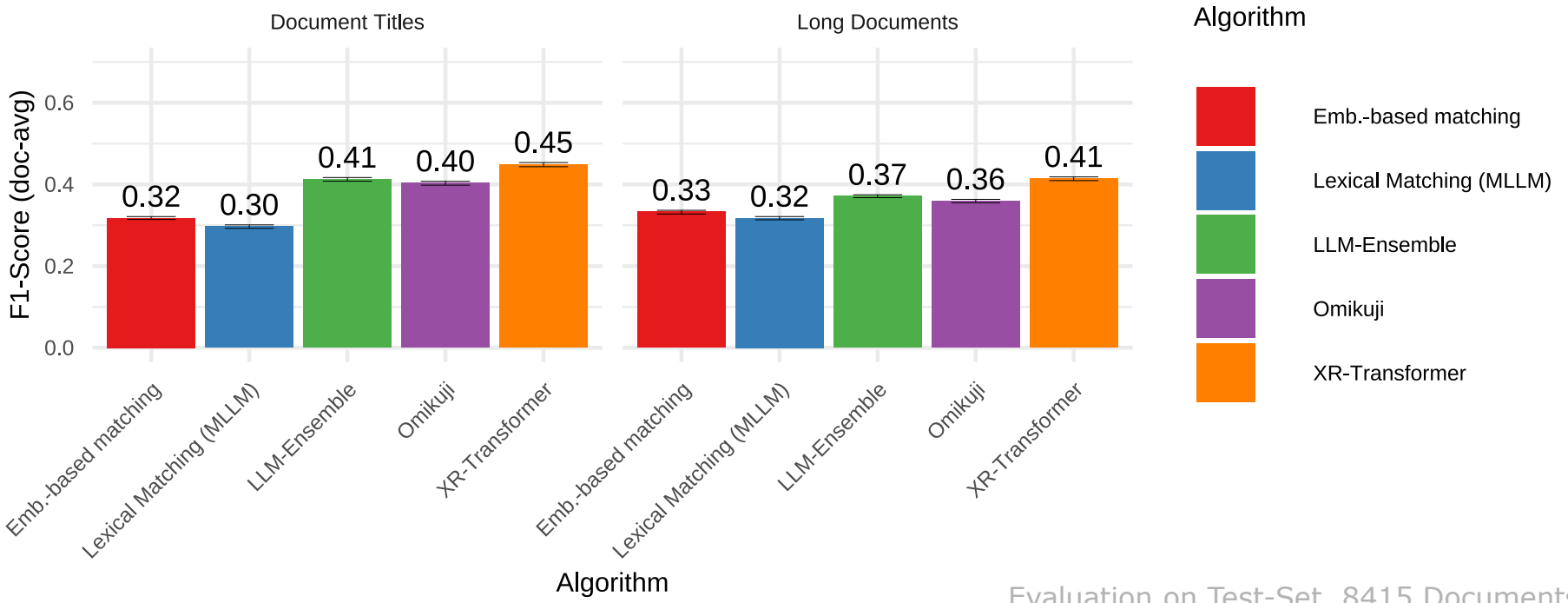
Example-Title: „Scientific Integrity in Qualitative Social Sciences and Public Health Research: Conflicts – Reflection – Expertise“
https://d-nb.info/1236332946

GND-Subject Term (Bold RSWK-Gold-Standard)	Omikuji	XR-Transformer	Embedding-bas. Matching	MLLM (lexikal. Matching)	LLM-Ensemble
Research	0.075	0.056	-	-	-
Social Conflict	0.078	-	-	-	-
Qualitative Social Science	0.097	0.053	0.431	-	0.567
Medicine	0.115	0.052	-	-	-
Healthcare System	0.141	0.050	-	-	-
Empirical Social Science	-	-	-	-	0,127
Research Ethics	-	-	0,073	-	0,573
Qualitative Method	-	0,050	0,125	-	-
Research Process	-	-	-	-	-
Public Health Research	-	-	-	-	0,089
Moral Action	-	-	-	-	-
Conflict	-	-	0.059	0.076	-
Research Method	-	-	0.085	-	-
Expertise	-	-	-	0.139	-
EXPERTIS	-	-	-	0.139	-

