Bi-Text Mining Across German Dialects:
On the Role of Synthetic Training Data for Dialect Adaptation

Jing Wang, Barbara Plank, Robert Litschko

Jing.Wang1@campus.lmu.de

LREC 2026 Palma

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CIS
nlp
mcml Munich Center for Machine Learning

Agenda

- Introduction
- Evaluation Protocol
- Experimental Setup
- Results
- Conclusion

Introduction

Dialects vs. Standard Languages

[en] Don’t hurry.

[swg] No ned huddla.

[de] Beeil dich nicht.

Introduction

Dialects vs. Standard Languages

[en] Don’t hurry.

[swg] No ned huddla.

[de] Beeil dich nicht.

Derived from “Huddel”: rags used by bakers to clean the oven.

→ Dialects capture culture-specific language use and regional spelling variations.

Introduction

Dialects vs. Standard Languages

[en] Don't hurry.

[swg] No ned huddla.

[de] Beeil dich nicht.

Derived from "Huddel": rags used by bakers to clean the oven.

→ Dialects capture culture-specific language use and regional spelling variations.

Multilingual encoders are predominantly trained and evaluated on standard languages.

Introduction

Trends and Gaps

- Dialect-aware MT evaluation and representation learning
  - Bavarian NMT case study (Her and Kruschwitz, 2024)
  - Creation of German dialect dictionaries (Litschko et al., 2025)

→ Underexplored applications on German dialect

- Lack of dialect-aware benchmarks for evaluating multilingual bi-encoders

Introduction

Trends and Gaps

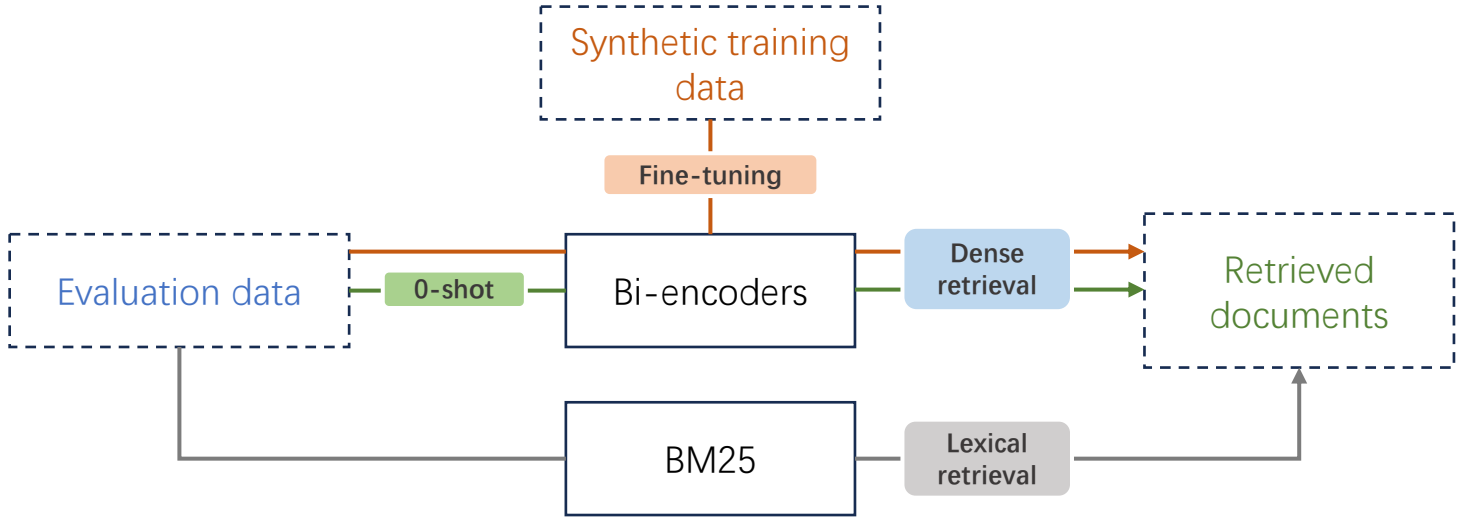
- Data scarcity → synthetic data augmentation
  - Synthetic query-document pair generation with LLMs (Jeronymo et al., 2023).
  - Morphologically-aware dictionary-based data augmentation for MT (Alam et al., 2024).

