Lisa Kluge

LLM Few-Shot Prompting for Automated Indexing

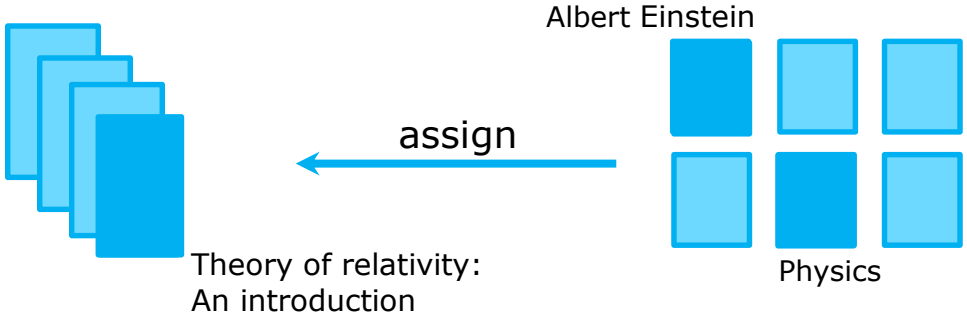
Overview

1. **Introduction**
2. **Experiments**
3. **Closed- vs. open-source LLMs**
4. **LLMs vs. other methods**
5. **Discussion & Outlook**

DNB‘s AI project (2021 – 2025)

- Funded by the German Minister for Culture and the Media as part of the AI strategy
- Project aims:
  - Improve automated subject indexing by testing methods from the field of NLP and AI
  - Knowledge transfer, build up know-how

LLM Few-Shot Prompting for Automated Indexing



Scientific digital publications

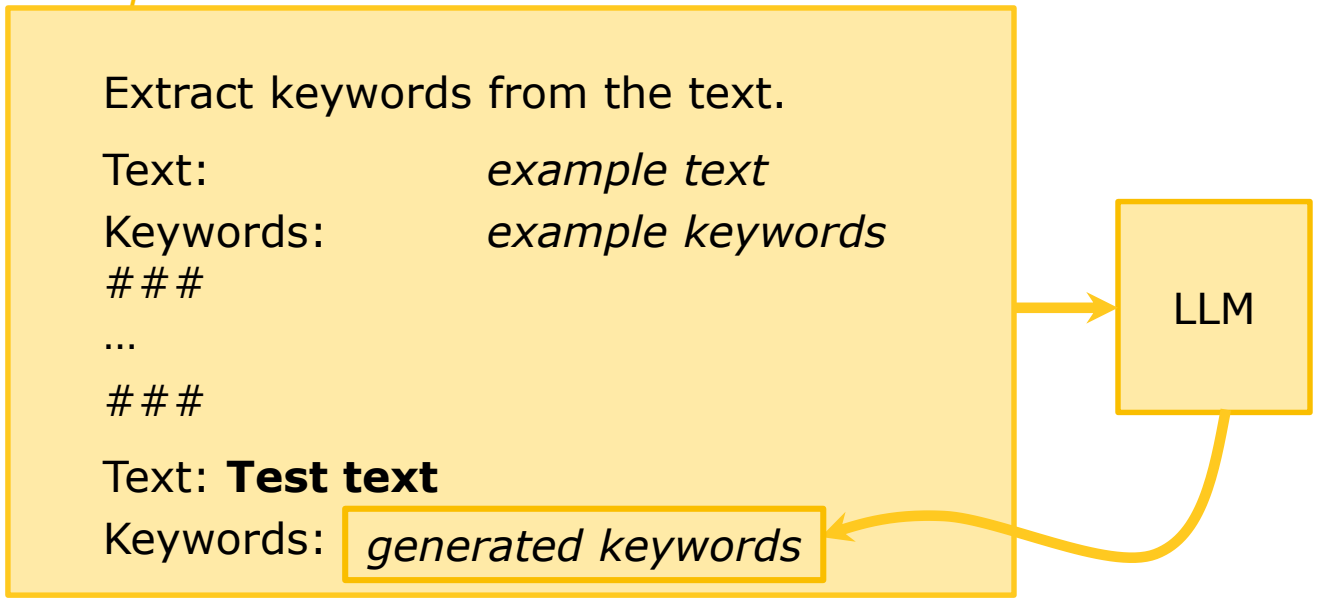
titles

full texts

Integrated Authority File (GND)

Subset of ~200k entities

LLM Few-Shot Prompting for Automated Indexing



Wanhae Lee, Minki Chun, Hyeonhak Jeong, and Hyunggu Jung. 2023. **Toward keyword generation through large language models**. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23 Companion*:37–40.

LLM Few-Shot Prompting for Automated Indexing

- Used embeddings: BAAI/bge-m3
  - $M3^2$: multi-functionality, multi-linguality, multi-granularity
  - 1024 dimensions

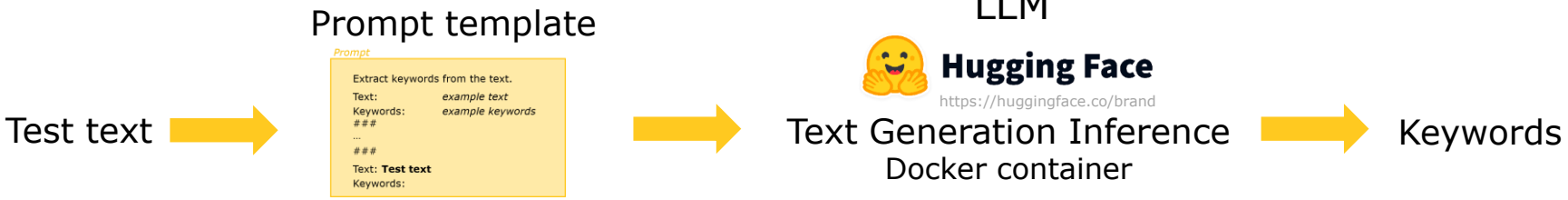
Yaxin Zhu and Hamed Zamani. 2024. **ICXML: An in-context learning framework for zero-shot extreme multi-label classification**. *arXiv preprint arXiv:2311.09649*.
Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian and Zheng, Liu. 2024. **M3-Embedding: Multi-linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation**. *Findings of the Association for Computational Linguistics ACL 2024:*2318-2335.

