# Cross-Dialect Information Retrieval: Information Access in Low-Resource and High-Variance Languages

Robert Litschko, Oliver Kraus, Verena Blaschke, Barbara Plank

COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

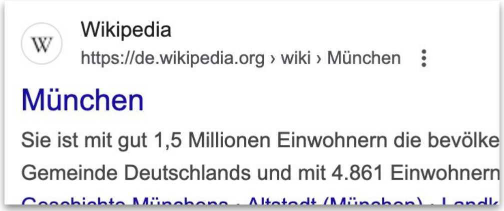
CIS

nlp

mcml
Munich Center for Machine Learning

🔍 München (“Munich”)

# München ("Munich")



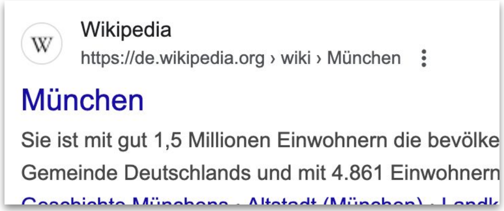
Wikipedia
https://de.wikipedia.org › wiki › München

München

Sie ist mit gut 1,5 Millionen Einwohnern die bevölke
Gemeinde Deutschlands und mit 4.861 Einwohnern
Geschichte Münchens · Altstadt (München) · Landk

# München (“Munich”)

What about culture-specific knowledge that can often be found in dialect Wikis?

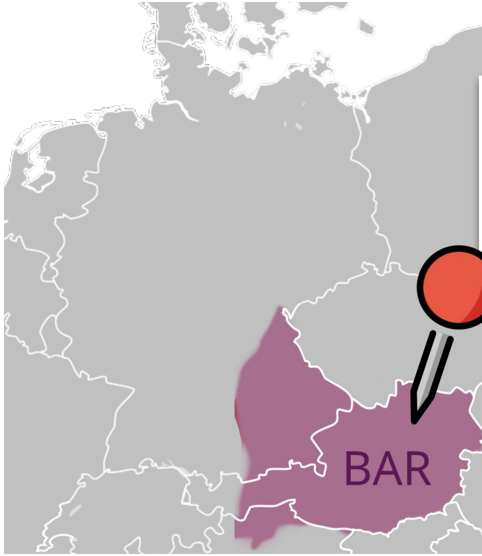


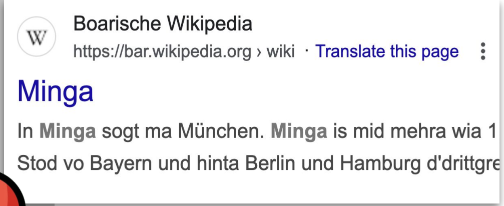
Wikipedia
https://de.wikipedia.org › wiki › München

München

Sie ist mit gut 1,5 Millionen Einwohnern die bevölke
Gemeinde Deutschlands und mit 4.861 Einwohnern
Geschichte Münchens · Altstadt (München) · Landk

# München (“Munich”)



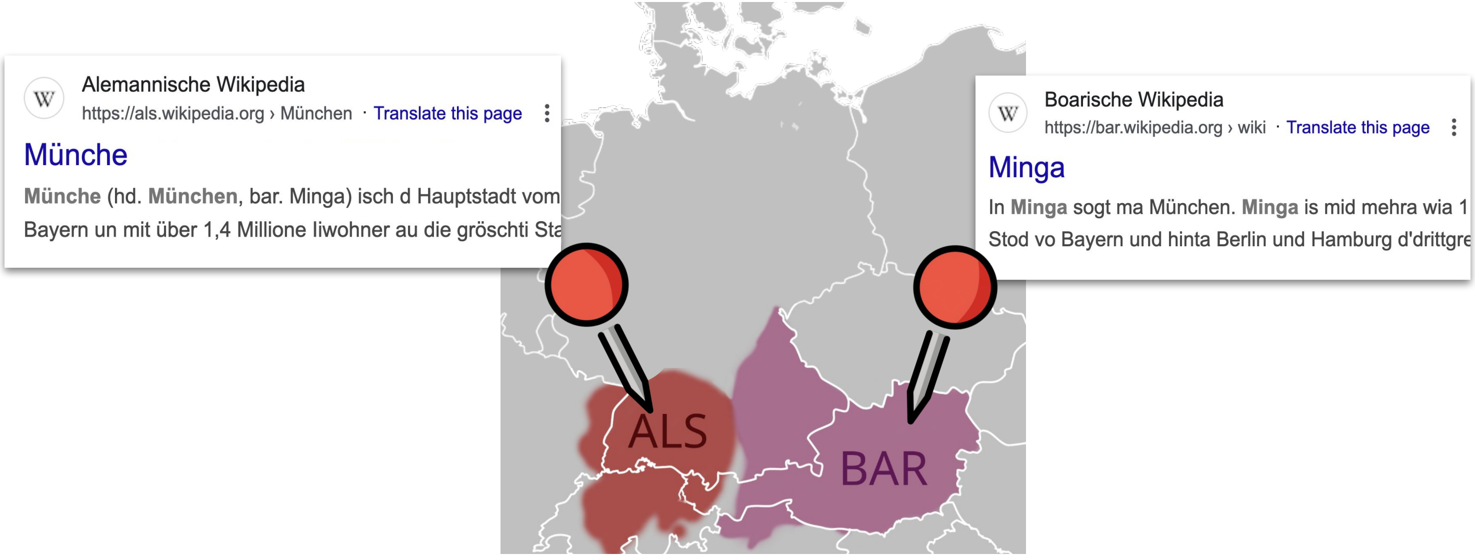

Boarische Wikipedia
https://bar.wikipedia.org › wiki · Translate this page

Minga

In **Minga** sogt ma München. **Minga** is mid mehra wia 1
Stod vo Bayern und hinta Berlin und Hamburg d'drittgre

icons by OpenMoji - CC BY-SA 4.0

# München (“Munich”)



Alemannische Wikipedia
https://als.wikipedia.org › München · Translate this page

Münche

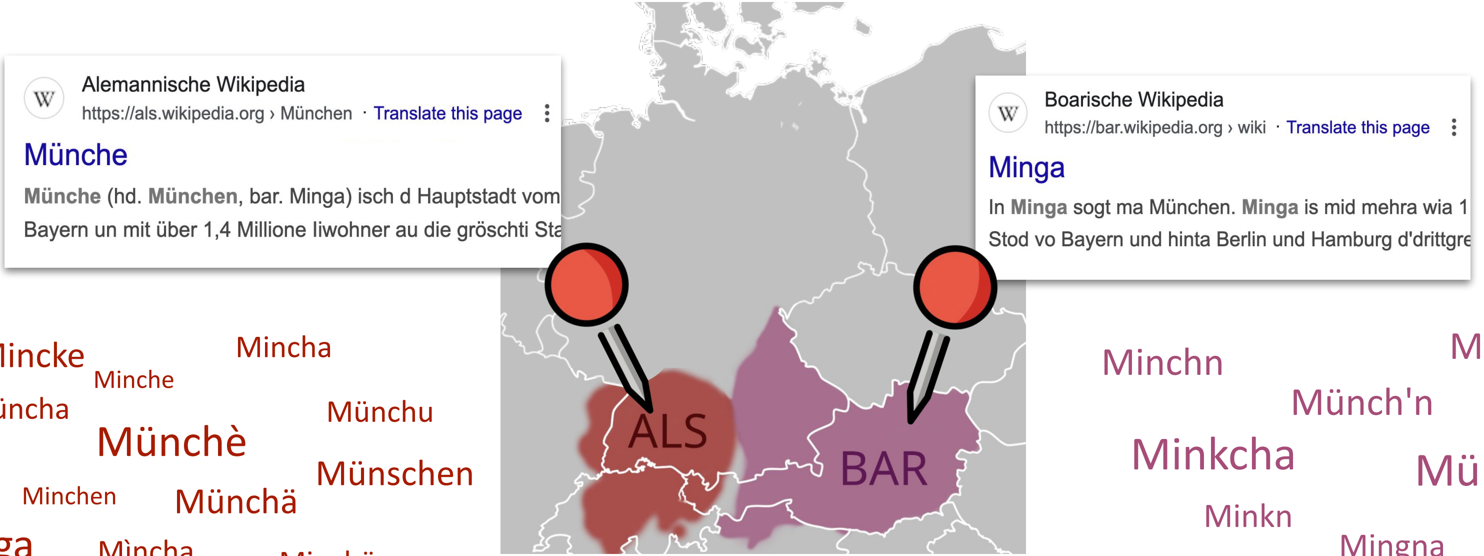
**Münche** (hd. **München**, bar. Minga) isch d Hauptstadt vom
Bayern un mit über 1,4 Millione Iiwohner au die gröschti Sta

Boarische Wikipedia
https://bar.wikipedia.org › wiki · Translate this page

Minga

In **Minga** sogt ma München. **Minga** is mid mehra wia 1
Stod vo Bayern und hinta Berlin und Hamburg d'drittgre

# München (“Munich”)



Alemannische Wikipedia
https://als.wikipedia.org › München · Translate this page

Münche

**Münche** (hd. **München**, bar. Minga) isch d Hauptstadt vom
Bayern un mit über 1,4 Millione Iiwohner au die gröschti Sta

Boarische Wikipedia
https://bar.wikipedia.org › wiki · Translate this page

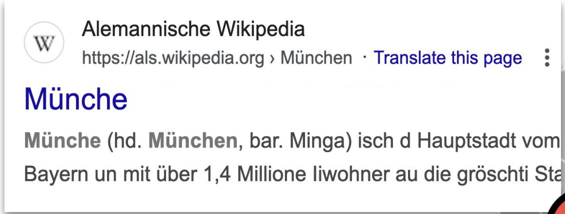
Minga

In **Minga** sogt ma München. **Minga** is mid mehra wia 1
Stod vo Bayern und hinta Berlin und Hamburg d'drittgre

Mincke Mincha Minche Müncha Münchu Minke Münchè Münschen Minchen Münchä Minga Mìncha Minchä

Minchn Minkhn Münch'n Minkcha Münchn Minkn Mingna

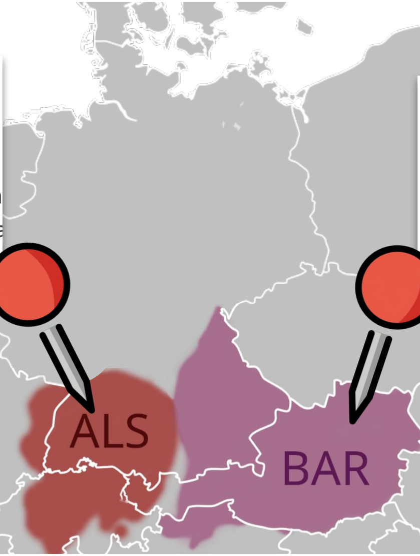
High Lexical variation due to regional word choices and different pronunciations.

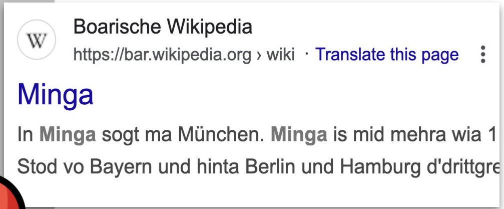


Alemannische Wikipedia
https://als.wikipedia.org › München · Translate this page

Münche

**Münche** (hd. **München**, bar. Minga) isch d Hauptstadt vom
Bayern un mit über 1,4 Millione Iiwohner au die gröschti Sta





Boarische Wikipedia
https://bar.wikipedia.org › wiki · Translate this page

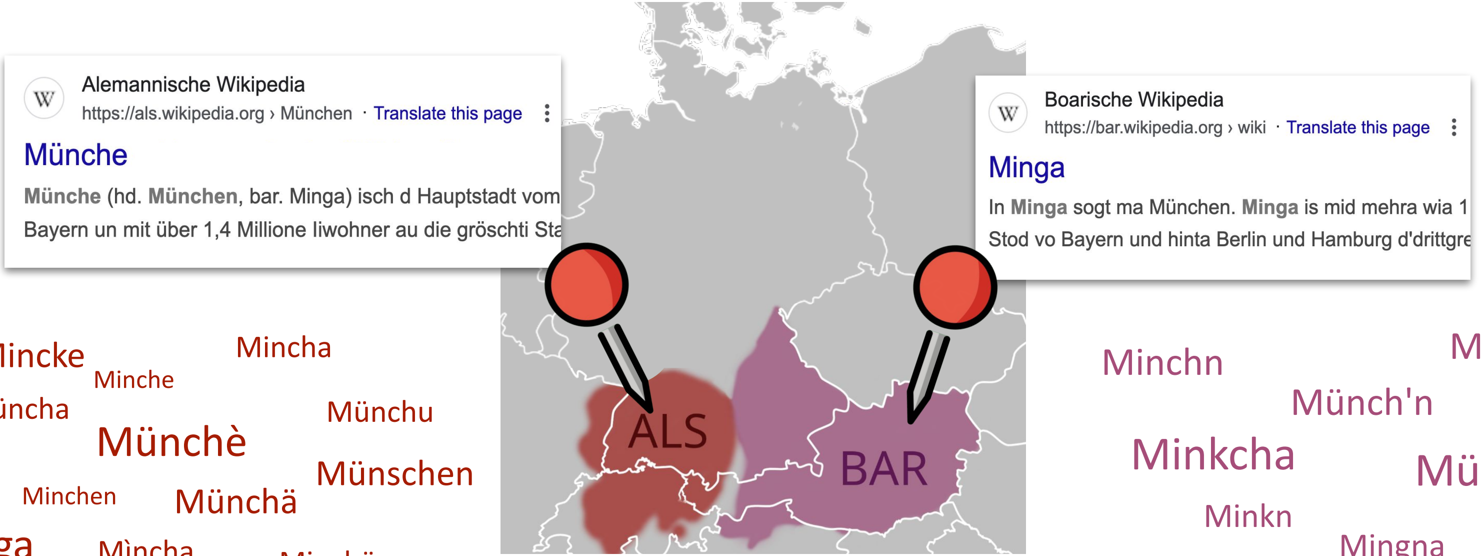
Minga

In **Minga** sogt ma München. **Minga** is mid mehra wia 1
Stod vo Bayern und hinta Berlin und Hamburg d'drittgre

Mincke Minche Mincha Müncha Münchè Münchu Minke Münschen Minchen Münchä Minga Mìncha Minchä

Minchn Minkhn Münch'n Minkcha Münchn Minkn Mingna

High Lexical variation due to regional word choices and different pronunciations.