True positives
False negatives
False positives
True negatives

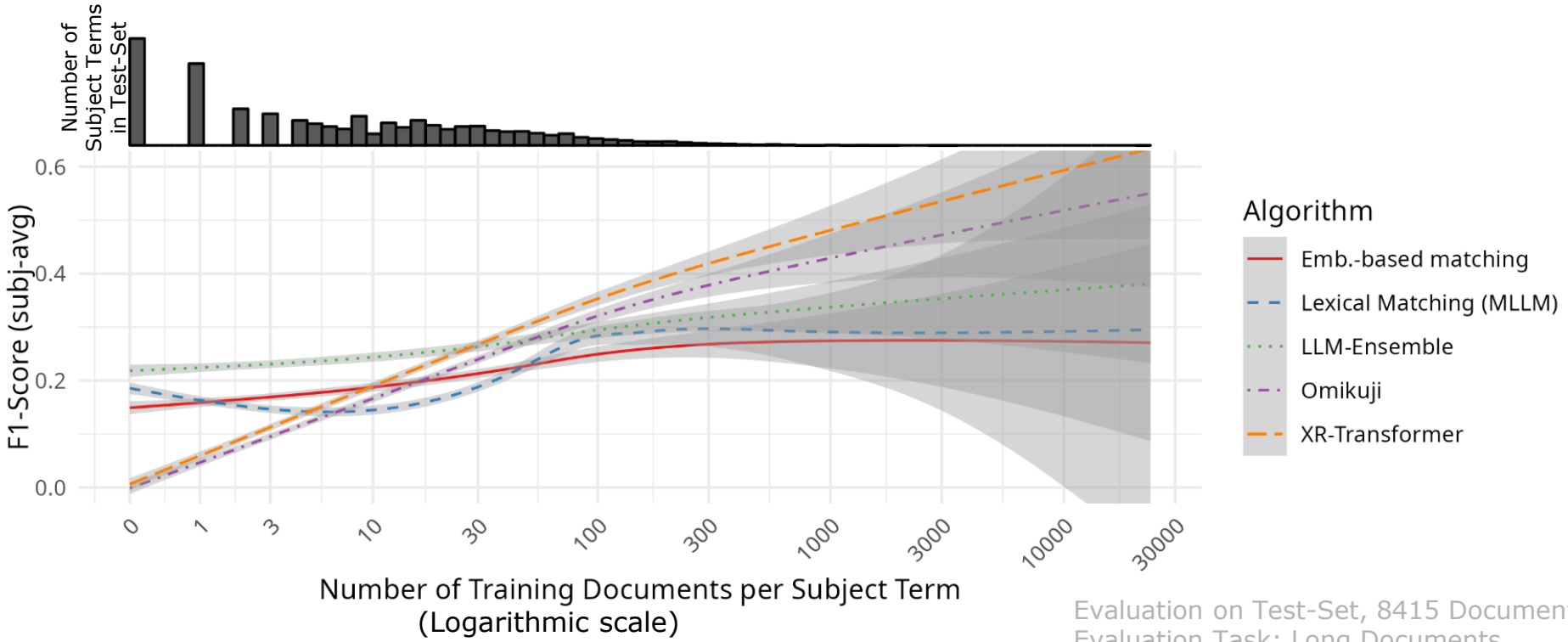
Numbers represent Confidence-Scores produced by the respective methods