LLM Few-Shot Prompting for Automated Indexing

Generation

Mapping

LLM Few-Shot Prompting for Automated Indexing



Mapping

LLM Few-Shot Prompting for Automated Indexing

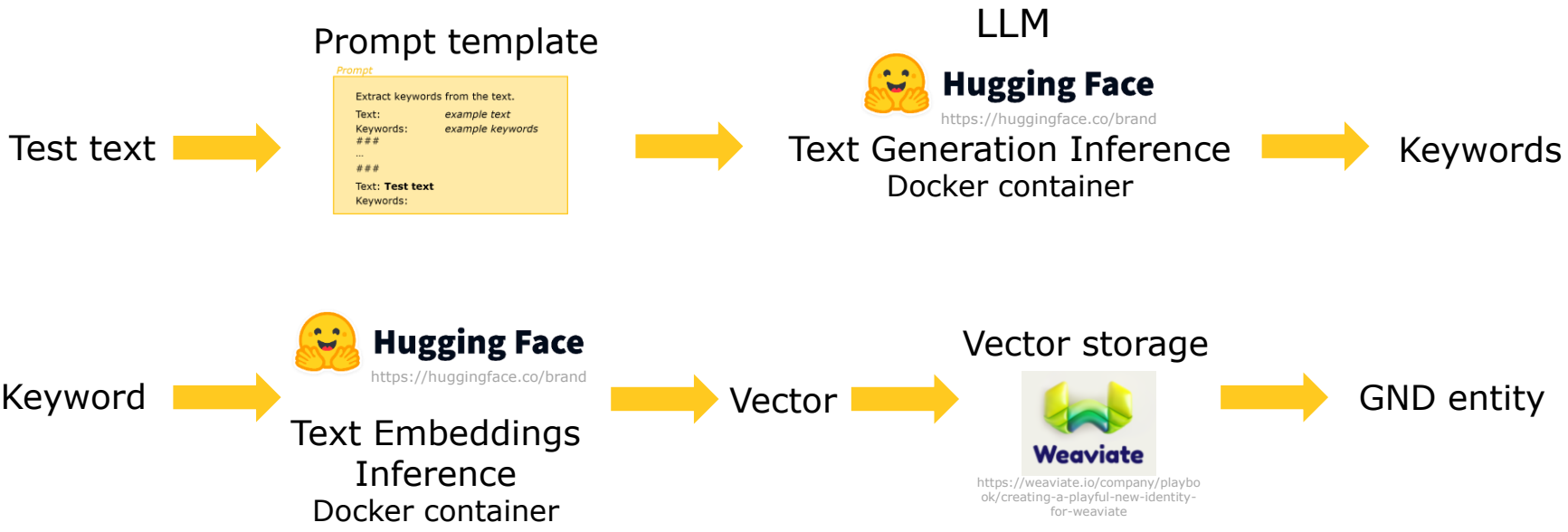
Generation

Keyword → 

Text Embeddings Inference
Docker container
→ Vector → Vector storage 

https://weaviate.io/company/playbook/creating-a-playful-new-identity-for-weaviate
→ GND entity

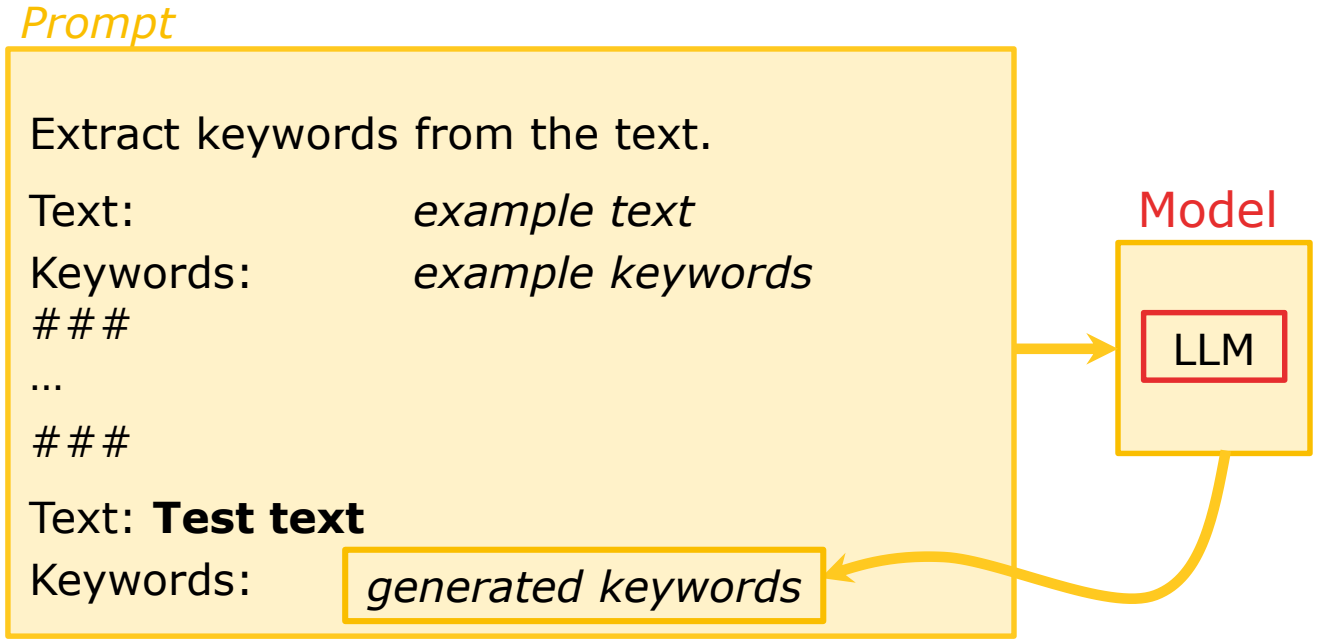
LLM Few-Shot Prompting for Automated Indexing



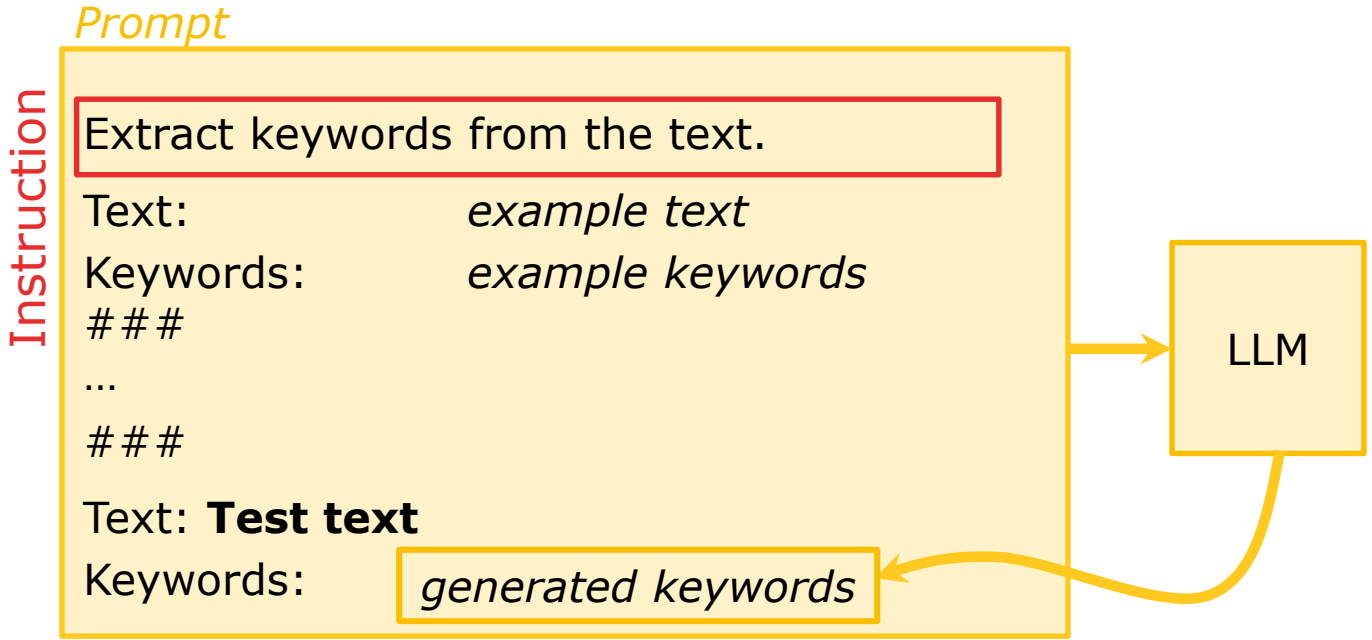
Overview

1. **Introduction**
2. **Experiments**
3. **Closed- vs. open-source LLMs**
4. **LLMs vs. other methods**
5. **Discussion & Outlook**