→ Effectiveness of synthetic data augmentation is unknown for bi-text mining.

Introduction

Contributions

- Evaluation protocol for dialect-aware translation retrieval in three German dialects
- Comprehensive evaluation of multilingual bi-encoders in both zero-shot and finetuned settings
- A study of synthetic data augmentation strategies for dialect retrieval and analysis on factors affecting task difficulty



Introduction

Research Questions

- RQ1: How well do SOTA bi-encoders perform in bi-text mining when queries are written in German dialects, compared to when they are written in English?
- RQ2: To what extent does training on translated data from dictionaries and LLMs improve the retrieval performance of bi-encoders?
- RQ3: How robust is the performance of bi-encoders with respect to different ratios of dialect code mixing?

Agenda

- Introduction
- Evaluation Protocol
- Experimental Setup
- Results
- Conclusion

Evaluation Protocol

Data Source

- Dialect–Standard German pairs:
  - **Low German** – Standard German (nds–de)
  - **Bavarian** – Standard German (bar–de)
  - **Alemannic** (Swiss German + Swabian) – Standard German (als–de)

- High-resource language pair:
  - **English**-Standard German (en–de)

- Dataset sources:
  - Tatoeba (Tiedemann, 2020)
  - WikiMatrix (Schwenk et al., 2021)
  - Wikimedia

Evaluation Protocol

Data Quality

- Manual investigation

DE side from bar-de-wikimatrix:	Bavarian side from bar-de-wikimatrix:
1980 Ein paar Schritte zurück.	1980 Ein paar Schritte zurück.
Weißt du, wie viel Sterne stehen?	Weißt du, wie viel Sterne stehen?
Was die Standbesitzer am liebsten kochen.	Was die Standbesitzer am liebsten kochen.
Ihm solle vielmehr im Himmel Göttlichkeit zuteil-werden.	An Heiland (Salvator Mundi) gibts aa in ondan Religionen.
Der gebürtige Zuger spielt mittlerweile im Sitzen.	De Grundschui sted aa nu in Aufham.
Patienten mit Erythrodermie sind typischerweise sehr warm angezogen.	De Etymologie vom Nama is ned genau kleat.