Alemannische Wikipedia
https://als.wikipedia.org › München · Translate this page

Münche

**Münche** (hd. **München**, bar. Minga) isch d Hauptstadt vom Bayern un mit über 1,4 Millione Iiwohner au die gröschti Sta

Boarische Wikipedia
https://bar.wikipedia.org › wiki · Translate this page

Minga

In **Minga** sogt ma München. **Minga** is mid mehra wia 1 Stod vo Bayern und hinta Berlin und Hamburg d'drittgre

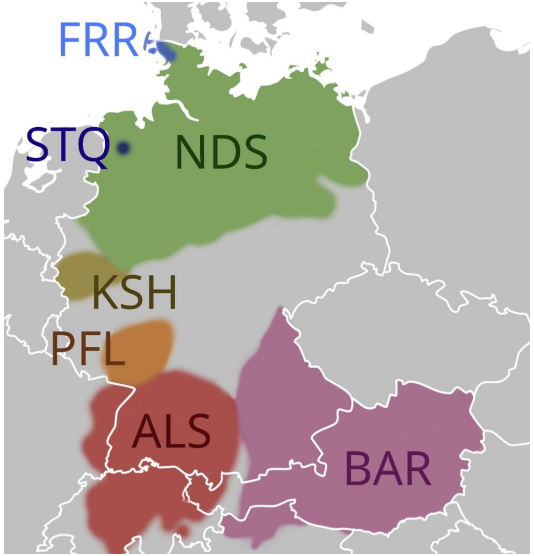
Mincke
Mincha
Minche
Müncha
Münchu
Münchè
Minke
Münschen
Minchen
Münchä
Minga
Mìncha
Minchä

Minkhn
Minchn
Münch'n
Minkcha
Münchn
Minkn
Mingna

Lexical retrieval falls short: Normalizers do not exists for most dialects.

icons by OpenMoji - CC BY-SA 4.0

High Lexical variation due to regional word choices and different pronunciations.



Low German (**nds**)

Alemannic (**als**)

Bavarian (**bar**)

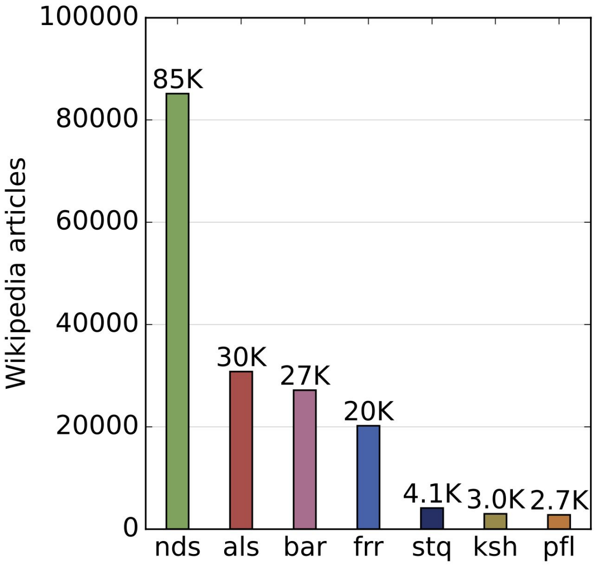
North Frisian (**frr**)

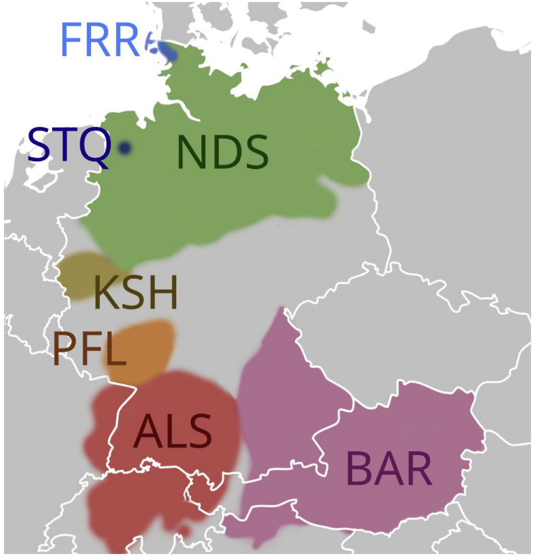
Saterfrisian (**stq**)

Ripuarian (**ksh**)

Rhine Franconian (**pfl**)

High Lexical variation due to regional word choices and different pronunciations.





Low German (**nds**)

Alemannic (**als**)

Bavarian (**bar**)

North Frisian (**frr**)

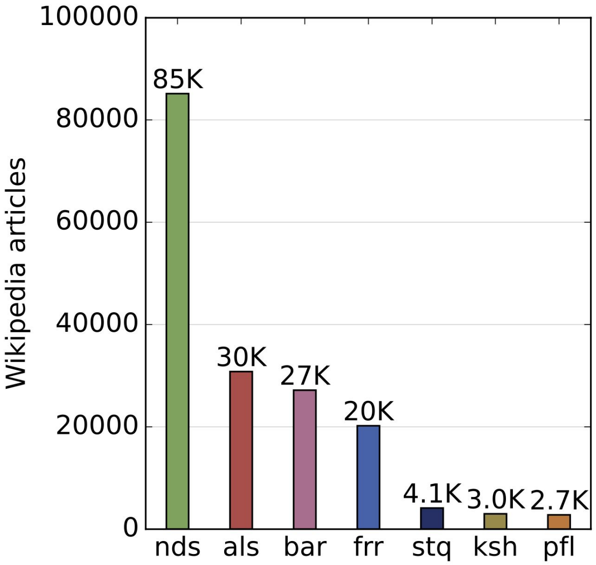
Saterfrisian (**stq**)

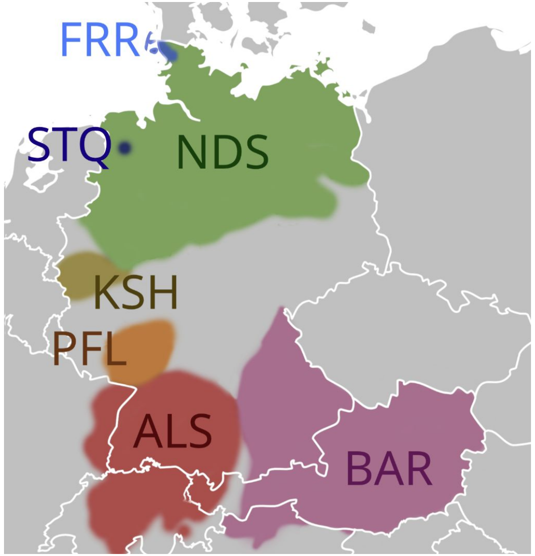
Ripuarian (**ksh**)

Rhine Franconian (**pfl**)

Standard German: 2.9M Wiki articles

High Lexical variation due to regional word choices and different pronunciations.





Low German (**nds**)

Alemannic (**als**)

Bavarian (**bar**)

North Frisian (**frr**)

Saterfrisian (**stq**)

Ripuarian (**ksh**)

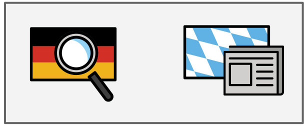
Rhine Franconian (**pfl**)

Low-resource: Very Limited resources data to train neural retrieval models.

# Contribution

- New task: Cross-dialect information retrieval
- New dataset: WikiDIR
- Dialect variation dictionaries
- Evaluation of IR models on WikiDIR

**Example**



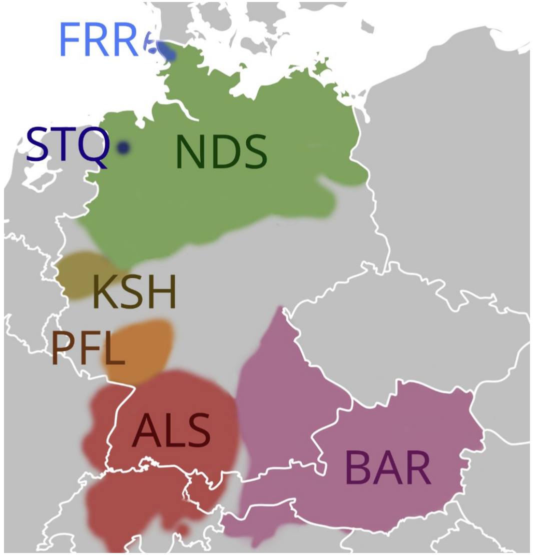


Image based on a map of Europe by Marian Sigler, CC BY-SA 3.0.

# Agenda

1. **Motivation**

2. WikiDIR dataset

3. Dialect dictionaries

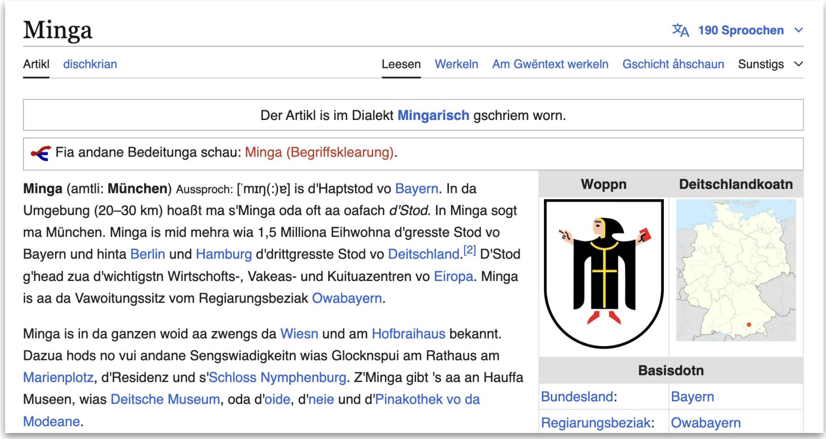
4. Models

5. Results

# Agenda

1. Motivation
2. **WikiDIR dataset**
3. Dialect dictionaries
4. Models
5. Results

# Dataset Pipeline



Minga

190 Sproochen

Artikl dischkrian

Leesen Werkeln Am Gwëntext werkeln Gschicht åhschaun Sunstigs

Der Artikl is im Dialekt **Mingarisch** gschriem worn.

Fia andane Bedeitunga schau: Minga (Begriffsklearung).

**Minga** (amtli: **München**) Aussproch: [ˈmɪŋ(ː)ɐ] is d'Haptstod vo Bayern. In da Umgebung (20–30 km) hoaßt ma s'Minga oda oft aa oafach *d'Stod*. In Minga sogt ma München. Minga is mid mehra wia 1,5 Milliona Eihwohna d'gresste Stod vo Bayern und hinta Berlin und Hamburg d'drittgresste Stod vo Deitschland.[2] D'Stod g'head zua d'wichtigstn Wirtschofts-, Vakeas- und Kuituazentren vo Eiropa. Minga is aa da Vawoitungssitz vom Regiarungsbeziak Owabayern.

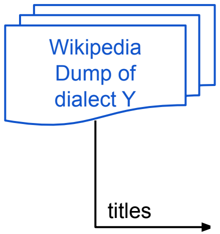
Minga is in da ganzen woid aa zwengs da Wiesn und am Hofbraihaus bekannt. Dazua hods no vui andane Sengswiadigkeitn wias Glocknspui am Rathaus am Marienplotz, d'Residenz und s'Schloss Nymphenburg. Z'Minga gibt 's aa an Hauffa Museen, wias Deitsche Museum, oda d'oide, d'neie und d'Pinakothek vo da Modeane.

| Woppn | Deitschlandkoatn |
|---|---|
| | |

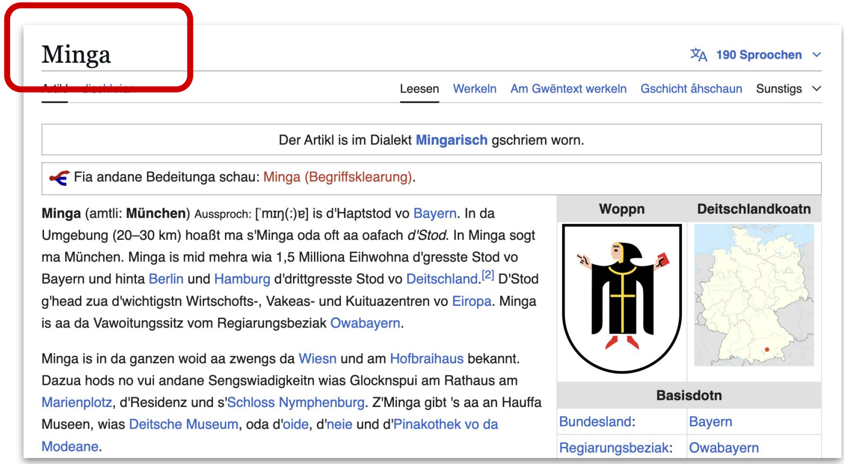
| Basisdotn | |
|---|---|
| Bundesland: | Bayern |
| Regiarungsbeziak: | Owabayern |

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Query $q_i$



Minga

Artikl dischkrian

190 Sproochen

Leesen Werkeln Am Gwëntext werkeln Gschicht åhschaun Sunstigs

Der Artikl is im Dialekt **Mingarisch** gschriem worn.

Fia andane Bedeitunga schau: Minga (Begriffsklearung).

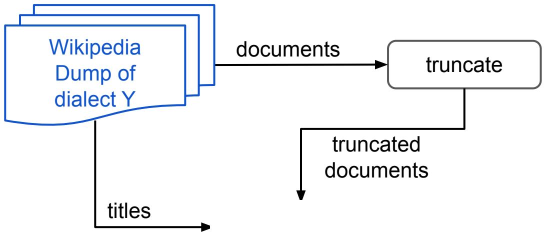
**Minga** (amtli: **München**) Aussproch: [ˈmɪŋ(ː)ɐ] is d'Haptstod vo Bayern. In da Umgebung (20–30 km) hoaßt ma s'Minga oda oft aa oafach *d'Stod*. In Minga sogt ma München. Minga is mid mehra wia 1,5 Milliona Eihwohna d'gresste Stod vo Bayern und hinta Berlin und Hamburg d'drittgresste Stod vo Deitschland.[2] D'Stod g'head zua d'wichtigstn Wirtschofts-, Vakeas- und Kuituazentren vo Eiropa. Minga is aa da Vawoitungssitz vom Regiarungsbeziak Owabayern.