Comparing Results with Gold-Standard



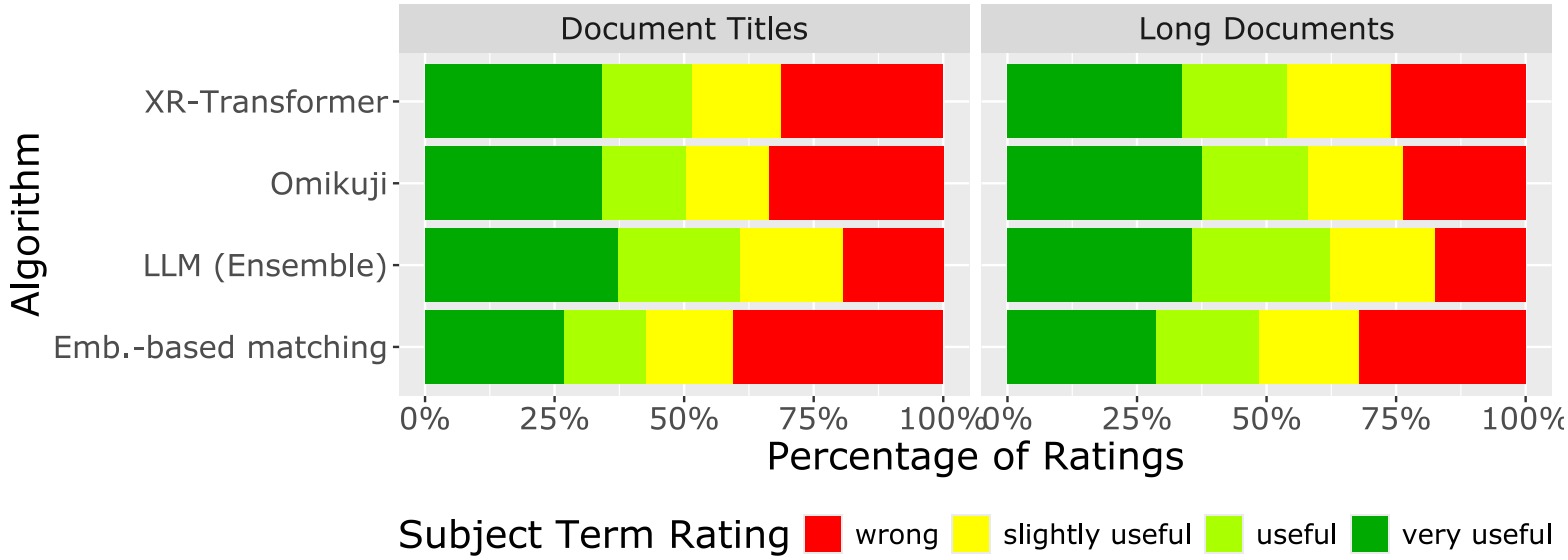
Evaluation on Test-Set, 8415 Documents

Quantitative Performance vs. Subject Term Frequency



Evaluation on Test-Set, 8415 Documents
Evaluation Task: Long Documents

Qualitative Rating by Subject Experts



Each Algorithm was evaluated on a distinct test-set of ~1100 German scientific documents across 20 subject groups

Metrics: Calculation of Precision and Recall with Graded Relevance

	Binary Relevance	Graded Relevance[11]
(Generalized) Precision	$\frac{tp}{tp + fp}$	$\frac{tp + \Delta_{rel}}{tp + fp}$
(Generalized) Recall	$\frac{tp}{tp + fn}$	$\frac{tp + \Delta_{rel}}{tp + fn + \Delta_{rel}}$

$$\Delta_{rel} = \sum_{i \in false\ positives} r_i$$

With relevance ratings $0 \leq r_i \leq 1$

tp - true positives

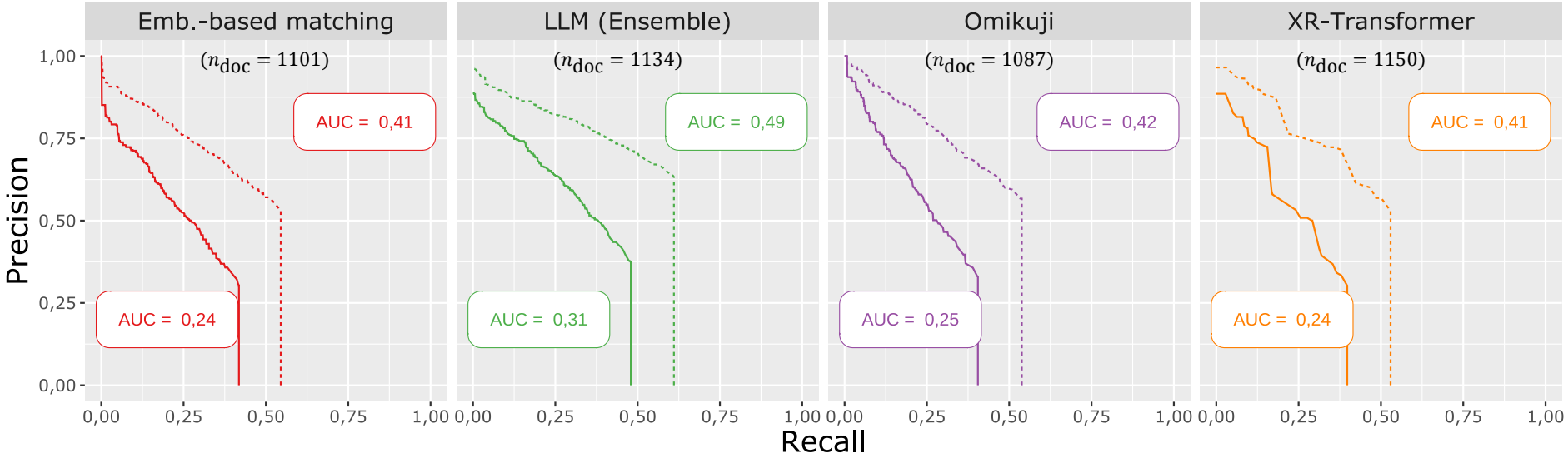
fp – false positives

fn – false negatives

[11] Kekäläinen, J. and Järvelin, K. (2002)