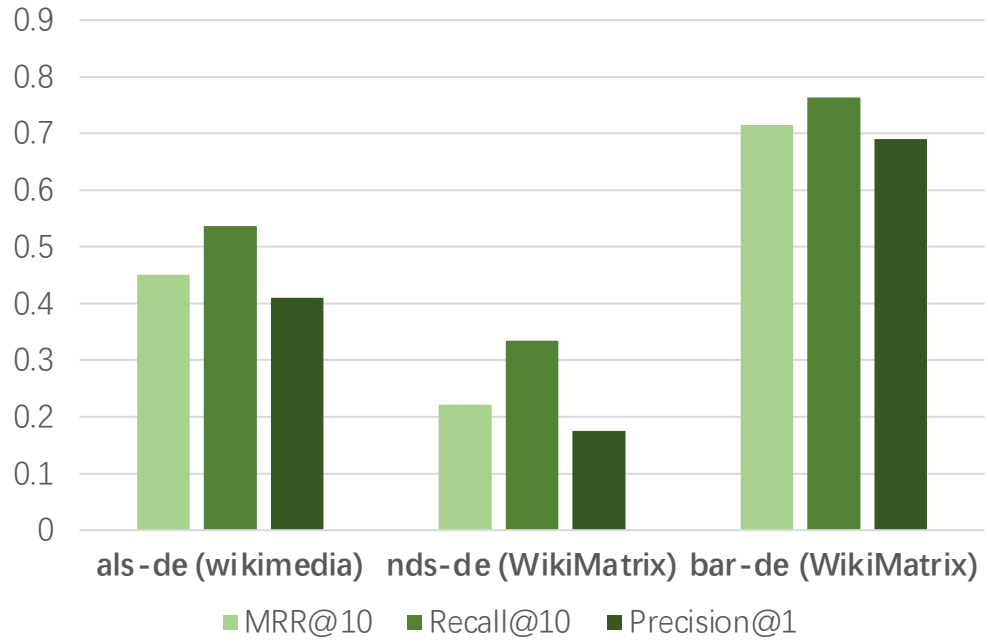
DE sentences on bar side

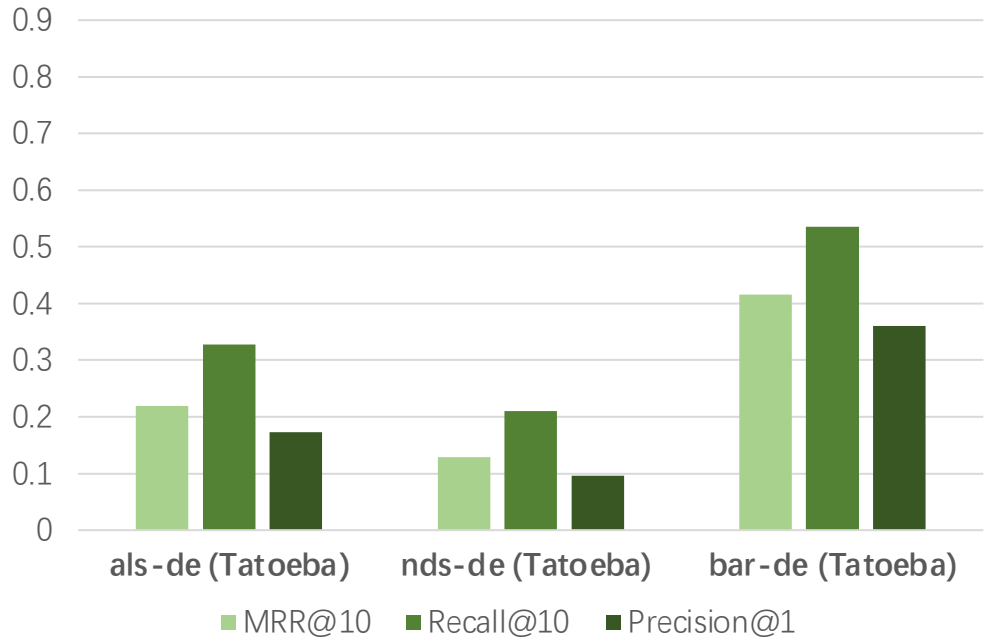
Semantic misalignments

Evaluation Protocol

Data Quality

- Quantifying the Lexical Overlap (BM25)