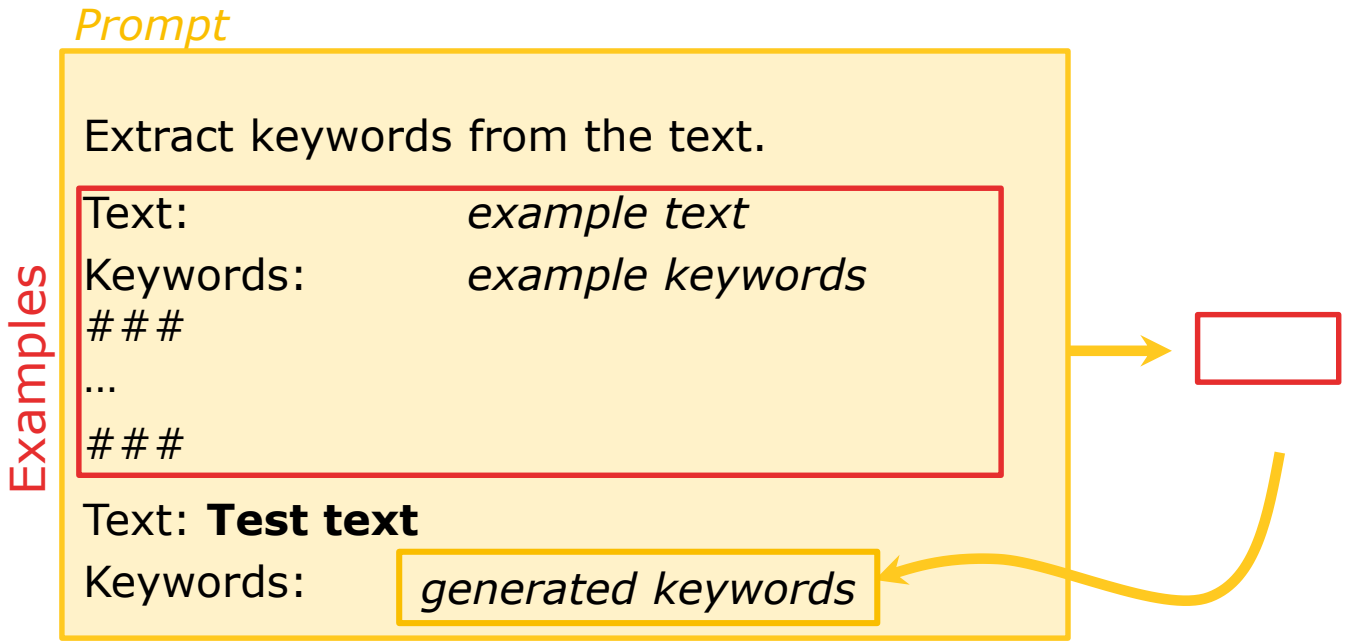
Experiments – What to vary?



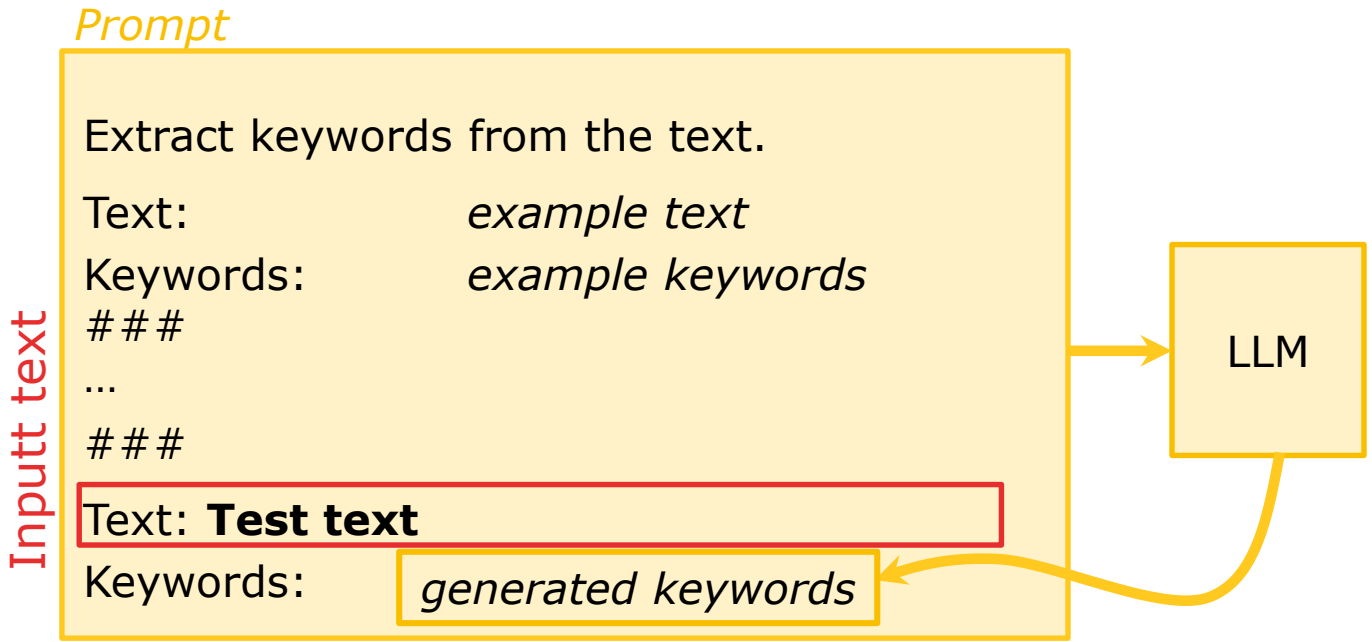
Experiments – What to vary?



Experiments – What to vary?

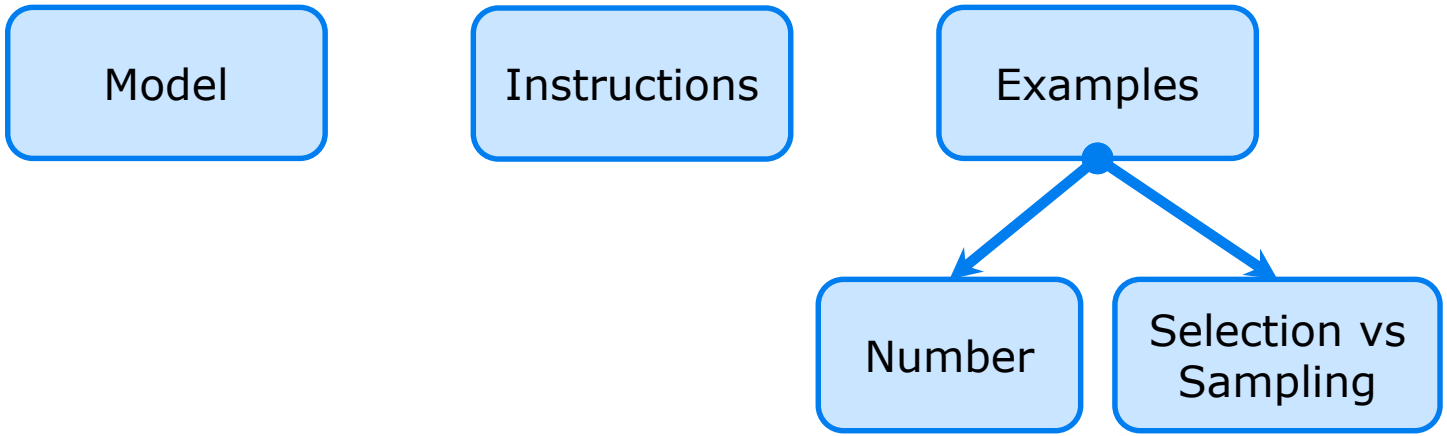


Experiments – What to vary?



Title experiments

- Parameters (selection)



Title experiments

Model

- Well-perfoming models[1]:
  - teknium/OpenHermes-13B
  - teknium/OpenHermes-2.5-Mistral-7B
  - meta-llama/Meta-Llama-3-8B-Instruct
  - openchat/openchat-3.5-0106

[1]All models available via Hugging Face – The AI community building the future.