Minga is in da ganzen woid aa zwengs da Wiesn und am Hofbraihaus bekannt. Dazua hods no vui andane Sengswiadigkeitn wias Glocknspui am Rathaus am Marienplotz, d'Residenz und s'Schloss Nymphenburg. Z'Minga gibt 's aa an Hauffa Museen, wias Deitsche Museum, oda d'oide, d'neie und d'Pinakothek vo da Modeane.

| Woppn | Deitschlandkoatn |
|---|---|
| Basisdotn | |
| Bundesland: | Bayern |
| Regiarungsbeziak: | Owabayern |

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Wikipedia Dump of dialect Y

documents

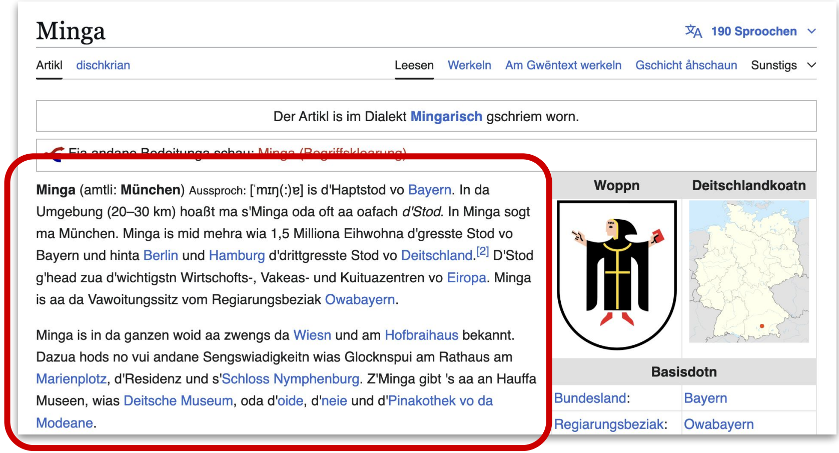
truncate

truncated documents

titles

Query $q_i$

Corpus $\mathcal{D}$



Minga

Der Artikl is im Dialekt **Mingarisch** gschriem worn.

Eia andane Bedeitunga schau: Minga (Begriffsklearung)

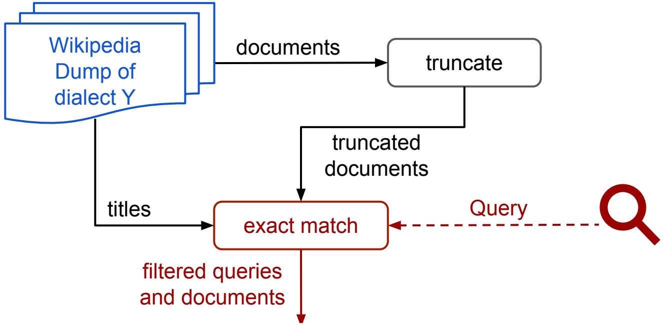
**Minga** (amtli: **München**) Aussproch: [ˈmɪŋ(:)ɐ] is d'Haptstod vo Bayern. In da Umgebung (20–30 km) hoaßt ma s'Minga oda oft aa oafach *d'Stod*. In Minga sogt ma München. Minga is mid mehra wia 1,5 Milliona Eihwohna d'gresste Stod vo Bayern und hinta Berlin und Hamburg d'drittgresste Stod vo Deitschland.[2] D'Stod g'head zua d'wichtigstn Wirtschofts-, Vakeas- und Kuituazentren vo Eiropa. Minga is aa da Vawoitungssitz vom Regiarungsbeziak Owabayern.

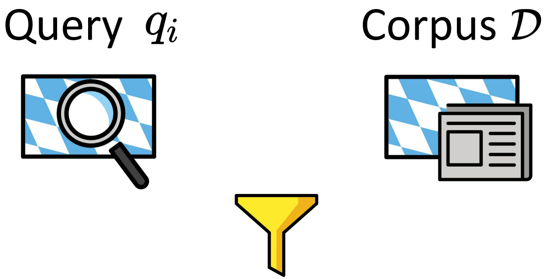
Minga is in da ganzen woid aa zwengs da Wiesn und am Hofbraihaus bekannt. Dazua hods no vui andane Sengswiadigkeitn wias Glocknspui am Rathaus am Marienplotz, d'Residenz und s'Schloss Nymphenburg. Z'Minga gibt 's aa an Hauffa Museen, wias Deitsche Museum, oda d'oide, d'neie und d'Pinakothek vo da Modeane.

| Woppn | Deitschlandkoatn |
|---|---|
| Basisdotn | |
| Bundesland: | Bayern |
| Regiarungsbeziak: | Owabayern |

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline

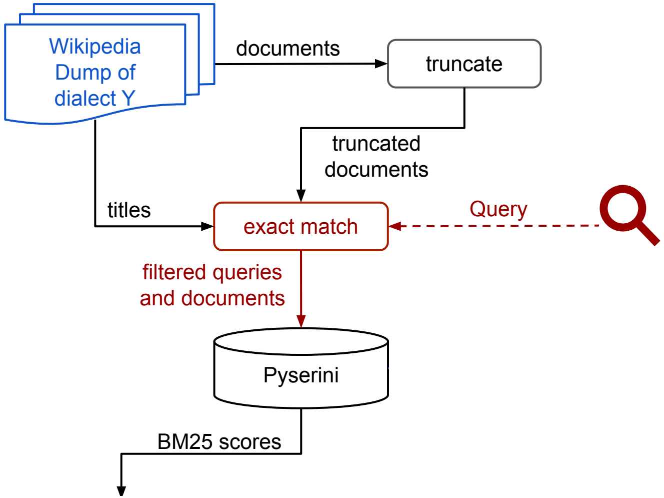




$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Wikipedia Dump of dialect Y

documents

truncate

truncated documents

titles

exact match

Query

filtered queries and documents

Pyserini

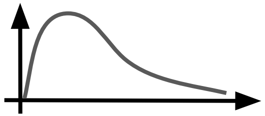
BM25 scores

Query $q_i$

Corpus $\mathcal{D}$

$$\mathcal{D}^{q_i}_{\text{rel}} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

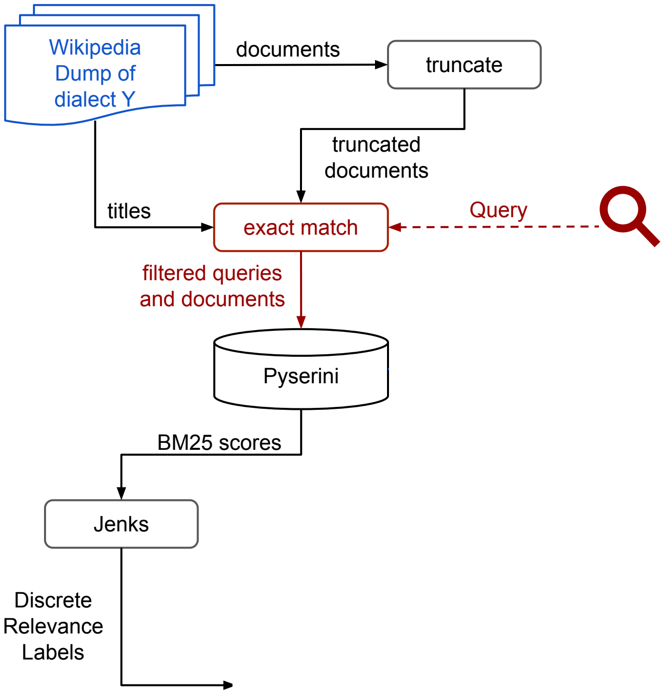
Lexical Similarity Scores



all (q,d)-pairs

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Wikipedia Dump of dialect Y

documents

truncate

truncated documents

titles

exact match

Query

filtered queries and documents

Pyserini

BM25 scores

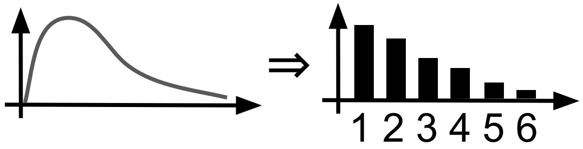
Jenks

Discrete Relevance Labels

Query $q_i$ Corpus $\mathcal{D}$

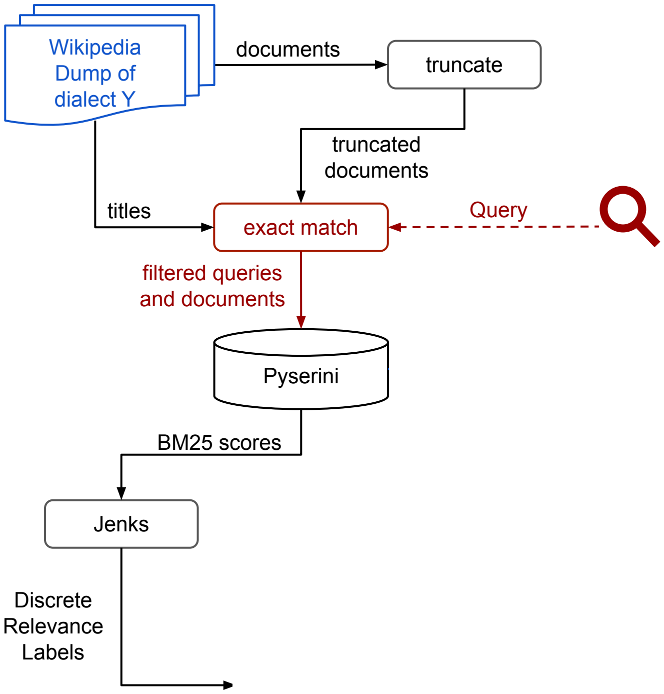
$$\mathcal{D}^{q_i}_{\text{rel}} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

Monolingual Relevance Labels



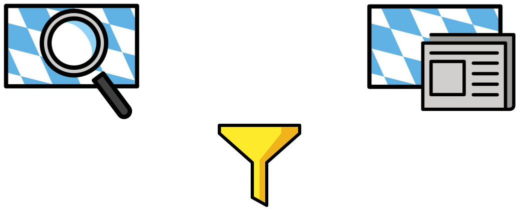
based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline


Wikipedia Dump of dialect Y
documents
truncate
truncated documents
titles
exact match
Query
filtered queries and documents
Pyserini
BM25 scores
Jenks
Discrete Relevance Labels

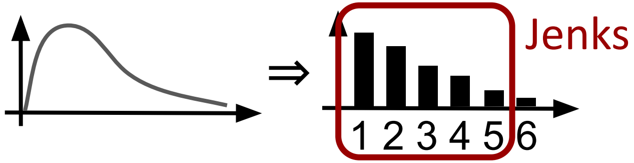
Query $q_i$

Corpus $\mathcal{D}$



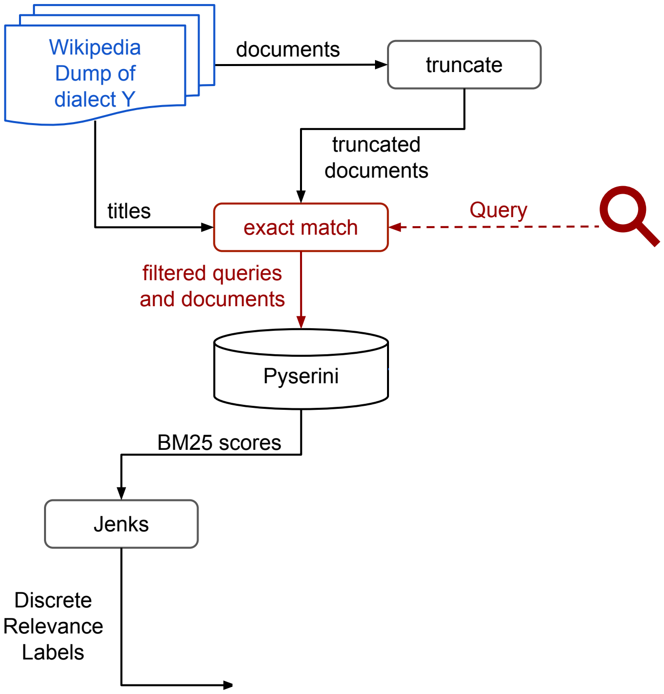
$$\mathcal{D}^{q_i}_{\text{rel}} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