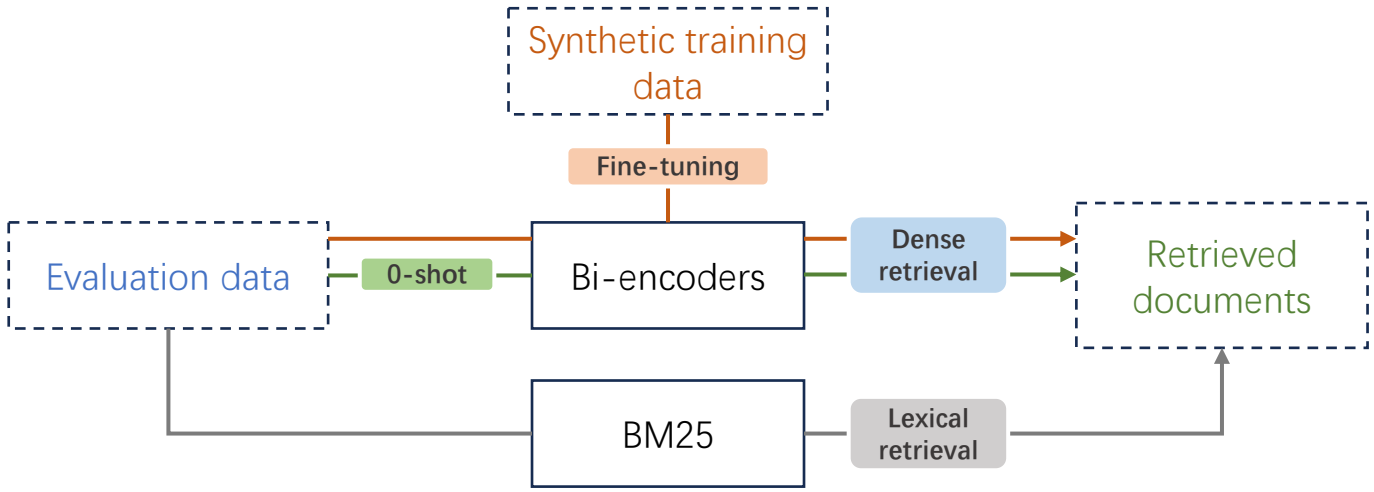




Higher risk of introducing lexical shortcuts during evaluation

Evaluation Protocol

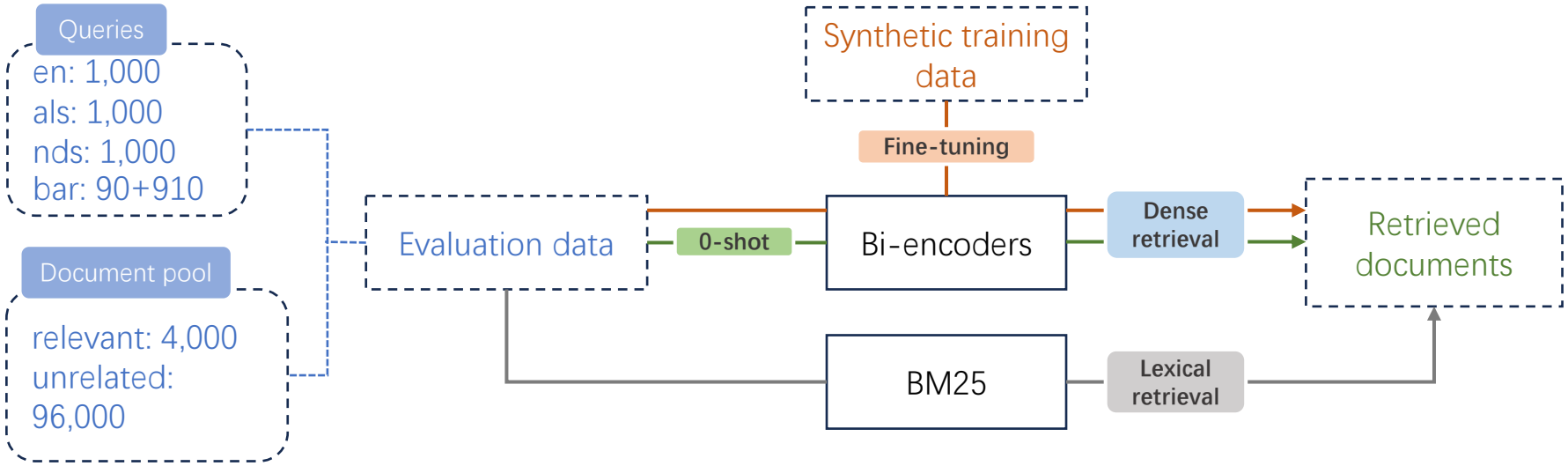
Dialect-to-German Evaluation



Evaluation Protocol

Dialect-to-German Evaluation

- Test Dataset:
  - For each language pair {en, als, nds}–de, we extract 1,000 parallel instances from Tatoeba.
  - Since bar–de contains only 90 instances, we add 910 GPT-translated instances of en-de split from Totoaba.
  - We augment 4,000 relevant German "documents" with 96,000 negatives (source: en-de split).



Evaluation Protocol

Ablation Study

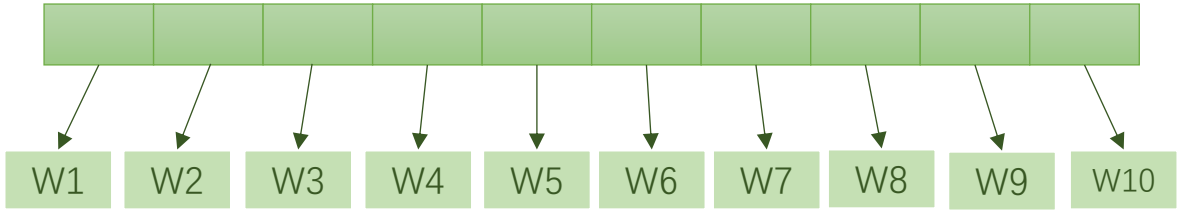
- Dialect-Standard mixtures

Select all German sentences with a length of 10 words (=39 sentences).