Title experiments

Instructions

- 4 slightly different instructions:
  - *No instruction.*
  - Extract keywords from titles.
  - This is a conversation between an intelligent, helpful AI-assistant and a user. *(The assistant is an expert in the field of assigning keywords to scientific publications.)* The assistant returns keywords upon the input text of the user.

Title experiments

Examples

- Between 1 and 12 examples
- Selection vs. sampling
  - Manual: guided by criteria regarding the similarity between title and labels
  - Random: sampled from the train split

Title experiments – example result

Example title

Text: Programming iOS apps with Swift
Keywords: Swift 3.0

- Suggestions:

Swift (Programming language)

iPhone

iOS

Programming

Development

Title experiments – results*

Model + Prompt	P	R	F1
teknium/OpenHermes-2.5-Mistral-7B, random prompt (6 examples)	0.345	0.324	**0.334**
teknium/OpenHermes-13B, manually created prompt with similar examples	**0.527**	0.159	0.244
meta-llama/Meta-Llama-3-8B-Instruct, manually created prompt with many labels per example	0.165	**0.354**	0.225

*Examplary intermediate results 09/24

Different LLMs work well with different settings.
Different settings optimise different metrics.

Title experiments – ensemble*[1]

i ≥	P	R	F1
1	0.169	**0.467**	0.249
2	0.265	0.411	0.322
3	0.318	0.378	0.346
…			
5	0.400	0.328	**0.360**
…			
10	0.602	0.153	0.244
11	**0.662**	0.081	0.144

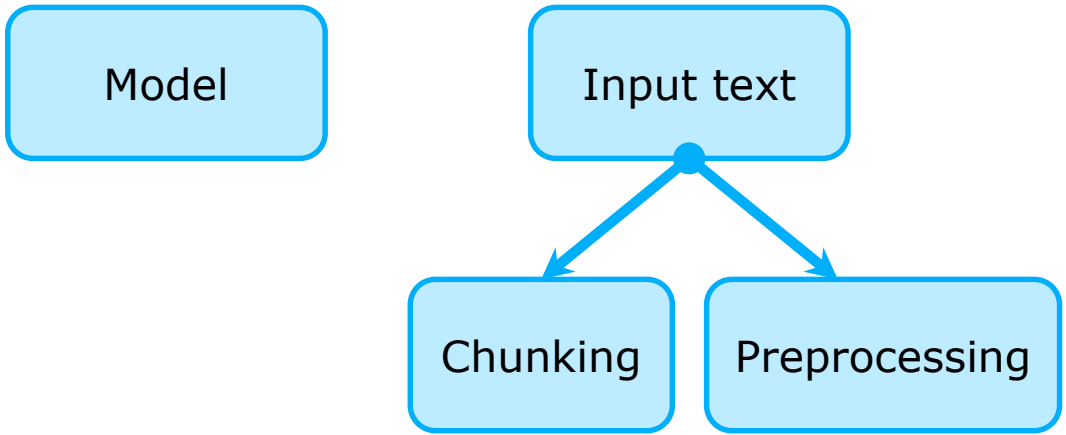
*Exemplary intermediate results 09/24

Combining the results of 11 models by counting at least how many models make a suggestion (**i**)

[1]Martin Toepfer and Christin Seifert. 2020. **Fusion architectures for automatic subject indexing under concept drift: Analysis and empirical results on short texts**. *International Journal on Digital Libraries*, 21(2):169–189.

Full text experiments

– Parameters (selection)



Full text experiments

Model

- Difficulty: find models with suitable context length
- Best performing model[1]:
  - meta-llama/Meta-Llama-3.1-8B-Instruct

[1]Model available via Hugging Face – The AI community building the future.

Full text experiments

Input text

- Chunking aspects:
  - Chunk length: 2.500 to 10.000 tokens
  - Number of chunks: 1 to 5
- Preprocessing of the text
  - Generate a summary
  - Extract the table-of-contents

Full text experiments – results*

Chunking and Preprocessing	P	R	F1
1 chunk with 10.000 tokens, prompt with 7 examples	**0.254**	0.342	**0.292**
2 chunks with 5000 tokens, prompt with 7 examples	0.166	**0.434**	0.240
Prompt with 7 examples, using summary as input text	0.223	0.353	0.274

*Exemplary intermediate results 09/24

DEUTSCHE NATIONAL BIBLIOTHEK

Overview

1. **Introduction**
2. **Experiments**
3. **Closed- vs. open-source LLMS**
4. **LLMs vs. other methods**
5. **Discussion & Outlook**

Closed- vs. Open-source LLMs

Hosting

Closed-Source

- Maintained by the vendor

Open-Source

- Local deployment possible

Closed- vs. Open-source LLMs

Cost scheme

Closed-Source

- Maintained by the vendor
- Cost per request

Open-Source

- Local deployment possible
- Hardware costs

Closed- vs. Open-source LLMs*

Text type	LLM	P	R	F1
Title	Luminous-base	0.275	0.281	0.278
	Teknium/OpenHermes-13B	**0.358**	**0.309**	**0.331**
Shortened full text	Luminous-base[1]	0.142[1]	**0.348[1]**	0.202[1]
	Meta-Llama-3.1-8B-Instruct	**0.289**	0.347	**0.315**

*Exemplary intermediate results 09/24

[1]Due to copyright, only *open-access* full texts were used here.

Closed- vs. Open-source LLMs*

Improvement with insights from open-source experiments

Text type	LLM	P	R	F1
Title	Luminous-base	0.306	0.292	0.299
	Teknium/OpenHermes-13B	**0.358**	**0.309**	**0.331**
Shortened full text	Luminous-base[1]	0.142[1]	**0.348[1]**	0.202[1]
	Meta-Llama-3.1-8B-Instruct	**0.289**	0.347	**0.315**

*Exemplary intermediate results 09/24

[1]Due to copyright, only *open-access* full texts were used here.

Closed- vs. Open-source LLMs

(Dis-)Advantages*

Closed-Source

- Maintained by the vendor
- Cost per request

– Copyrighted texts potentially not usable

\+ Stable performance

Open-Source

- Local deployment possible
- Hardware costs

– Set-up infrastructure

\+ Test many models

*from our subjective point of view

DEUTSCHE NATIONAL BIBLIOTHEK

Overview

1. **Introduction**
2. **Experiments**
3. **Closed- vs. open-source LLMs**
4. **LLMs vs. other methods**
5. **Discussion & Outlook**

Comparison against other methods

Type of method

LLMs

- Transformer[1] architecture

Omikuji[2]

- Partitioned Label Trees

MLLM[3]

- Lexical Method based on Maui[4]

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. **Attention is all you need**. *Proceedings of the 31st International Conference on in Neural Information Processing Systems*.
[2] https://github.com/tomtung/omikuji
[3] https://github.com/NatLibFi/Annif/wiki/Backend:-MLLM
[4] Olena Medelyan. 2009. **Human-competitive automatic topic indexing**. Ph.D. thesis, The University of Waikato, New Zealand.