Monolingual Relevance Labels



based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Wikipedia Dump of dialect Y

documents

truncate

truncated documents

titles

exact match

Query

filtered queries and documents

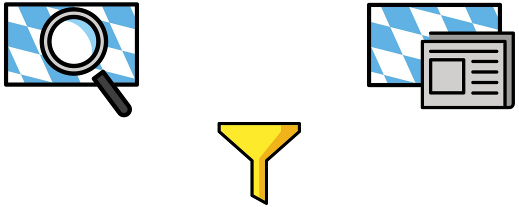
Pyserini

BM25 scores

Jenks

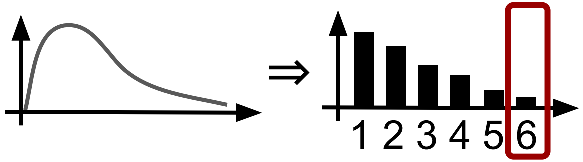
Discrete Relevance Labels

Query $q_i$ Corpus $\mathcal{D}$



$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

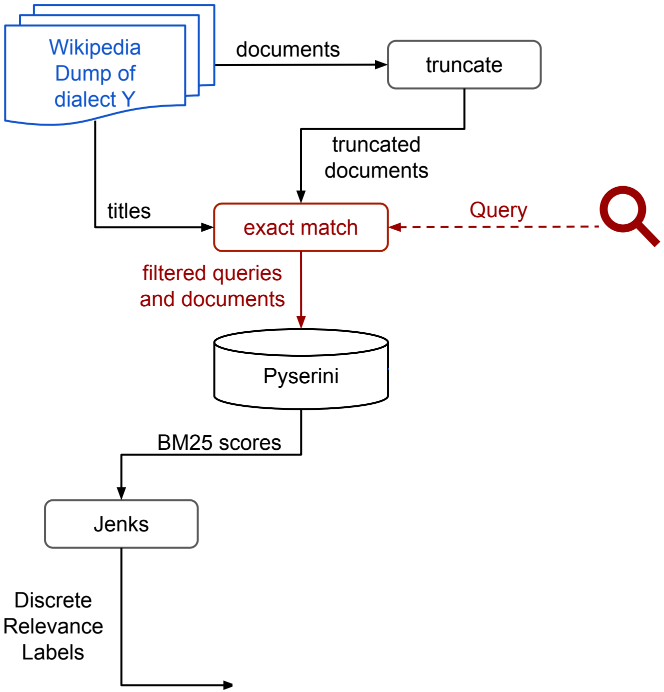
Monolingual Relevance Labels



same article

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Wikipedia Dump of dialect Y

documents

truncate

truncated documents

titles

exact match

Query

filtered queries and documents

Pyserini

BM25 scores

Jenks

Discrete Relevance Labels

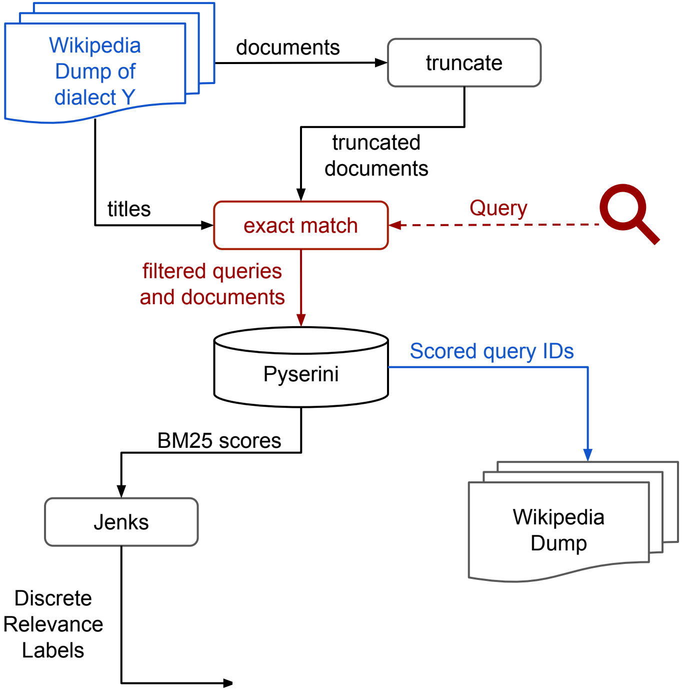
Query $q_i$

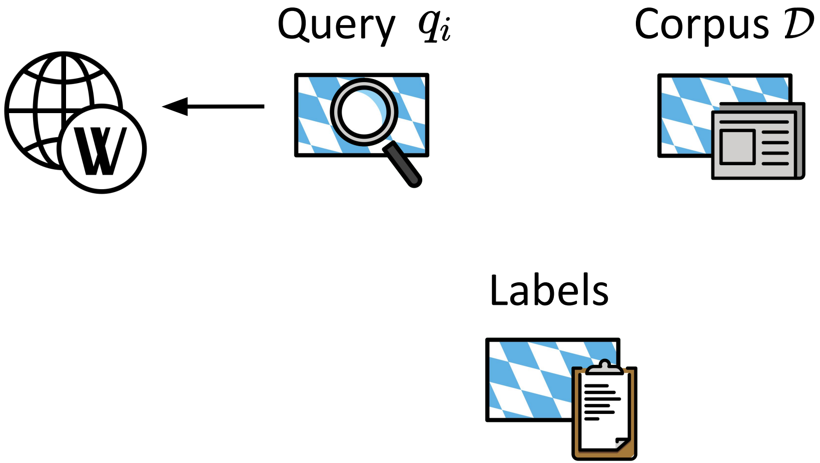
Corpus $\mathcal{D}$

Labels

based on CLIRMatrix (Sun and Duh, 2018)

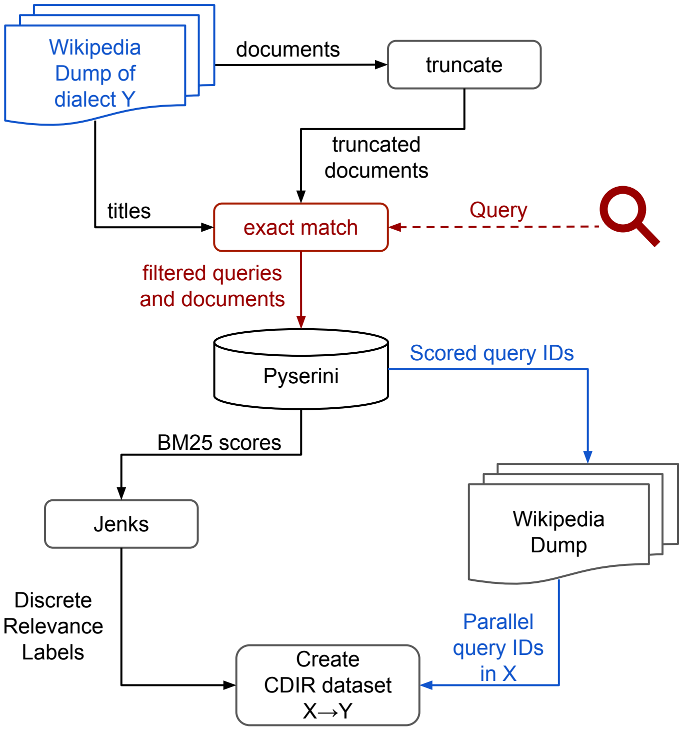
# Dataset Pipeline





based on CLIRMatrix (Sun and Duh, 2018), wiki icon by Putu Kharismayad

# Dataset Pipeline



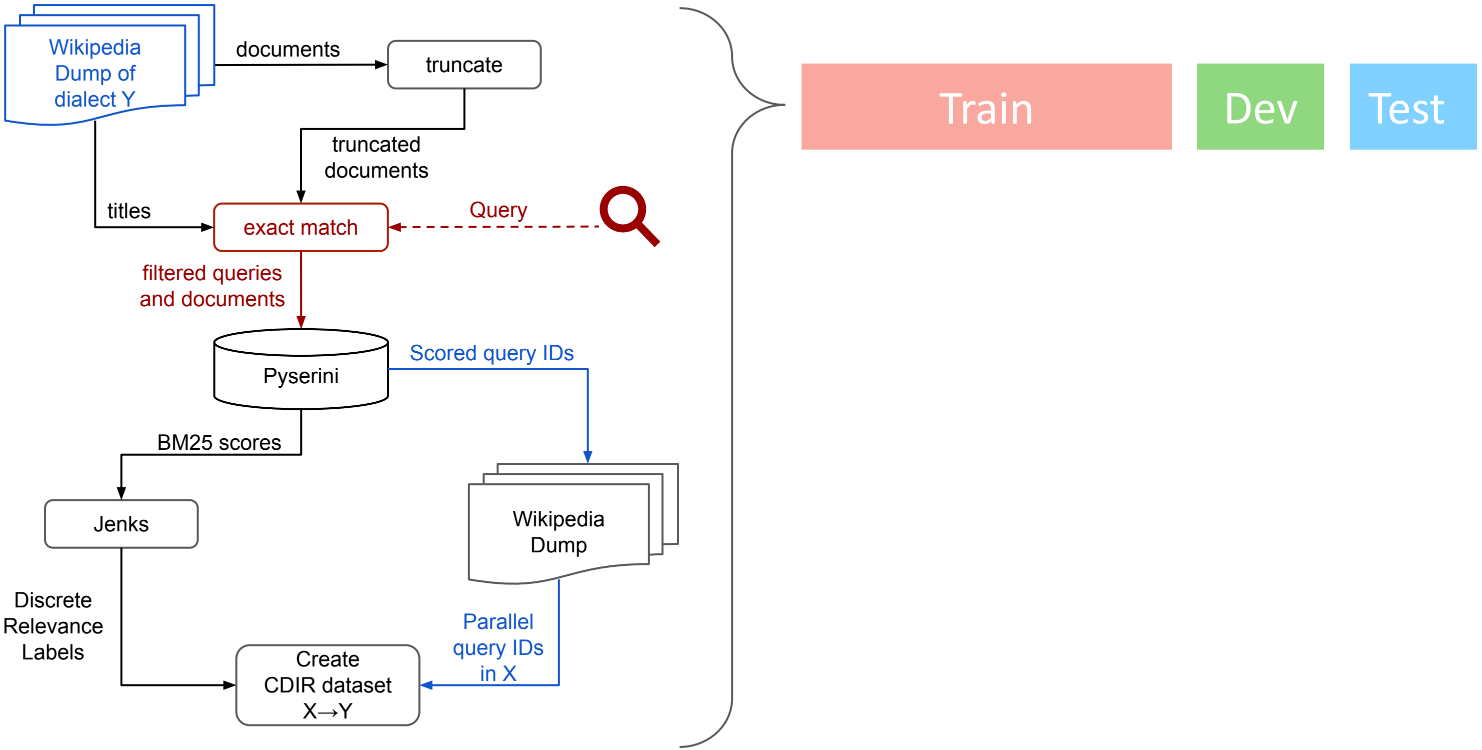
Query $q_i$

Corpus $\mathcal{D}$

Labels

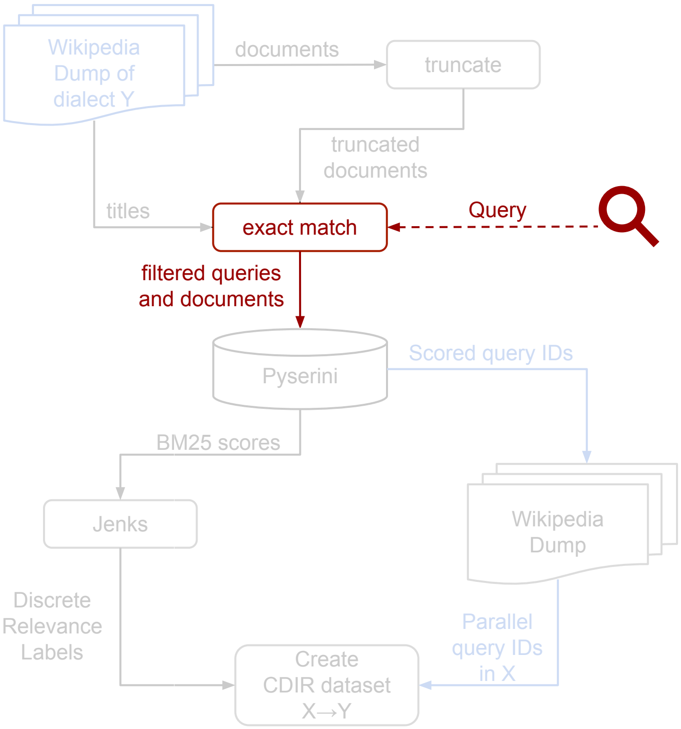
based on CLIRMatrix (Sun and Duh, 2018), wiki icon by Putu Kharismayad

# Dataset Pipeline



based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



Wikipedia Dump of dialect Y

documents

truncate

truncated documents

titles

exact match

Query

filtered queries and documents

Pyserini

Scored query IDs

BM25 scores

Jenks

Wikipedia Dump

Discrete Relevance Labels

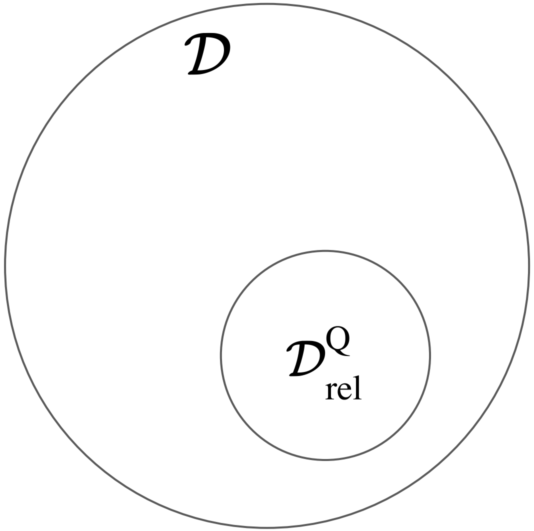
Parallel query IDs in X

Create CDIR dataset X→Y

| Train | Dev | Test |
|---|---|---|

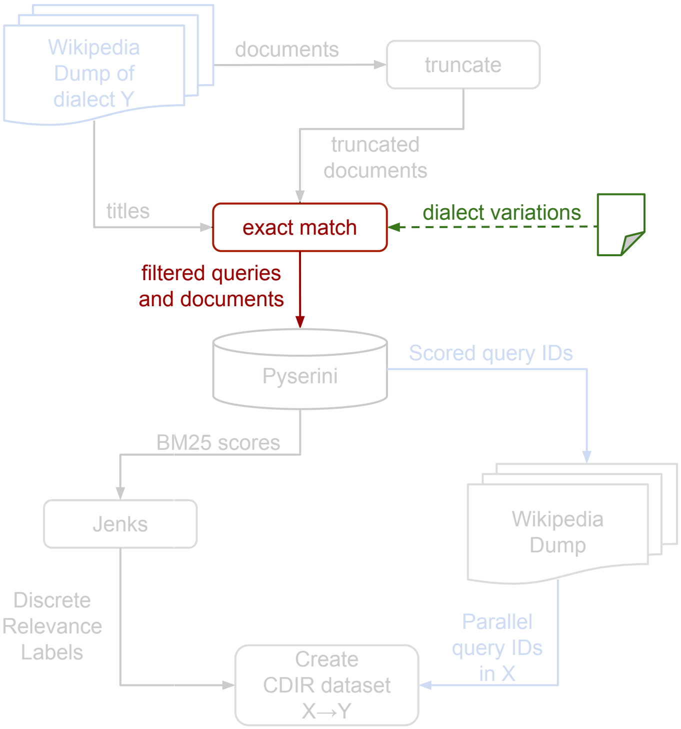
**Set of rel. docs.**

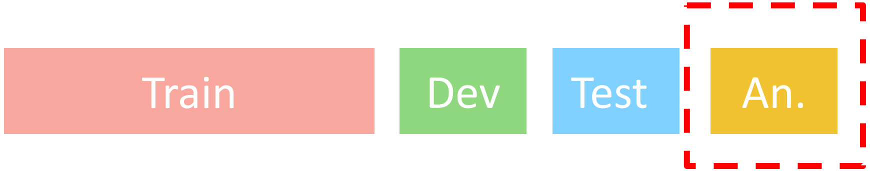
All documents that contain a query.