Evaluation Protocol

Ablation Study

- Dialect-Standard mixtures

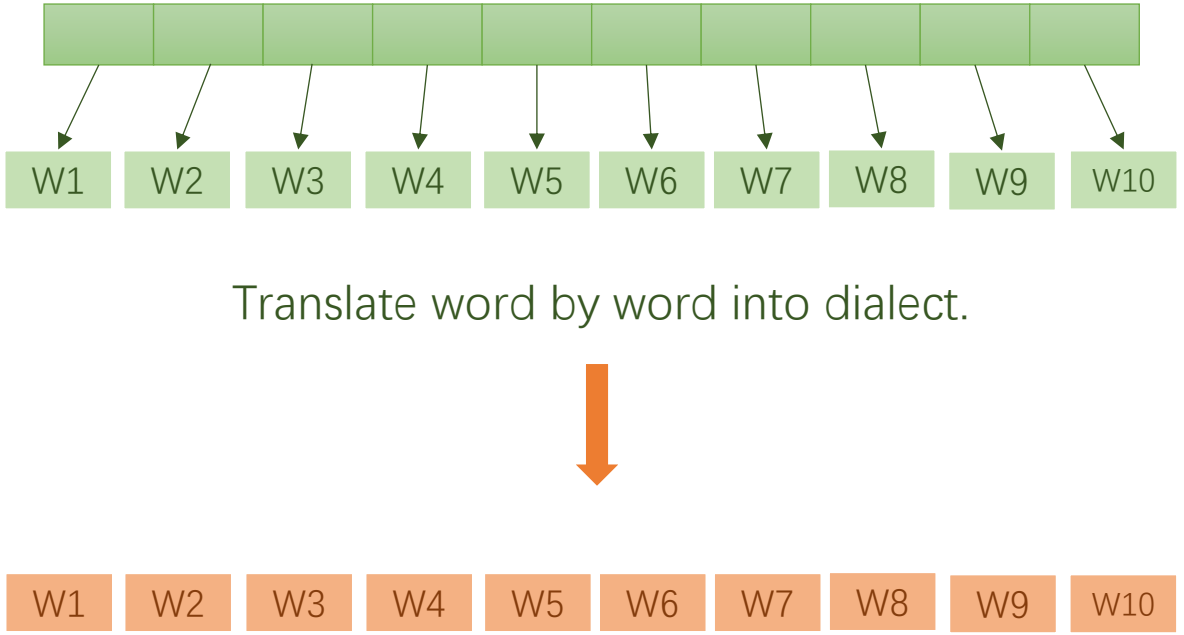


Tokenize German sentence.