Comparison against other methods

Text representation

LLMs

- Transformer[1] architecture
- Natural language text → Embeddings

Omikuji[2]

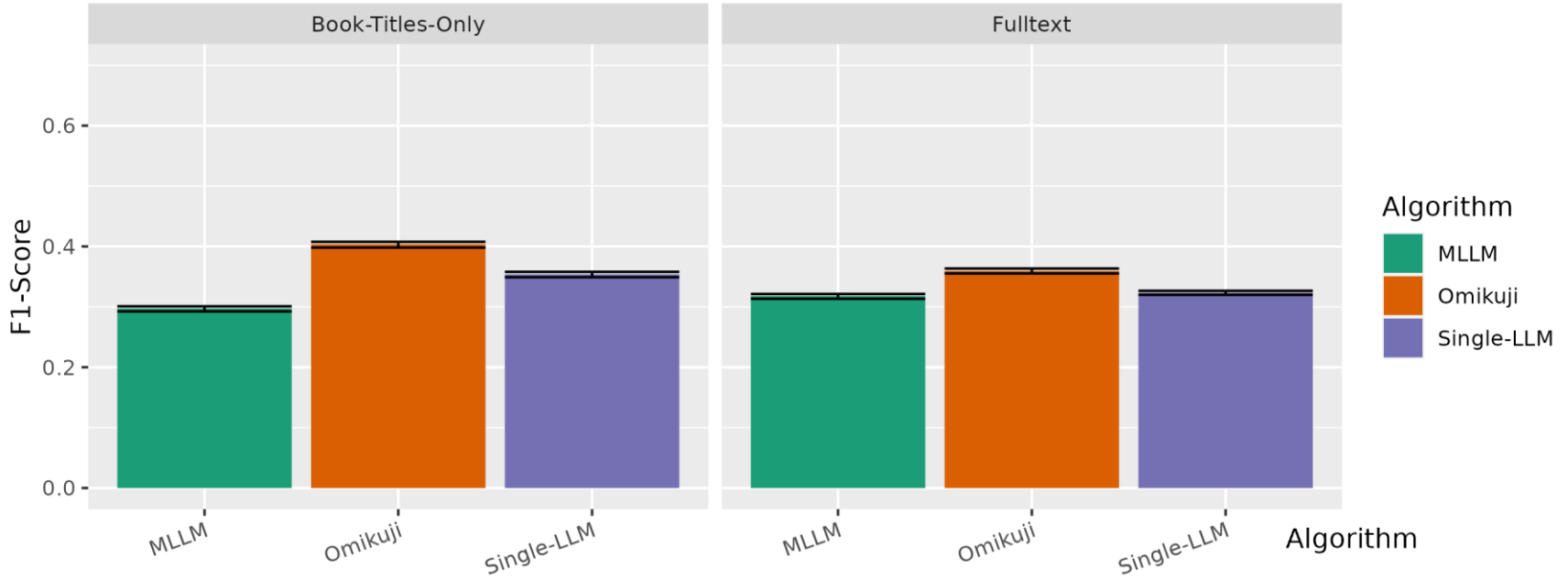
- Partitioned Label Trees
- TF-IDF-Matrices

MLLM[3]

- Lexical Method based on Maui[4]
- Preprocessed Text

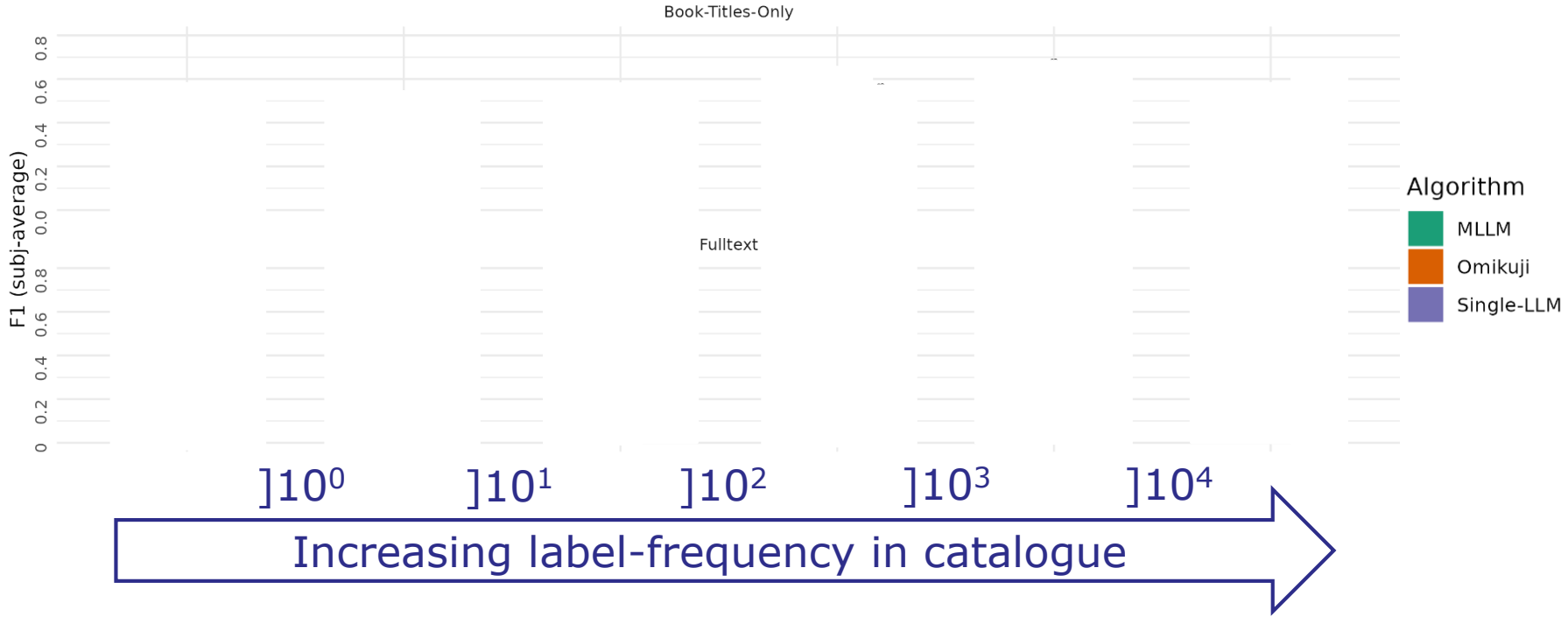
[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. **Attention is all you need**. *Proceedings of the 31st International Conference on in Neural Information Processing Systems*.
[2] https://github.com/tomtung/omikuji
[3] https://github.com/NatLibFi/Annif/wiki/Backend:-MLLM
[4] Olena Medelyan. 2009. **Human-competitive automatic topic indexing**. Ph.D. thesis, The University of Waikato, New Zealand.