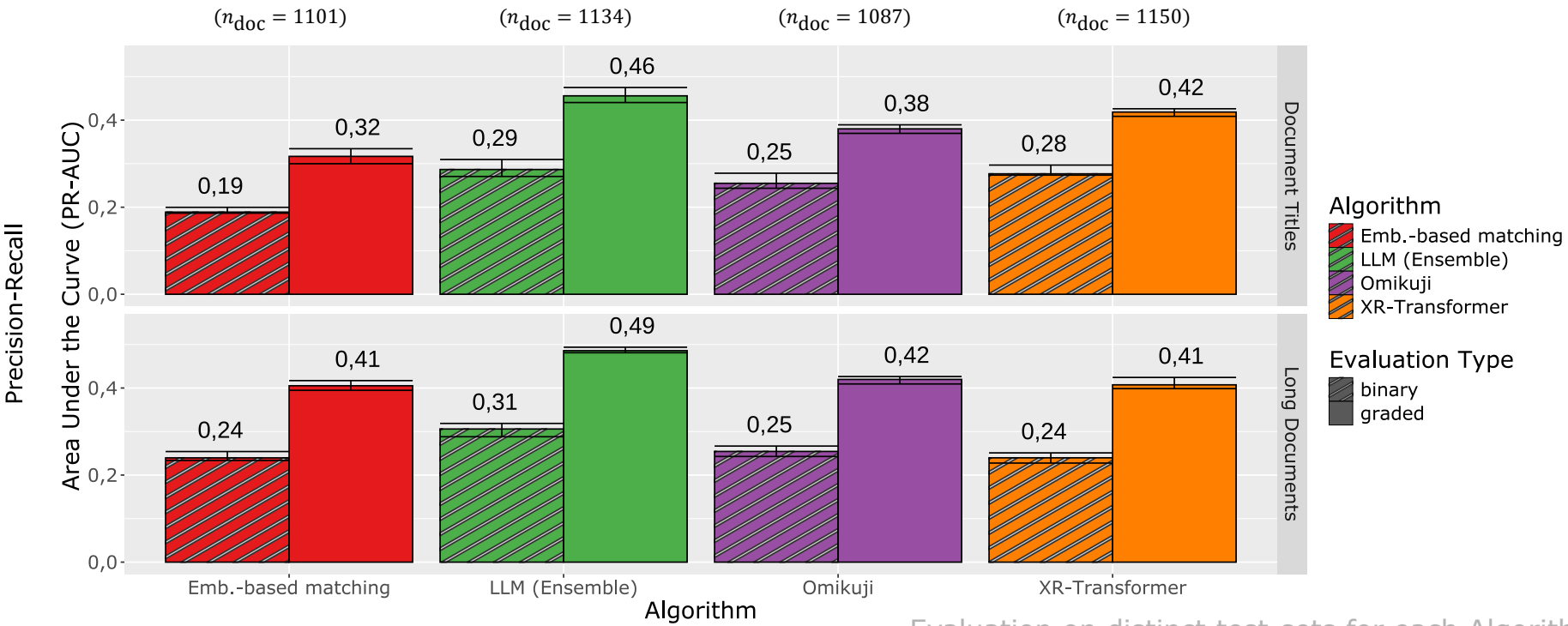
Binary and Graded Relevance

Task: Long Documents



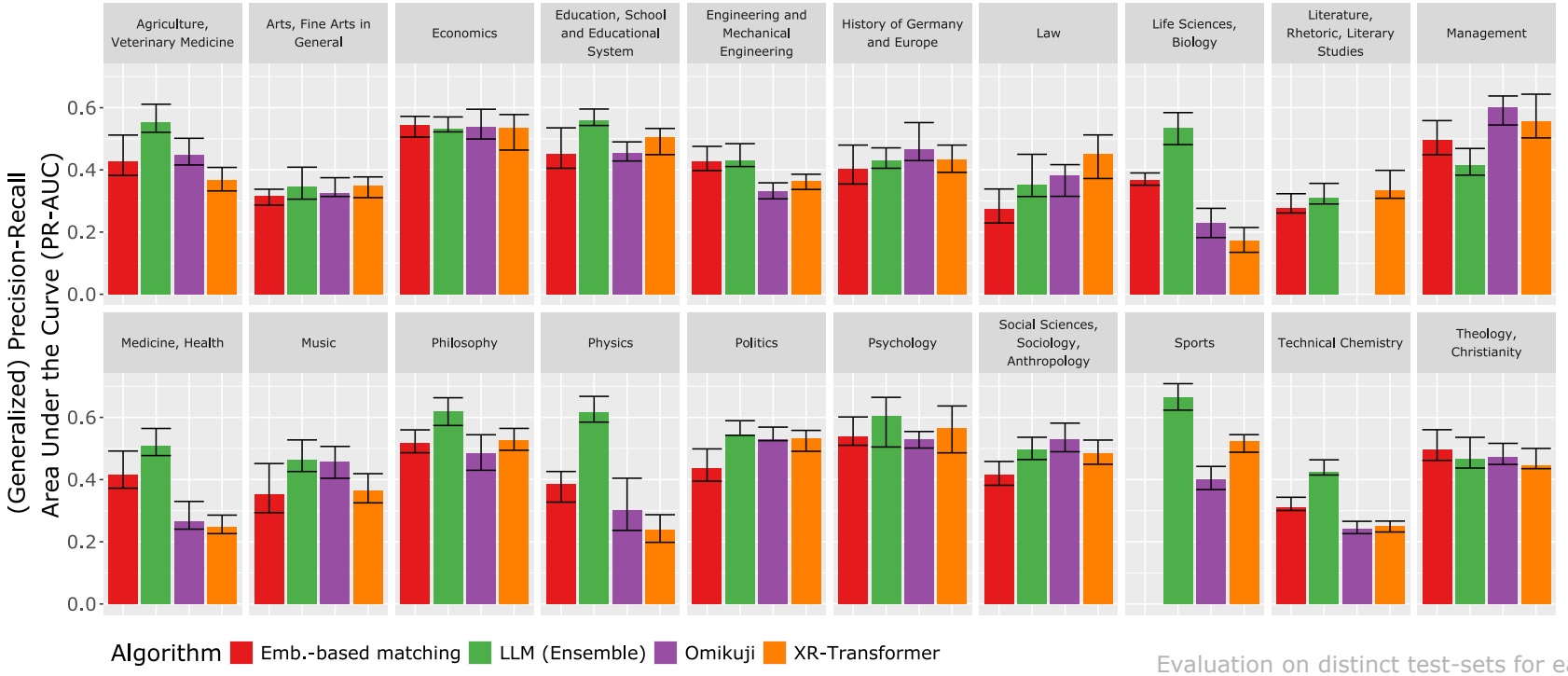
Evaluation on distinct test-sets for each Algorithm

Binary and Graded Relevance



Evaluation on distinct test-sets for each Algorithm

Comparative Results per Subject Group, Qualitative Ratings

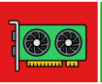


Evaluation on distinct test-sets for each Algorithm
Evaluation Task: Long Documents

Summary

DEUTSCHE NATIONAL BIBLIOTHEK

Overview of Strengths and Weaknesses in Automated Indexing Algorithms

	Training-costs (with self-curated data)	Inference -costs	Few- and Zero-Shot Predictions	Processing long documents	Disambiguation	Generalization
Omikuji	-	++	--	+*	+ (Label names irrelevant)	o (for frequent labels)
Lexical Matching (MLLM)	+	++	o	++	-	--
Embedding bas. Match. (EBM)	+ 	o 	o	++	o	--
XR-Transformer	-- 	o 	--	o (Only TFIDF-Features)	+ (Label names irrelevant)	o (for frequent labels)
LLM-Ensemble	++	-- 	++	+*	+	+