Evaluation Protocol

Ablation Study

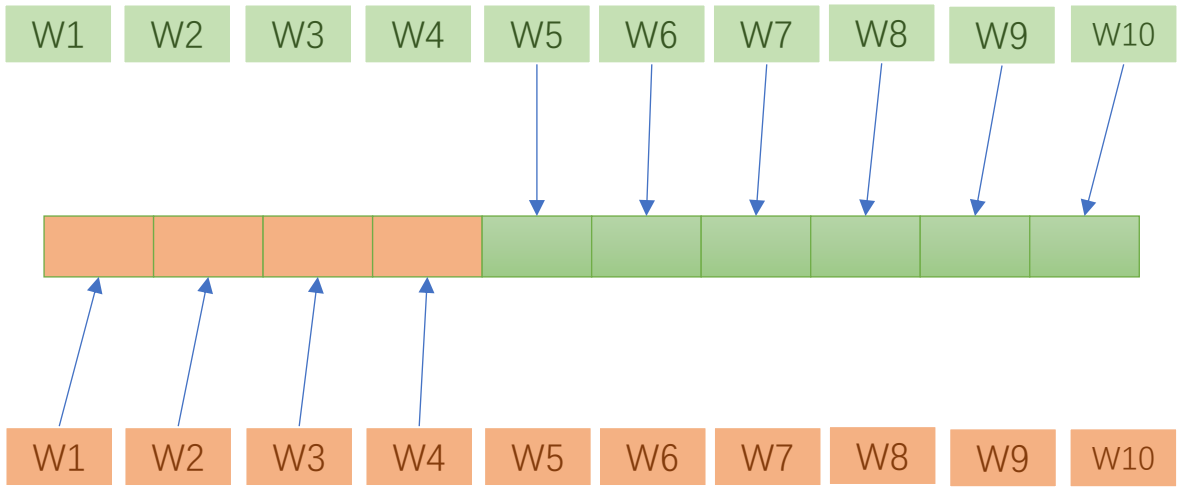
- Dialect-Standard mixtures



Evaluation Protocol

Ablation Study

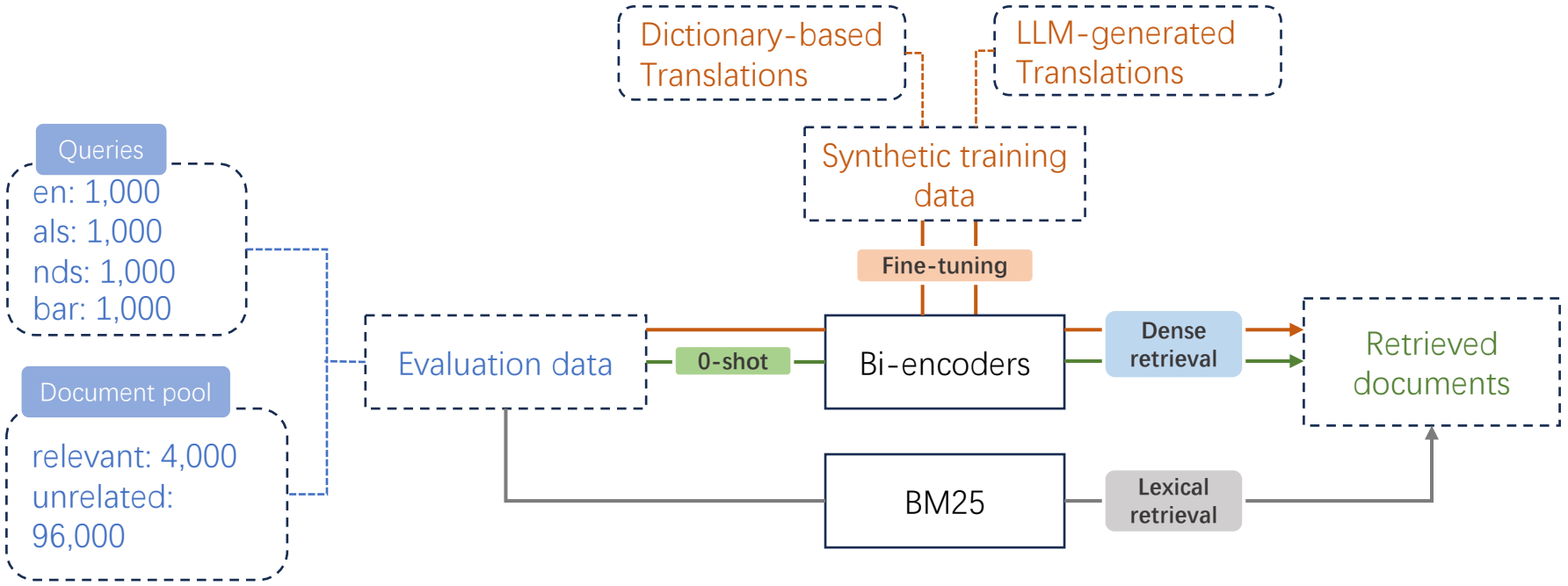
- How does the retrieval performance vary with different code mixing ratios?
  - We evaluate sentence retrieval with different proportions of Bavarian words (20%, 40%, …, 100%).



Substitute different proportions of words and form a new sentence.