Comparison against other methods*



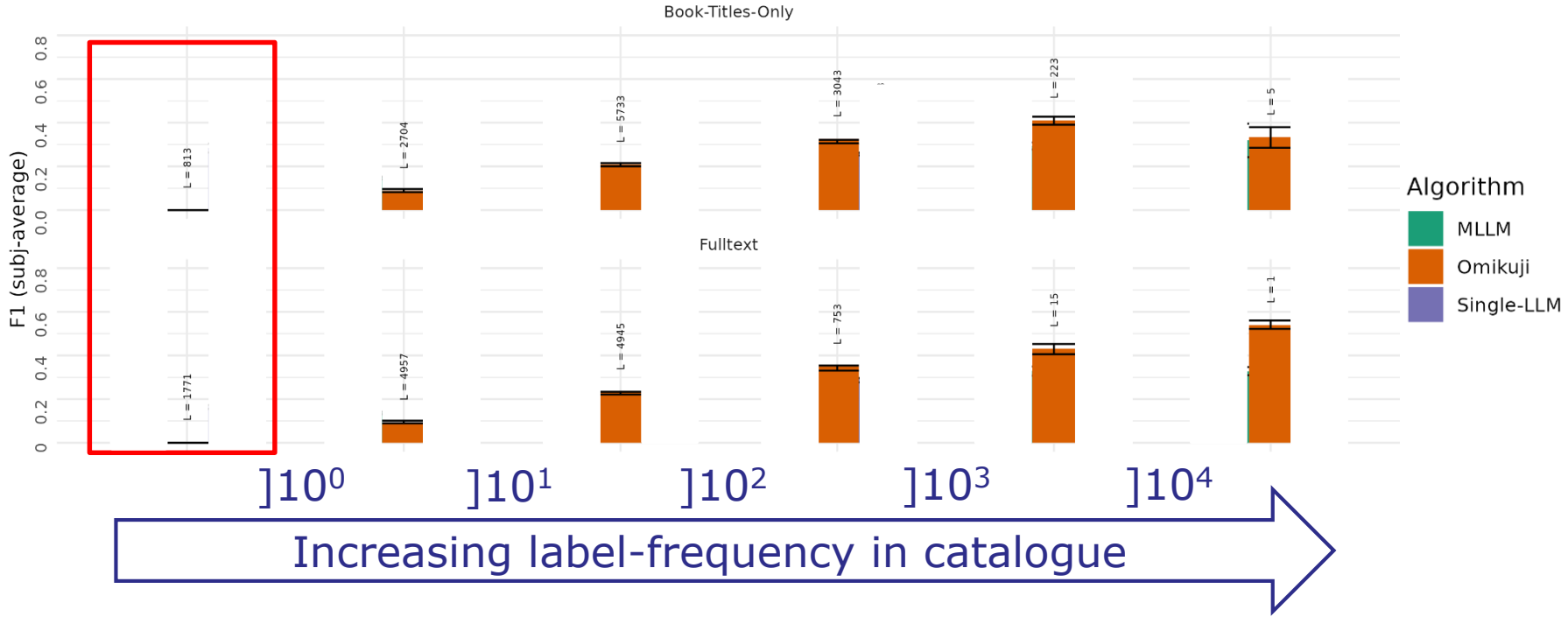
*Exemplary intermediate results 09/24

Comparison against other methods*



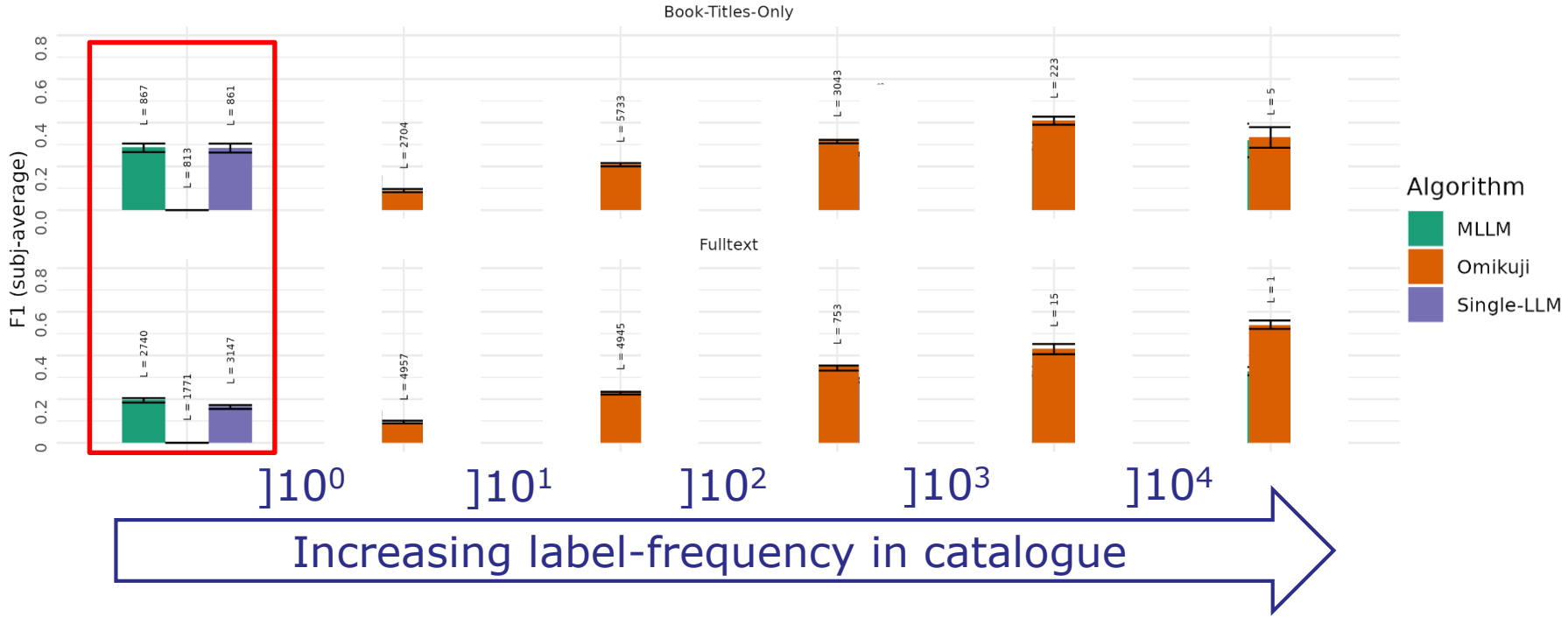
*Exemplary intermediate results 09/24

Comparison against other methods*



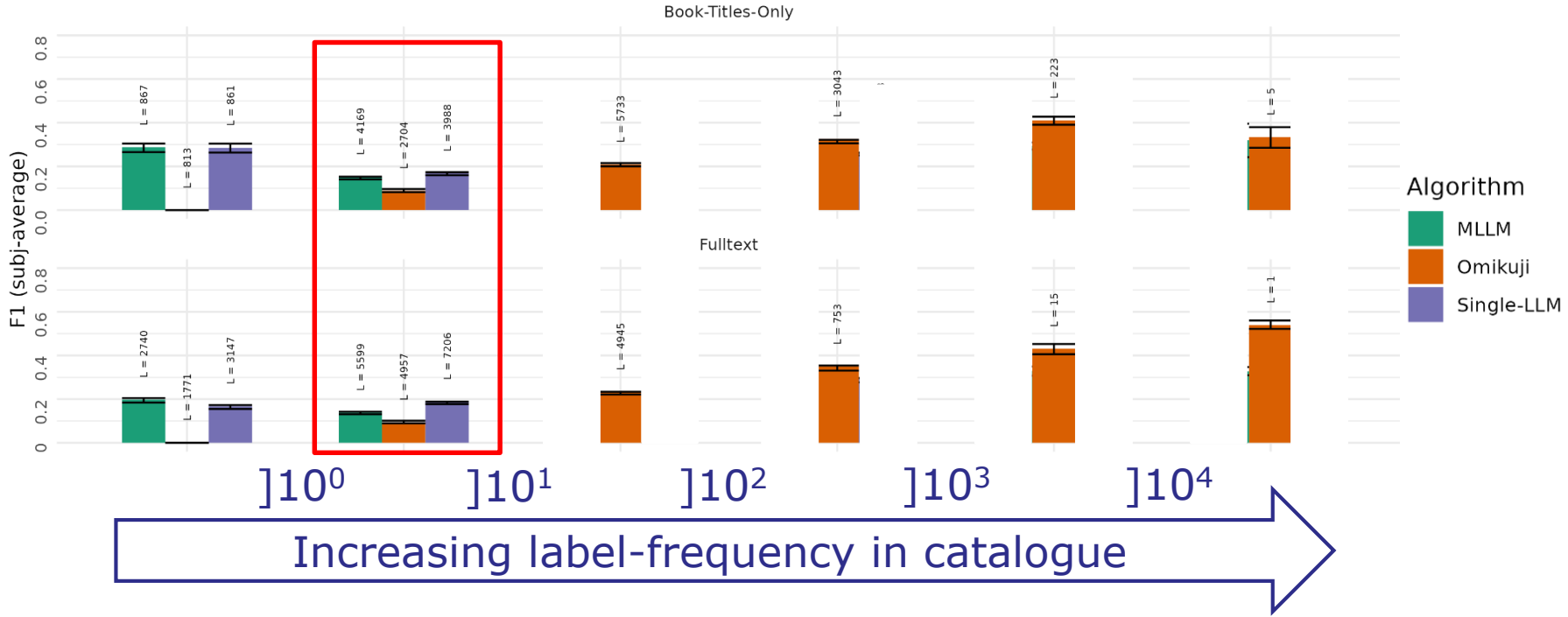
*Exemplary intermediate results 09/24

Comparison against other methods*



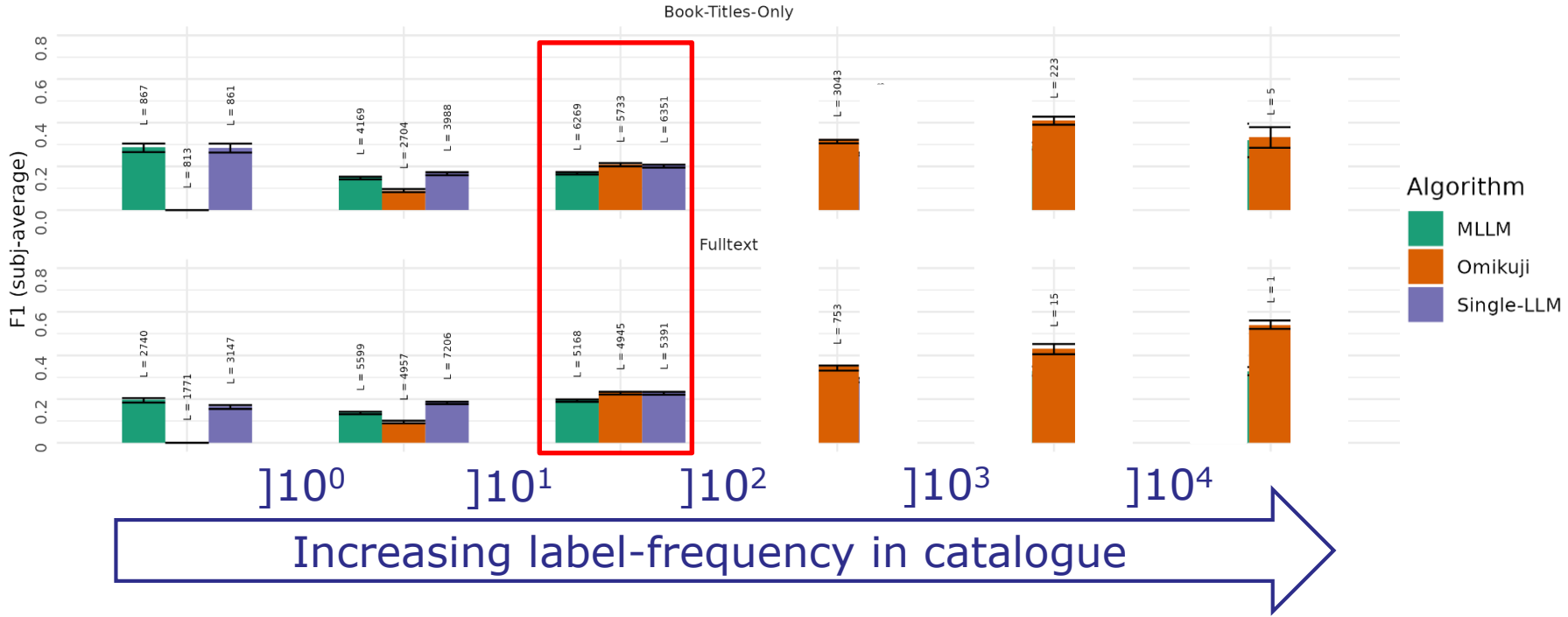
*Exemplary intermediate results 09/24

Comparison against other methods*



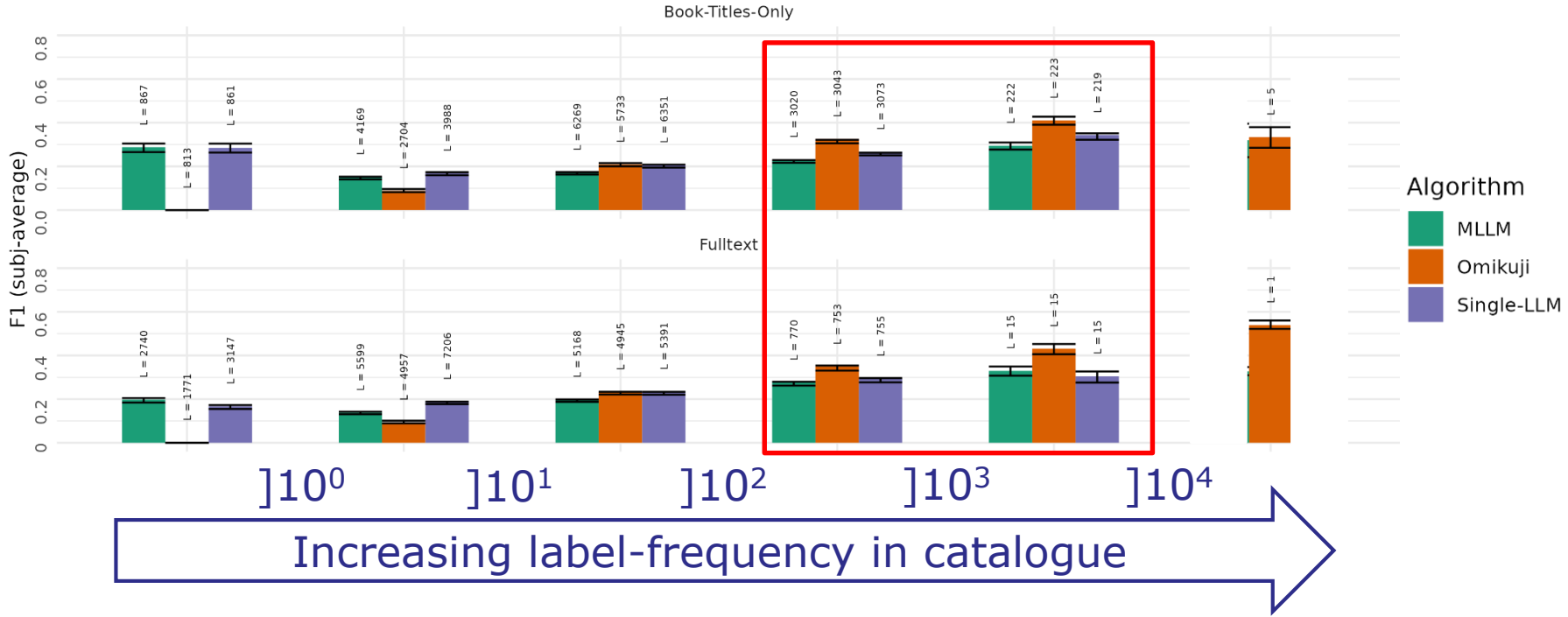
*Exemplary intermediate results 09/24

Comparison against other methods*



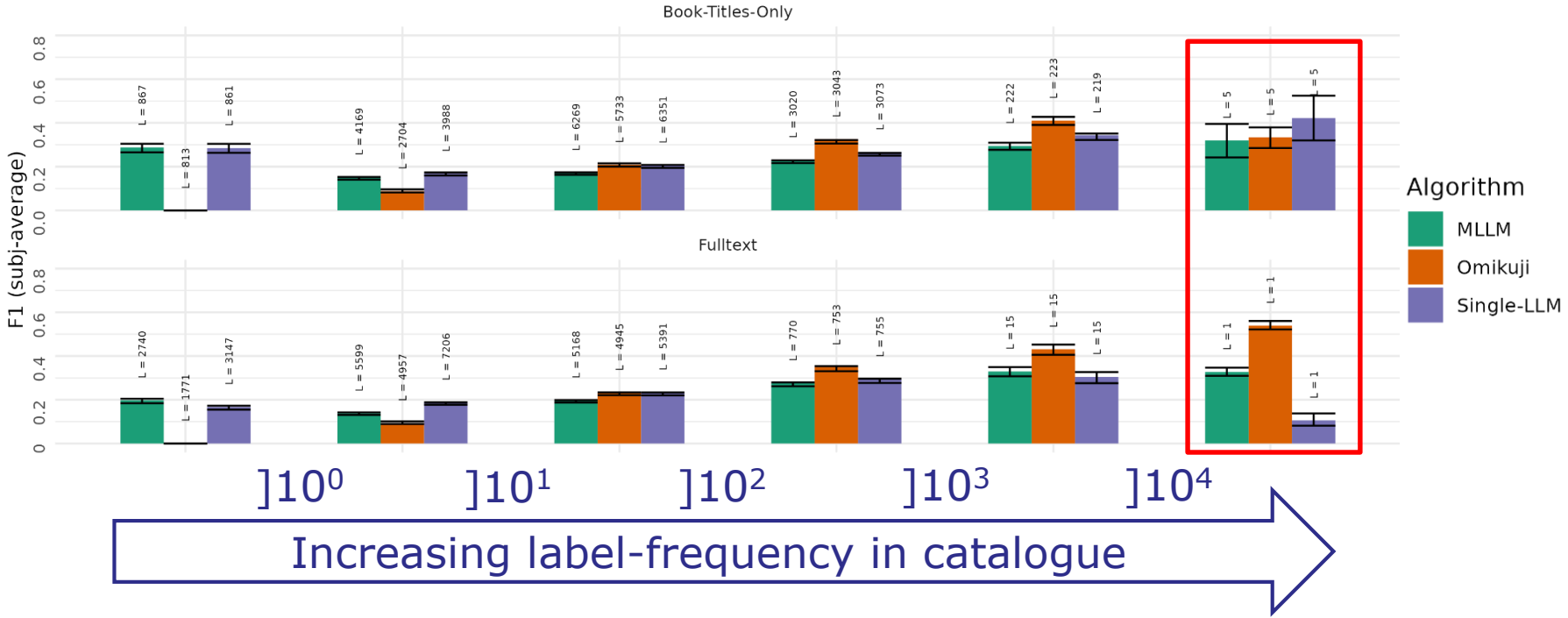
*Exemplary intermediate results 09/24

Comparison against other methods*



*Exemplary intermediate results 09/24

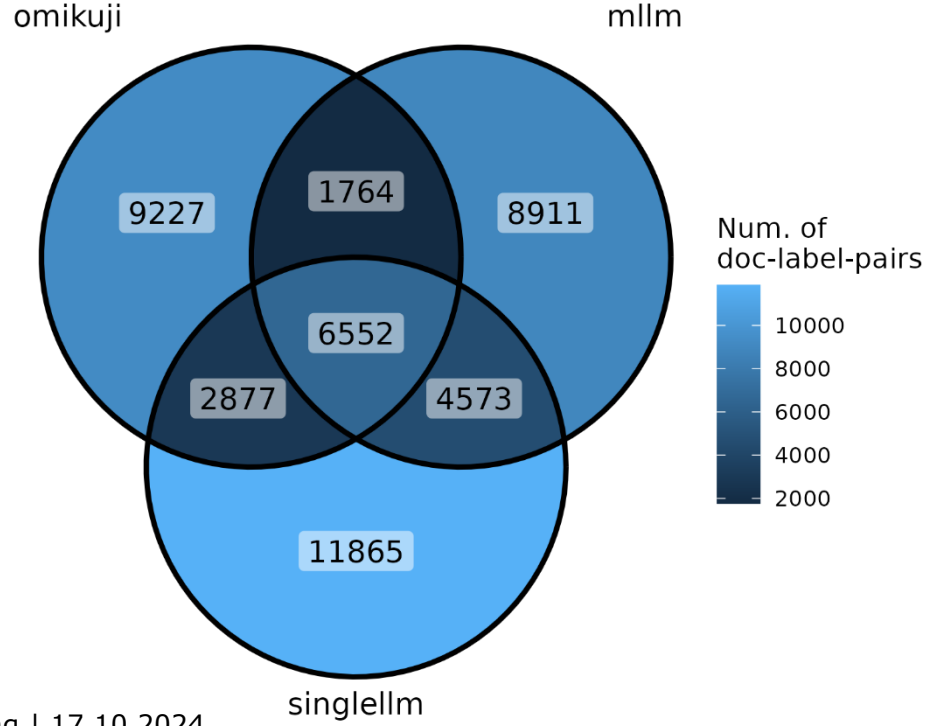
Comparison against other methods*



*Exemplary intermediate results 09/24

Comparison against other methods*

Venn-Diagram of suggestions of Omikuji, MLLM and the best open-source LLM experiment on the title corpus



*Exemplary intermediate results 09/24

Comparison against other methods*

Example title: *´Research morality in qualitative social and health research Conflicts – Reflection – Expertise´*