based on CLIRMatrix (Sun and Duh, 2018)

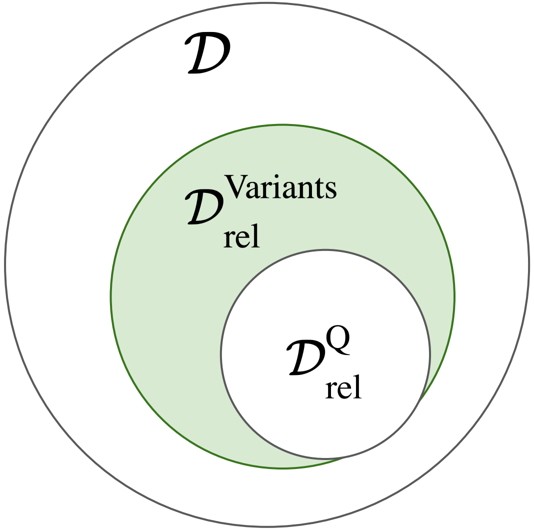
# Dataset Pipeline





## Analysis Split

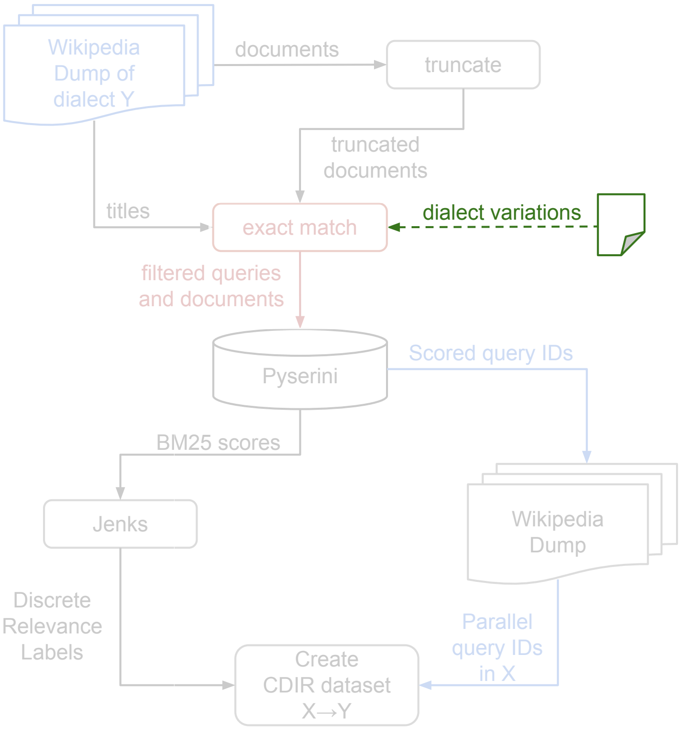
All documents that contain a query or any of its dialect variations.



$\mathcal{D}$
$\mathcal{D}^{\text{Variants}}_{\text{rel}}$
$\mathcal{D}^{\text{Q}}_{\text{rel}}$

based on CLIRMatrix (Sun and Duh, 2018)

# Dataset Pipeline



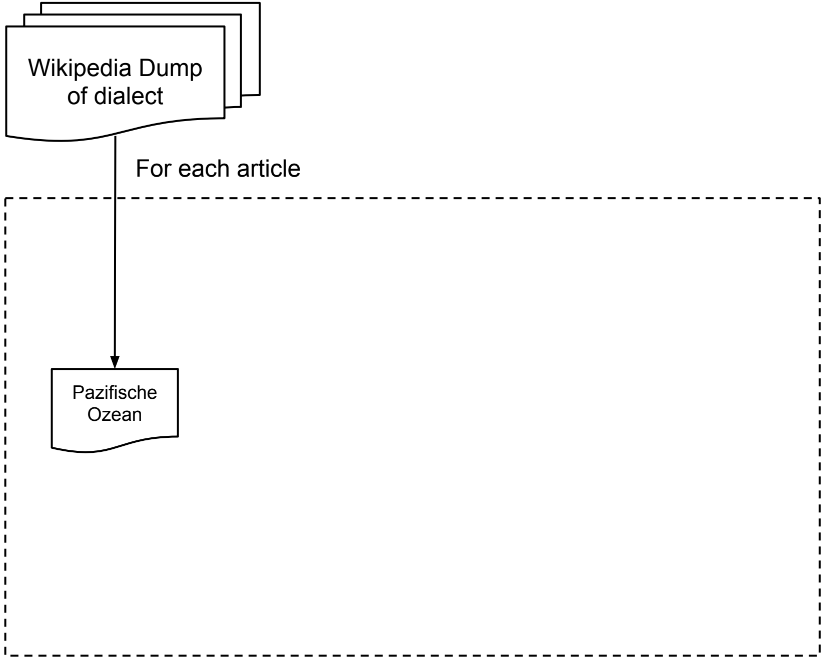
Where do **dialect variations** come from?

based on CLIRMatrix (Sun and Duh, 2018)

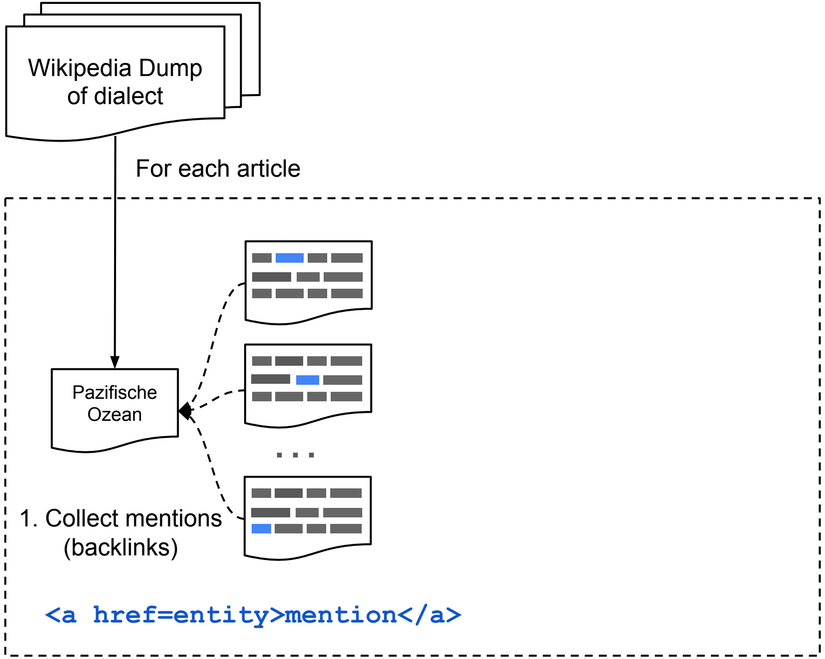
# Agenda

1. Motivation
2. WikiDIR Dataset
3. **Dialect dictionaries**
4. Models
5. Results

# Dialect variation dictionaries

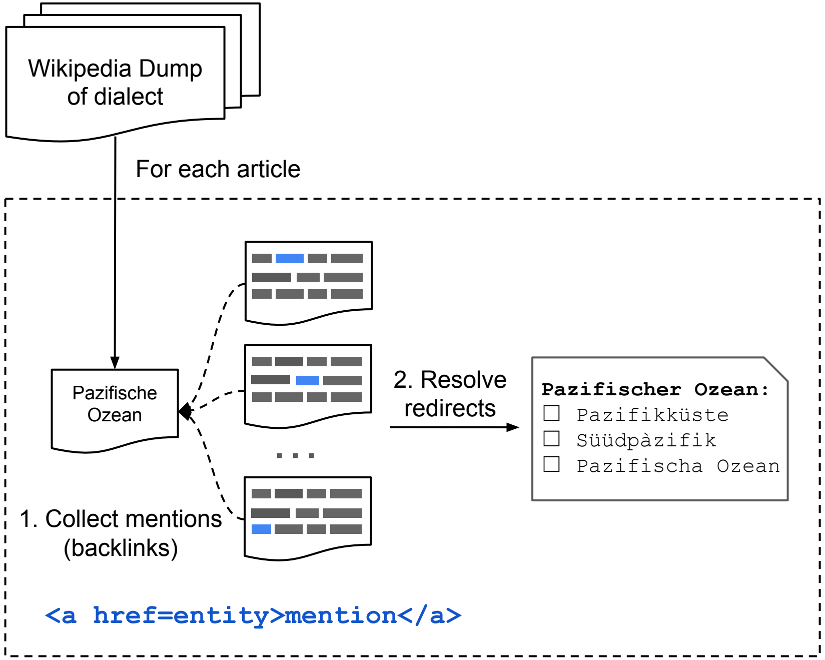


# Dialect variation dictionaries



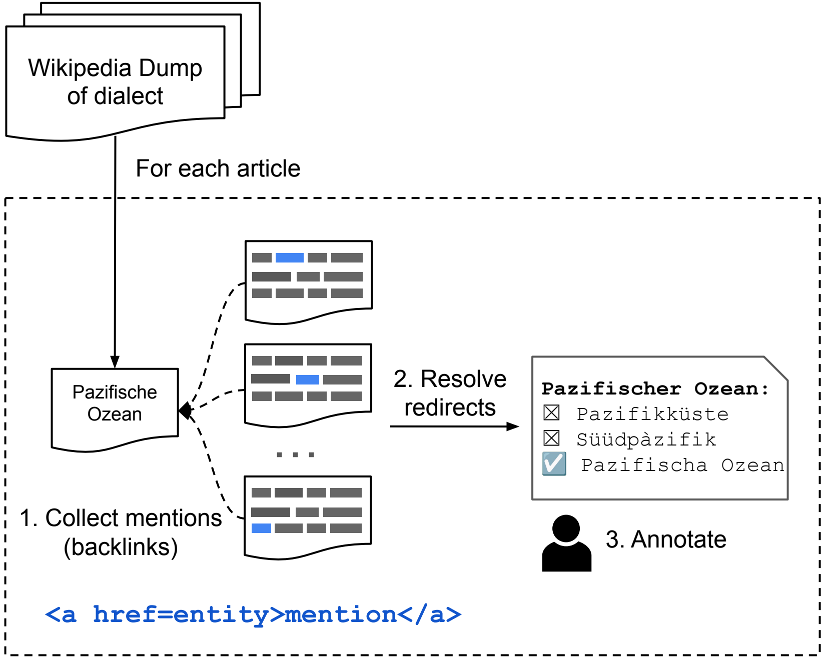
```
<a href=entity>mention</a>
```

# Dialect variation dictionaries



```
<a href=entity>mention</a>
```

# Dialect variation dictionaries



Wikipedia Dump of dialect

For each article

Pazifische Ozean

1. Collect mentions (backlinks)

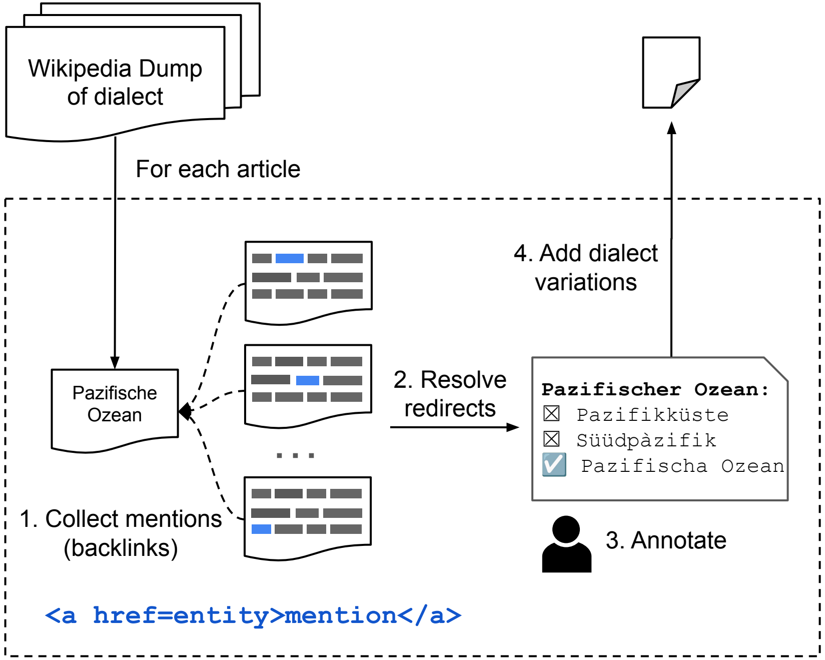
2. Resolve redirects

**Pazifischer Ozean:**
☒ Pazifikküste
☒ Süüdpàzifik
☑ Pazifischa Ozean

3. Annotate

`<a href=entity>mention</a>`

# Dialect variation dictionaries



Wikipedia Dump of dialect

For each article

Pazifische Ozean

1. Collect mentions (backlinks)

2. Resolve redirects

**Pazifischer Ozean:**

☒ Pazifikküste

☒ Süüdpàzifik

☑ Pazifischa Ozean

3. Annotate

4. Add dialect variations

```
<a href=entity>mention</a>
```

# Example Record (Bavarian dictionary)

```
{
  "de_id": "3215",
  "de_title": "München",
  "dial_id": "12259",
  "dial_title": "Minga",
  "variants": ["Münch'n", "Minkcha", "Minkn", "Minchn", "Mingna", "Minkhn", "Münchn"]
}
```