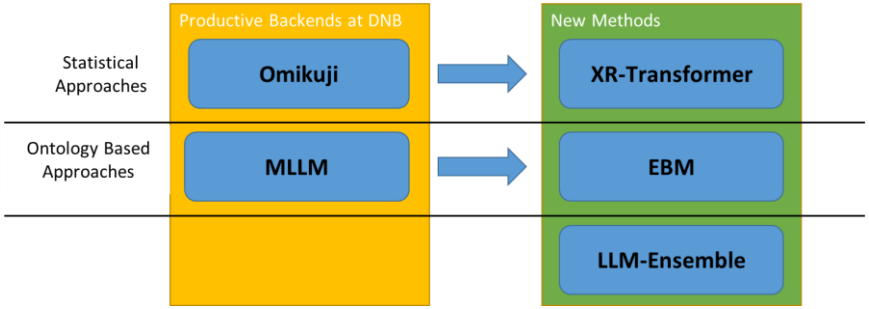
 GPU-Usage recommended  GPU-Usage necessary *larger RAM or GPU-RAM necessary

Precise measurements depend on application scenario and hardware configuration

GPU-image: Flaticon.com

Summary

- Outfitting existing methods (MLLM, Omikuji) with transformer based text representations (EBM, XR-Transformer) brings performance boosts
- Qualitative evaluation suggests that LLM-based methods offer great potential, but come with higher inference costs and hardware requirements



DEUTSCHE NATIONAL BIBLIOTHEK

More on LLMs for Subject Indexing

- Previous CENL Webinar Talk by Lisa Kluge:
  - LLM Few-Shot Prompting for Automated Indexing
- Two Publications created for LLMs4Subjects:
  - An LLM-Ensemble Approach for Automated Subject Indexing[7]
  - KIFSPrompt - Knowledge-Injected Few-Shot Prompting[12]
- See also: LLMs4Subjects – Task Overview Paper[13]

[7]Kluge, L. and Kähler, M. (2025)
[12]Kähler, M. et al. (2025)
[13]D'souza, J. et al. (2025)

References

1. Gupta, N. et al. (2021) "Generalized Zero-Shot Extreme Multi-label Learning," Available at: https://doi.org/10.1145/3447548.3467426
2. Omikuji (2018) Available at: https://github.com/tomtung/omikuji
3. Schultheis, E. and Babbar, R. (2021) "Speeding-up One-vs-All Training for Extreme Classification via Smart Initialization." Available at: https://doi.org/10.48550/arxiv.2109.13122
4. Zhang, J. et al. (2021) "Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification." Available at: https://arxiv.org/abs/2110.00685v2
5. You, R. et al. (2019) "AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification," Available at: https://doi.org/10.5555/3454287.3454810
6. Dahiya, K. et al. (2023) "NGAME: Negative Mining-aware Mini-batching for Extreme Classification," Available at: https://doi.org/10.1145/3539597.3570392
7. Kluge, L. and Kähler, M. (2025) "DNB-AI-Project at SemEval-2025 Task 5: An LLM-Ensemble Approach for Automated Subject Indexing," Available at: https://aclanthology.org/2025.semeval-1.148/
8. Kähler, M. and Rietdorf, C. (2025) Embedding Based Matching for Automated Subject Indexing. Available at: https://github.com/deutsche-nationalbibliothek/ebm4subjects
9. Suominen, O. (2021) Maui Like Lexical Matching, https://github.com/NatLibFi/Annif/wiki/Backend%3A-MLLM.
10. https://www.elastic.co/elasticsearch
11. Kekäläinen, J. and Järvelin, K. (2002) "Using graded relevance assessments in IR evaluation," Available at: https://doi.org/10.1002/asi.10137.
12. Kähler, M. et al. (2025) "DNB-AI-Project at the GermEval-2025 LLMs4Subjects Task: KIFSPrompt - Knowledge-Injected Few-Shot Prompting," Available at: https://doi.org/10.25968/opus-3679
13. D'souza, J. et al. (2025) "SemEval-2025 Task 5: LLMs4Subjects - LLM-based Automated Subject Tagging for a National Technical Library's Open-Access Catalog," Available at: https://aclanthology.org/2025.semeval-1.328/

DEUTSCHE NATIONAL BIBLIOTHEK

Thank you!

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

Please get in touch for further questions and discussion:

Maximilian Kähler

m.kaehler@dnb.de

@mfakaehler@openbiblio.social

Our Project@DNB:

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

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

DEUTSCHE NATIONAL BIBLIOTHEK

Appendix

DEUTSCHE NATIONAL BIBLIOTHEK

Ensemble Contribution:

Beispiel: Expected contribution of models in a potential future ensemble