LLMs	Omikuji	MLLM
Conflict	Research	Conflict
Expertise	Social conflict	Expertise
Empirical social research	Qualitative social research	EXPERTIS
Research ethics	Medicine	
Health Sciences	Healthcare	

Missing labels:

Qualitative method, research process, moral behaviour

Incorrect suggestions
Correct suggestions

*Exemplary intermediate results 09/24

Comparison against other methods

(Dis-)Advantages*

LLMs

- Transformer architecture
- Natural language text → Embeddings

+ Zero-shot handling
- Resource intensity
+ - Off-the-shelf solution

Omikuji

- Partitioned Label Trees
- TF-IDF-Matrices

+ Good performance
- No use of label text

MLLM

- Lexical Method based on Maui
- Preprocessed Text

+ Zero-shot handling
- Depends on lexical overlap

*from our subjective point of view

Overview

1. **Introduction**
2. **Experiments**
3. **Closed- vs. open-source LLMs**
4. **LLMs vs. other methods**
5. **Discussion & Outlook**

Discussion – Are LLMs worth it?

- How do LLMs compare to other methods?
  - So far, they can't beat the other tested methods numerically.
  - Experts will assess how close suggestions are to gold standard.
- Strengths: zero-shot handling, no problems with synonyms, ability to generalise
- Capacity: 1 GPU (A100): run models up to 13B params (so far)

Outlook

- Fine-tune (smaller) LLMs
  - Parameter-efficient fine-tuning (PEFT)[1]
- D'Oosterlinck et al. (2024)[2]: InferRetrieveRank
  - Use DSPy[3] framework
  - Automate prompt tuning
- XR-Transformer[4]

[1]Neil Housby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan and Sylvain Gelly. 2019. **Parameter-efficient transfer learning for NLP.** *International conference on machine learning:*2790-2799.

[2]Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. 2024. **In-context learning for extreme multi-label classification**. *arXiv preprint arXiv:2401.12178.*
[3]DSPy Documentation | DSPy (dspy-docs.vercel.app)

[4]Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, Inderjit Dhillon. 2021. **Fast multi-resolution transformer fine-tuning for extreme multi-label text classification.** *Advances in Neural Information Processing Systems 34:* 7267-7280.

Thanks for your attention!

Project Team: Maximilian Kähler, Katja Konermann, Nico Wagner, Markus Schumacher

Contact: Lisa Kluge (l.kluge@dnb.de)

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