# Agenda

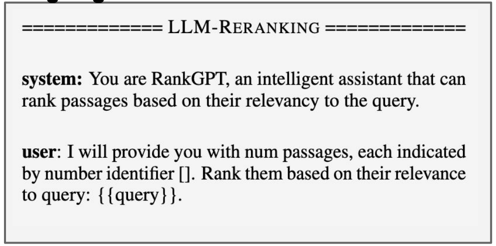
1. Motivation
2. WikiDIR dataset
3. Dialect dictionaries
4. **Models**
5. Results

# Models

**Baseline:** BM25 (Robertson, 1995)

# Models

RankGPT (Llama 3.1)



============= LLM-RERANKING =============

**system:** You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

**user**: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

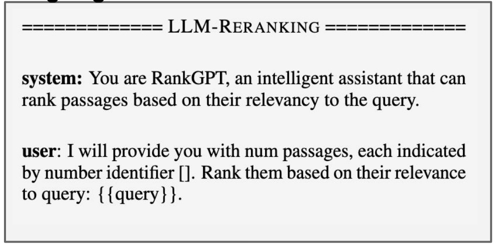
(Sun et al., 2023)

**Baseline:** BM25 (Robertson, 1995)

Llama icon created by Freepik - Flaticon

# Models

RankGPT (Llama 3.1)

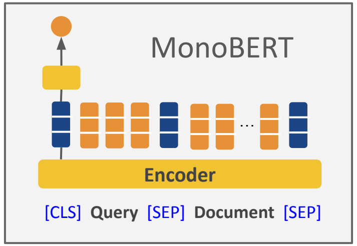


============== LLM-RERANKING ==============

**system:** You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

**user**: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)



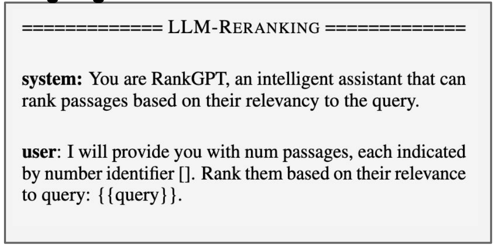
(Nogueira et al., 2019)

**Baseline:** BM25 (Robertson, 1995)

Model diagrams taken from (Litschko, 2024)

# Models

RankGPT (Llama 3.1)

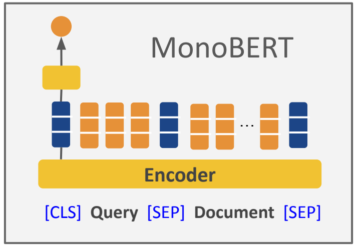


============= LLM-RERANKING =============

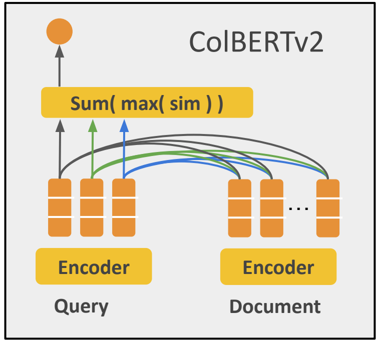
**system:** You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

**user**: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)



(Nogueira et al., 2019)



(Santhanam et al., 2022)

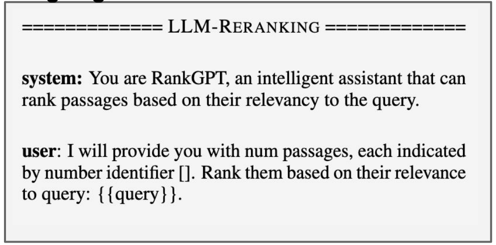
**Baseline:** BM25 (Robertson, 1995)

Model diagrams taken from (Litschko, 2024)

# Models

## Rerank top 100

RankGPT (Llama 3.1)

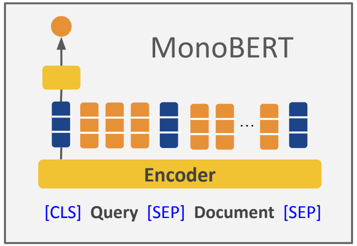


============= LLM-RERANKING =============

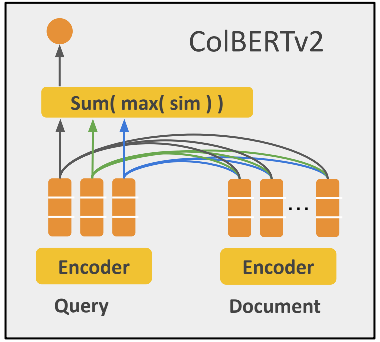
**system:** You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

**user**: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)



(Nogueira et al., 2019)

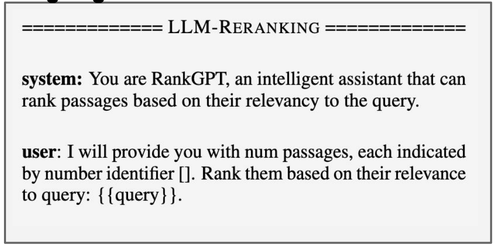


(Santhanam et al., 2022)

**Baseline:** BM25 (Robertson, 1995)

# Models

RankGPT (Llama 3.1)

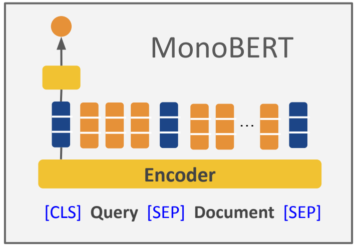


============= LLM-RERANKING =============

**system:** You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

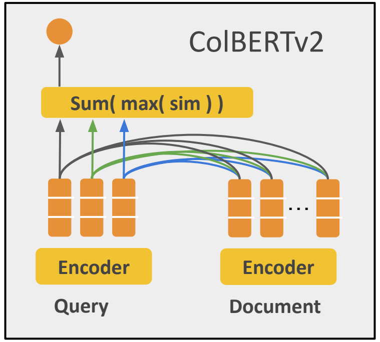
**user**: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)



(Nogueira et al., 2019)

Retrieval

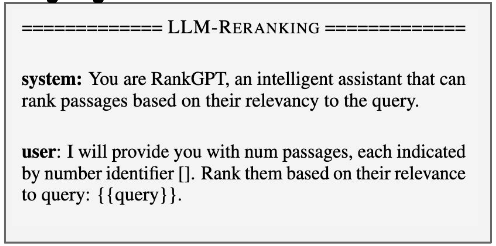


(Santhanam et al., 2022)

**Baseline:** BM25 (Robertson, 1995)

# Models

RankGPT (Llama 3.1)

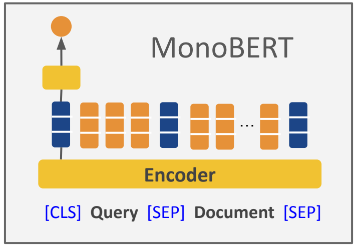


============= LLM-RERANKING =============

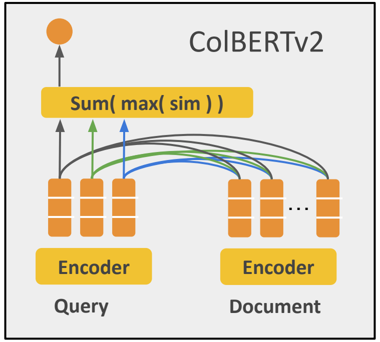
**system:** You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

**user**: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)

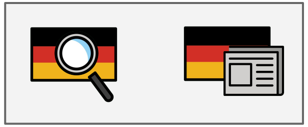


(Nogueira et al., 2019)

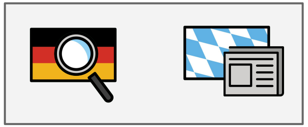


(Santhanam et al., 2022)

Zero-shot Transfer



Fine-tuning



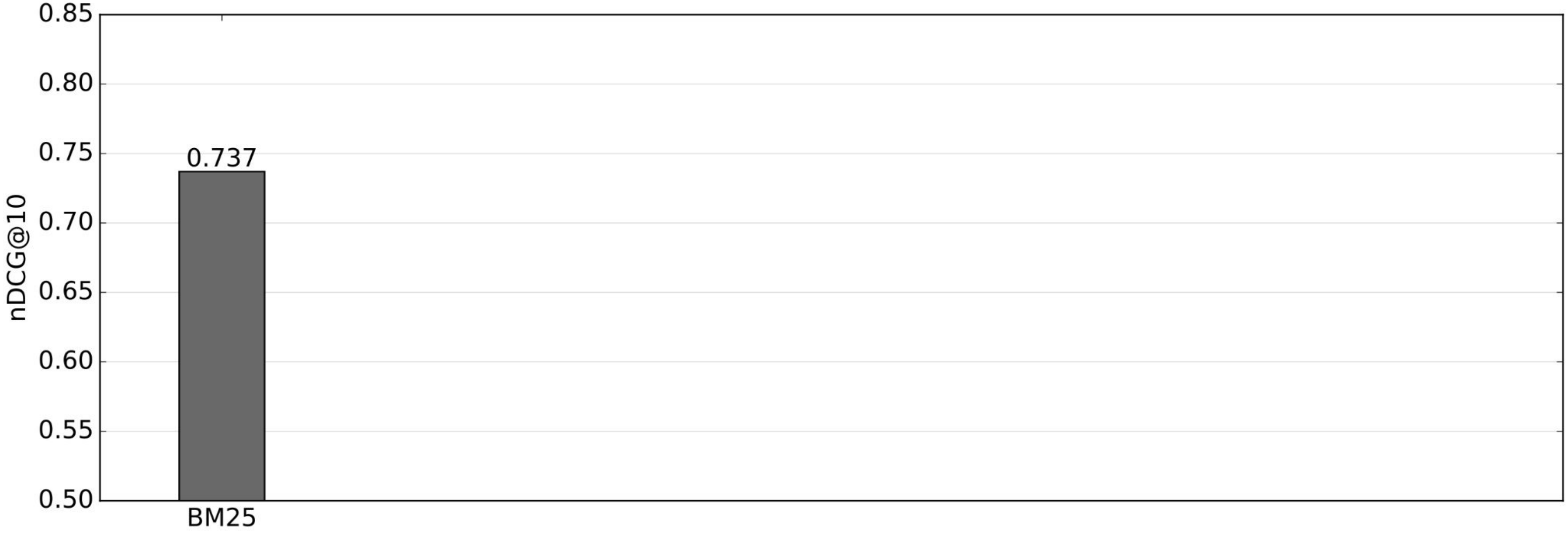
**Baseline:** BM25 (Robertson, 1995)

# Agenda

1. Motivation
2. WikiDIR dataset
3. Dialect dictionaries
4. Models
5. **Results**

Test

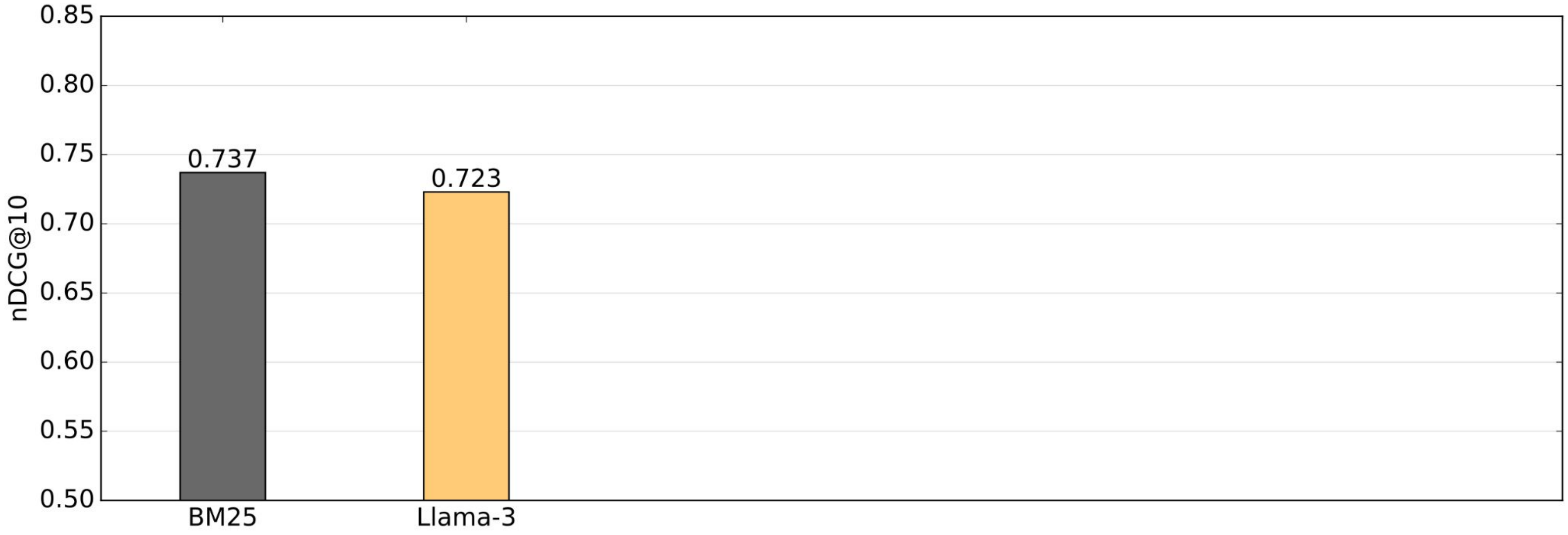
# Main results



*average over 7 dialects

Test

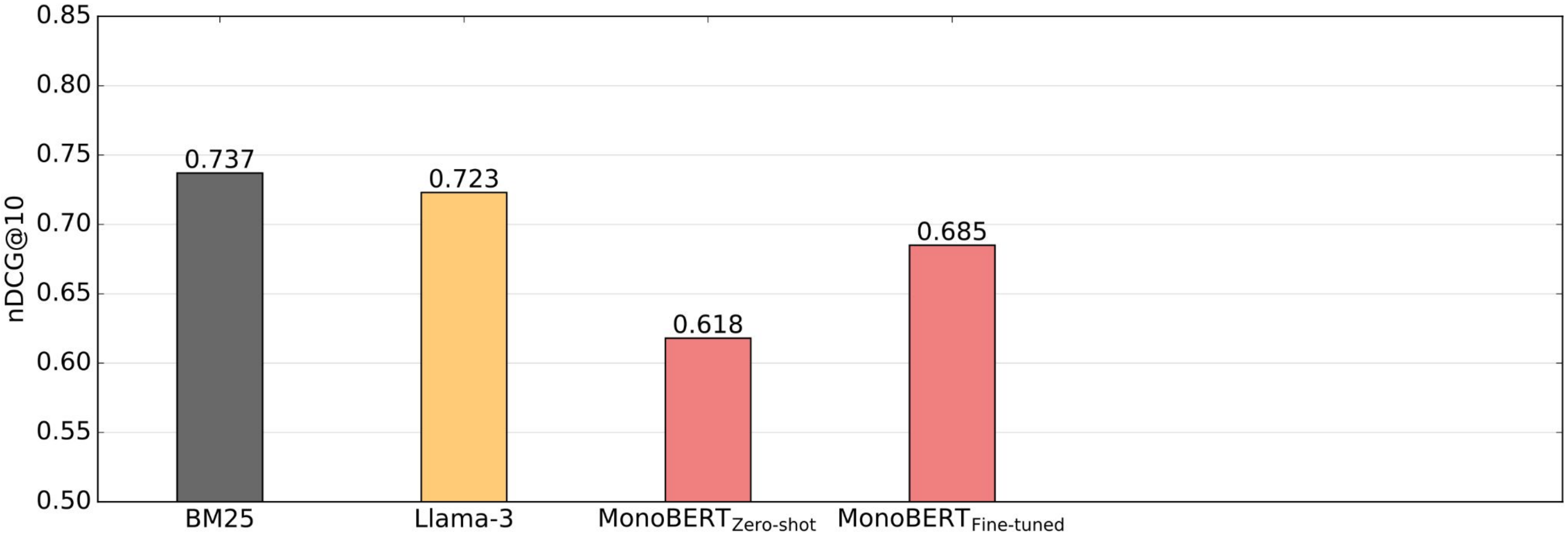
# Main results



*average over 7 dialects

Test

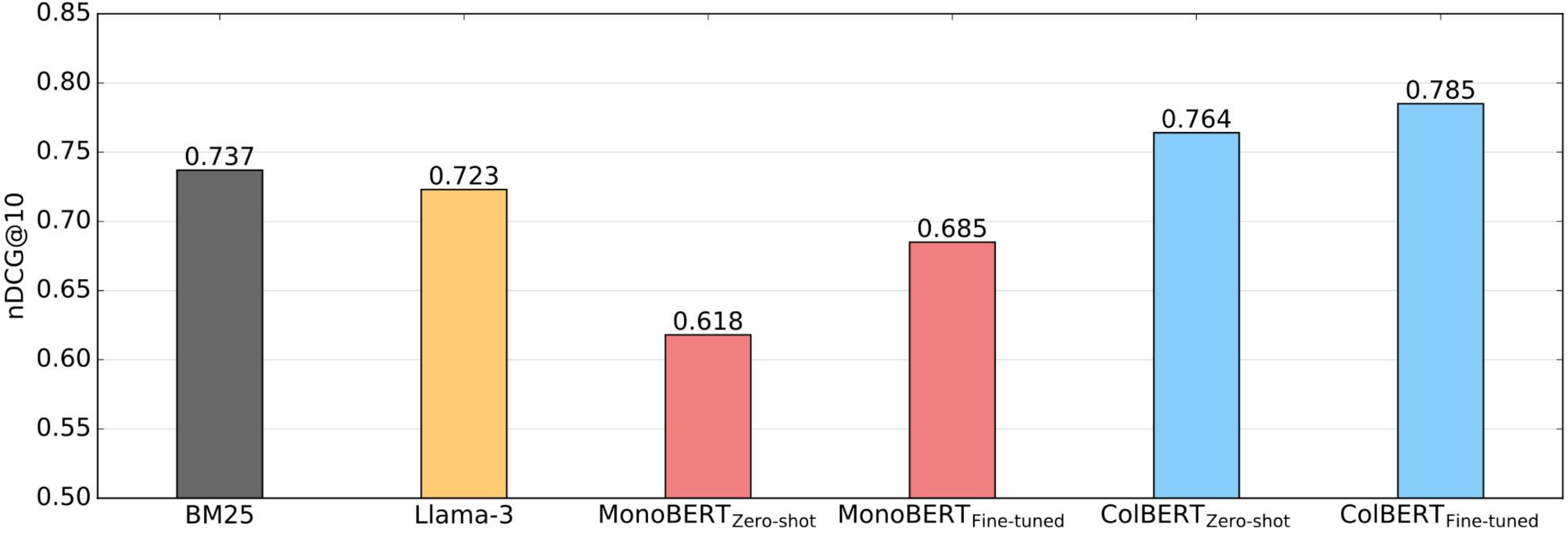
# Main results



*average over 7 dialects

Test

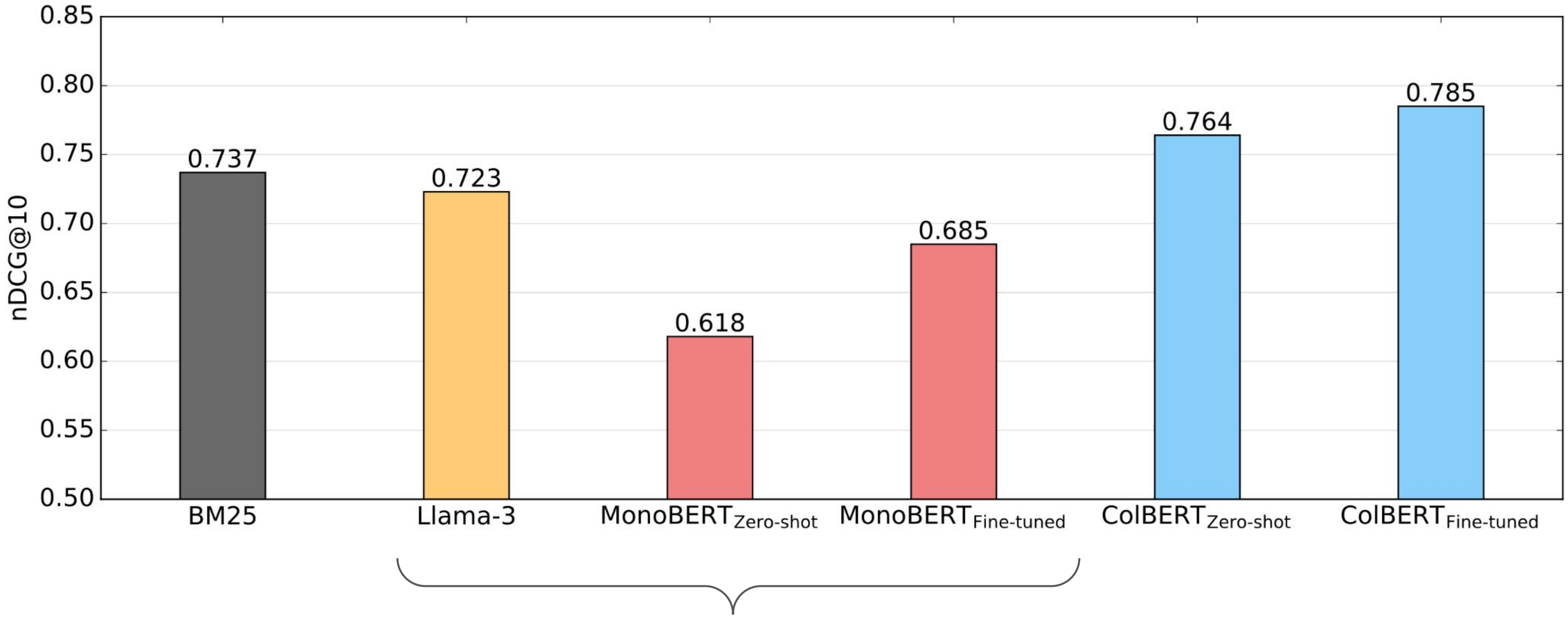
# Main results



*average over 7 dialects

Test

# Main results

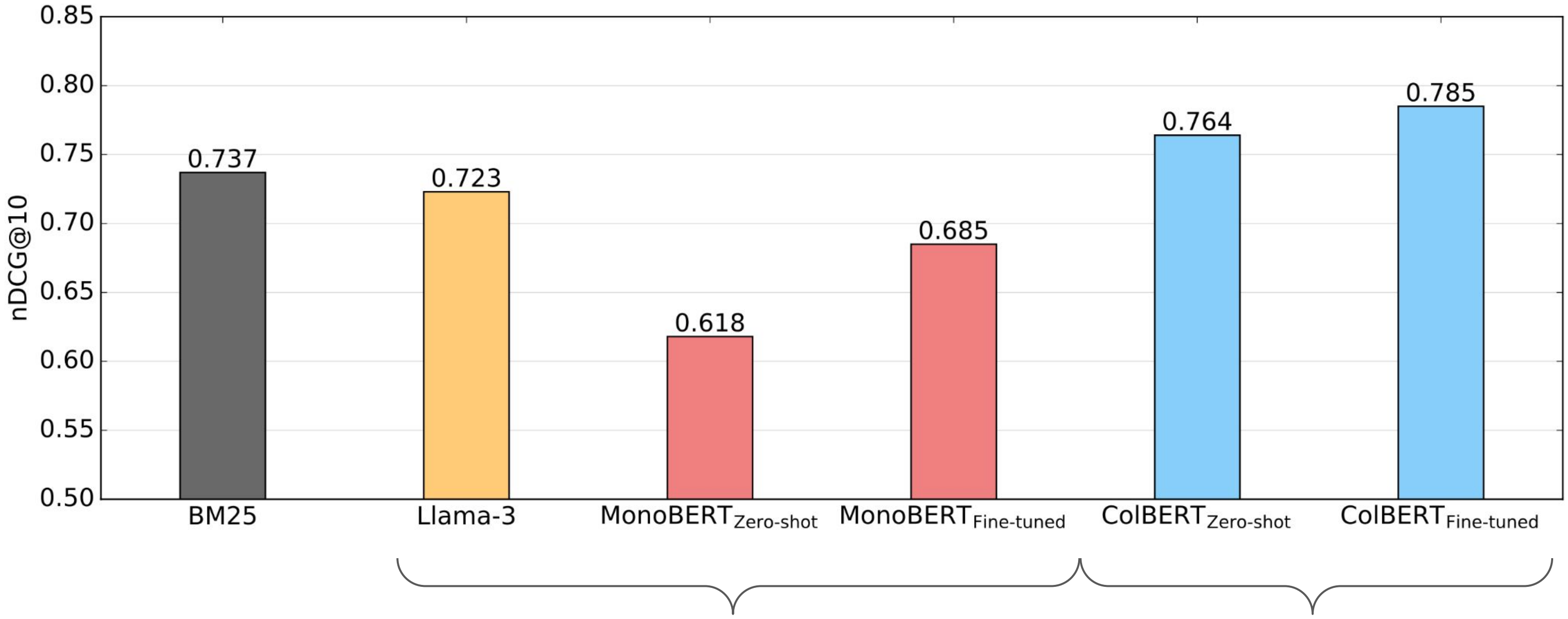


Reranking

*average over 7 dialects

Test

# Main results



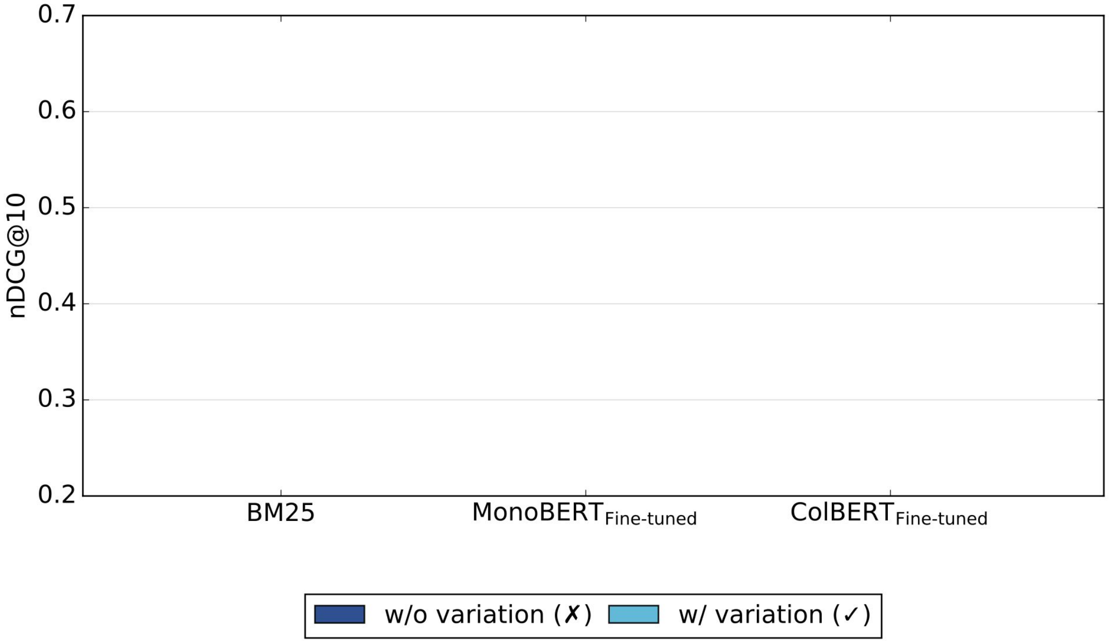
Reranking

Retrieval

*average over 7 dialects

An.

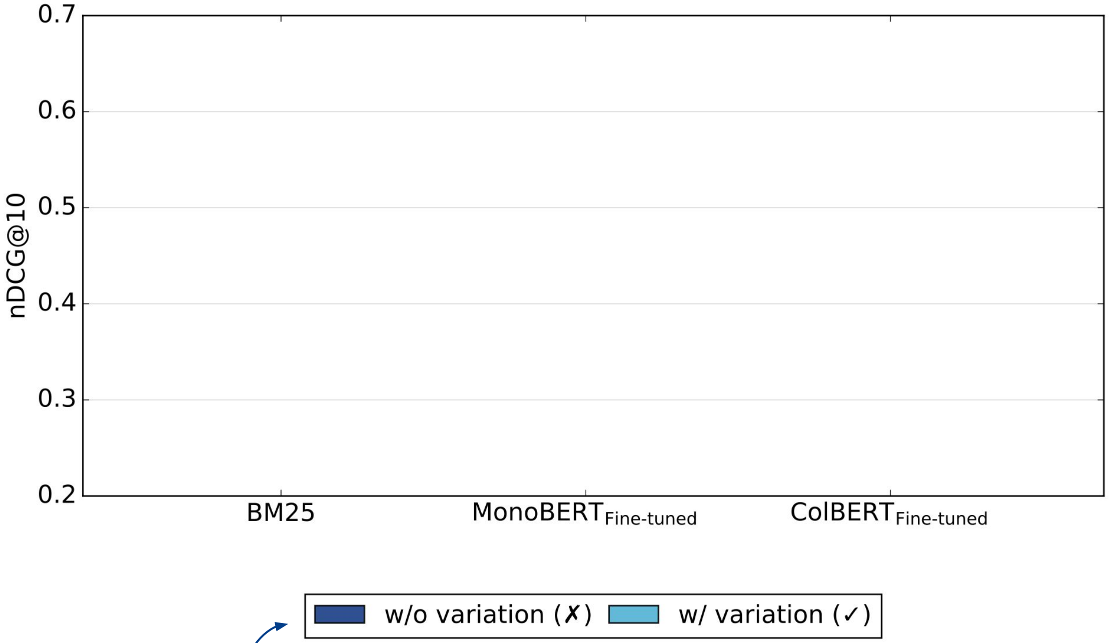
# Dialect variation results



*average over 5 dialects

An.

# Dialect variation results

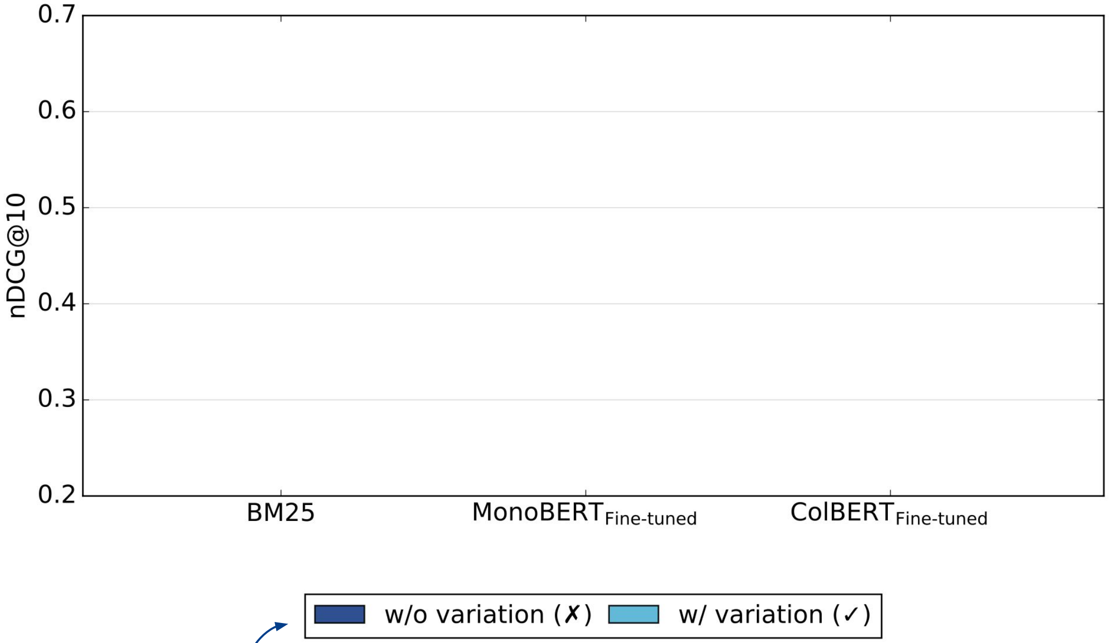


Exclude documents containing variations

*average over 5 dialects

An.

# Dialect variation results

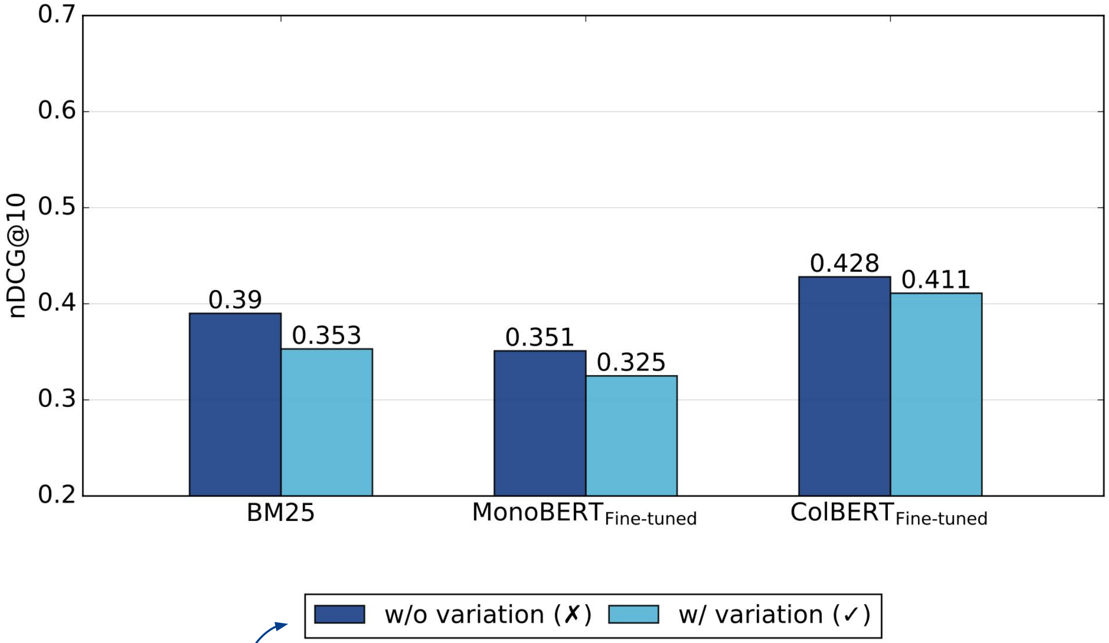


Exclude documents containing variations

full analysis split

*average over 5 dialects

# Dialect variation results



*average over 5 dialects

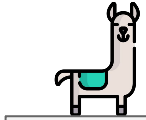
Exclude documents containing variations

full analysis split

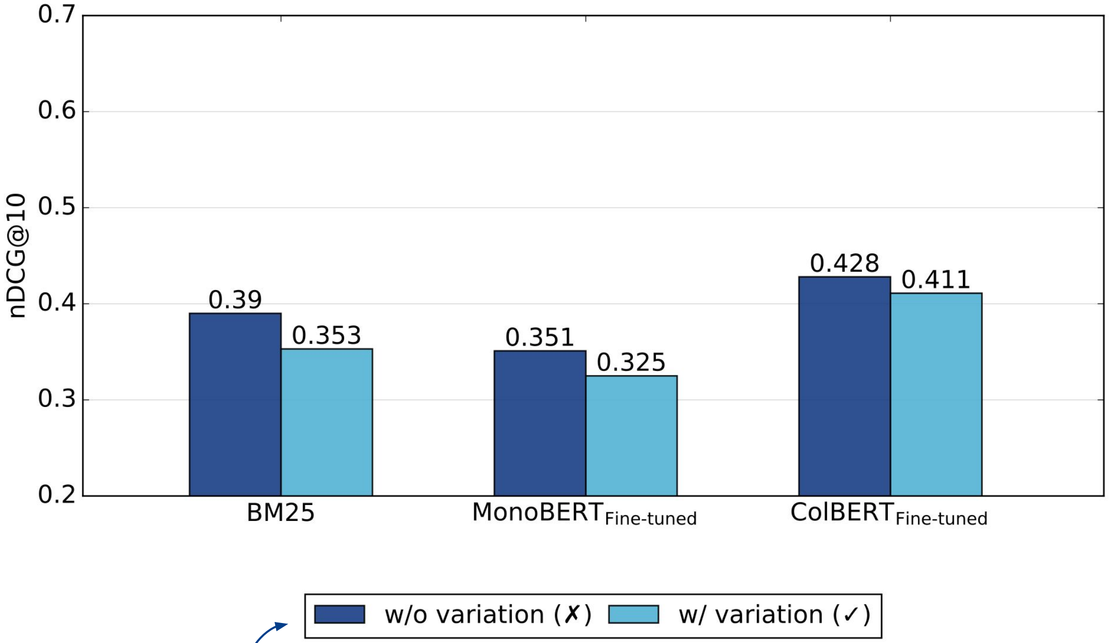
An.

# Document translation results

Can we use LLMs to close the dialect gap?



Document transl.
Dialect → DE



Exclude documents containing variations

full analysis split

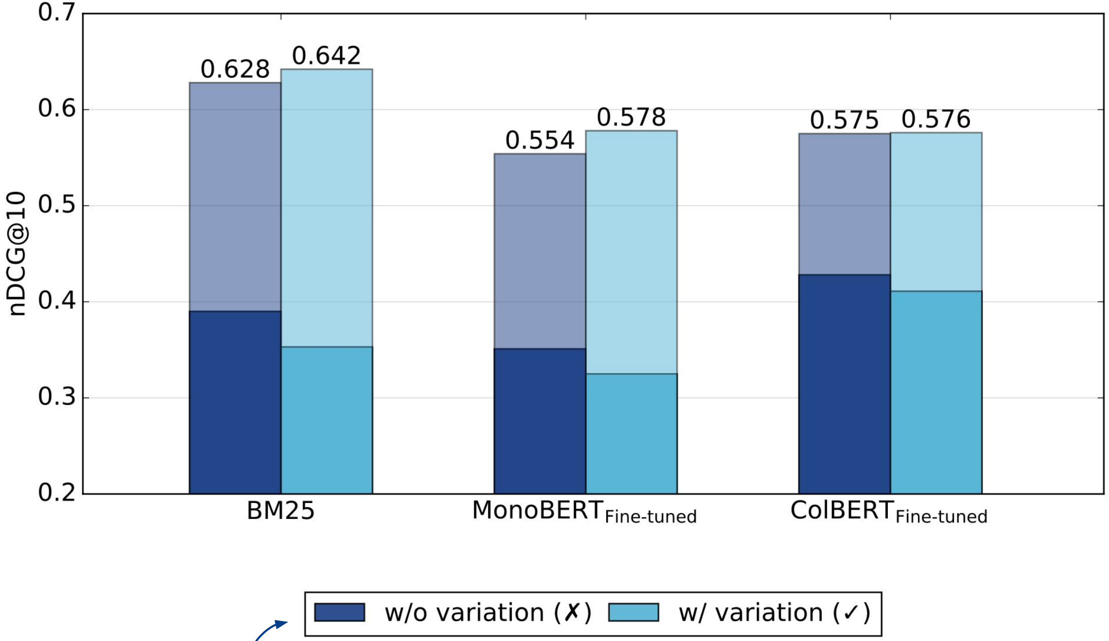
*average over 5 dialects

An.

# Document translation results

Can we use LLMs to close the dialect gap?

Document transl.
Dialect → DE



*average over 5 dialects

Exclude documents containing variations

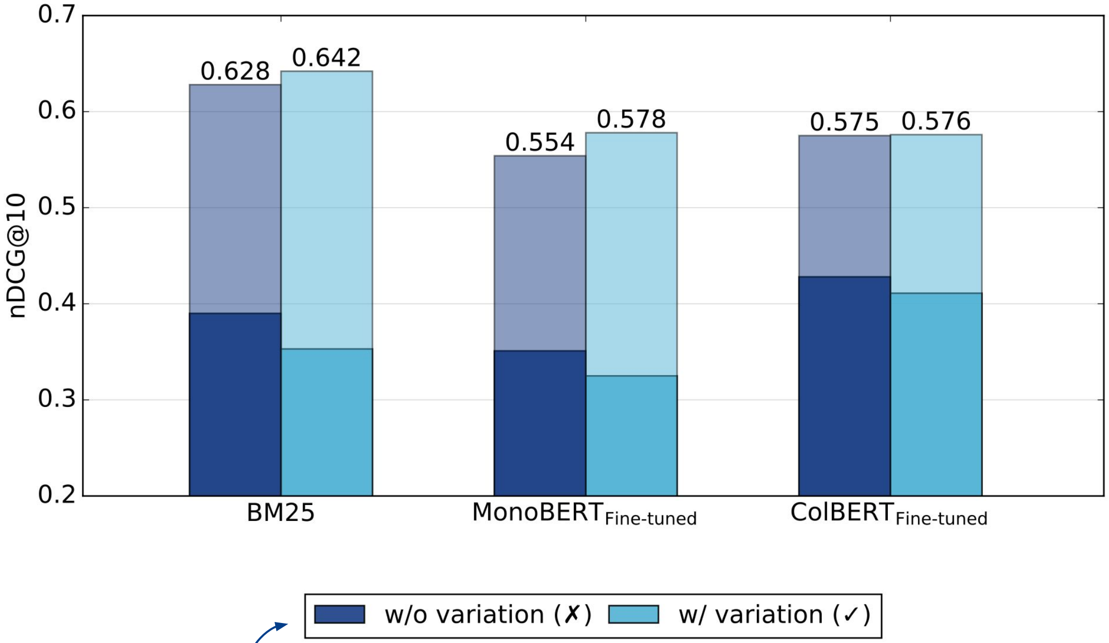
full analysis split

An.

# Document translation results

Can we use LLMs to close the dialect gap?

Document transl.
Dialect → DE



Exclude documents containing variations

full analysis split

There are still large gaps!

*average over 5 dialects

# Conclusion

- We introduce WikiDIR, a cross-dialect information retrieval dataset.
- We release dialect variation dictionaries for German dialects.
- More results and analyses in the paper.

GitHub

# Conclusion

- We introduce WikiDIR, a cross-dialect information retrieval dataset.
- We release dialect variation dictionaries for German dialects.
- More results and analyses in the paper.

CDIR is novel and challenging task!

→ Low-resource

→ High-Variance

The gaps are still large.

GitHub