Evaluation Protocol

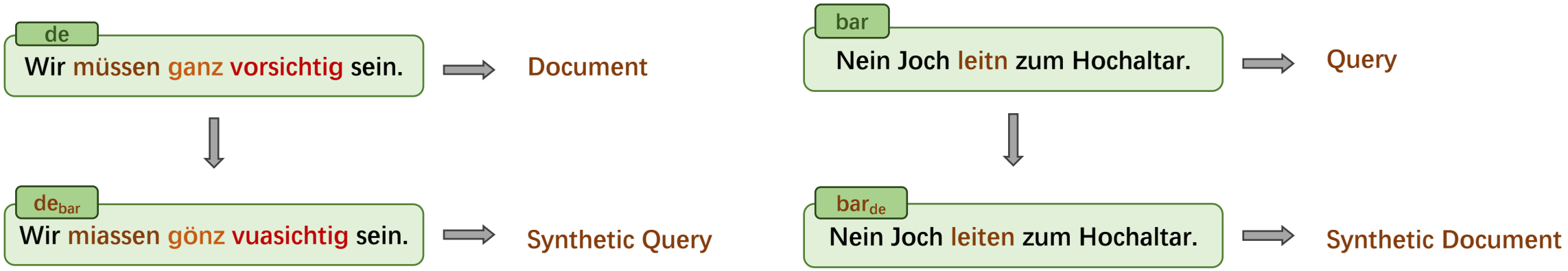
Synthetic Training Data



Evaluation Protocol

Synthetic Training Data

- Dictionary-Based Translations:
  - **DiaLemma:** Bavarian dialect variation dictionary (Litschko et al., 2025)
  - Source: bar and de texts from WikiMatrix
  - **Word-level** code-switching
  - 32,458 pairs
  - Low vocabulary coverage



Evaluation Protocol

Synthetic Training Data

- LLM-Generated Translations:
  - Source: de sentences from Tatoeba (de-en)
  - Translated to all **3 dialects** and merged into **one mixed set** (GPT-4o)
  - 27,000 pairs in total

Prompt

Translate the following Standard German sentence into natural, fluent {target dialect}. Only output the translation. Try to aim for diverse translations.

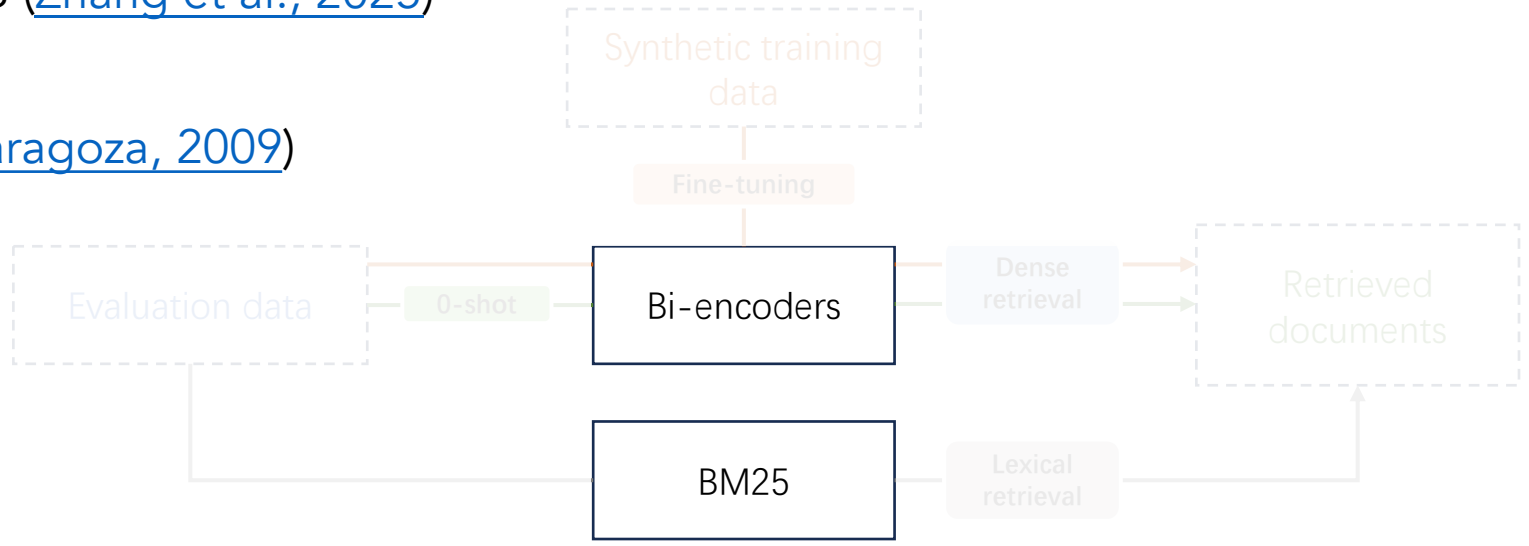
Agenda

- Introduction
- Evaluation Protocol
- Experimental Setup
- Results
- Conclusion

Experimental Setup

Models

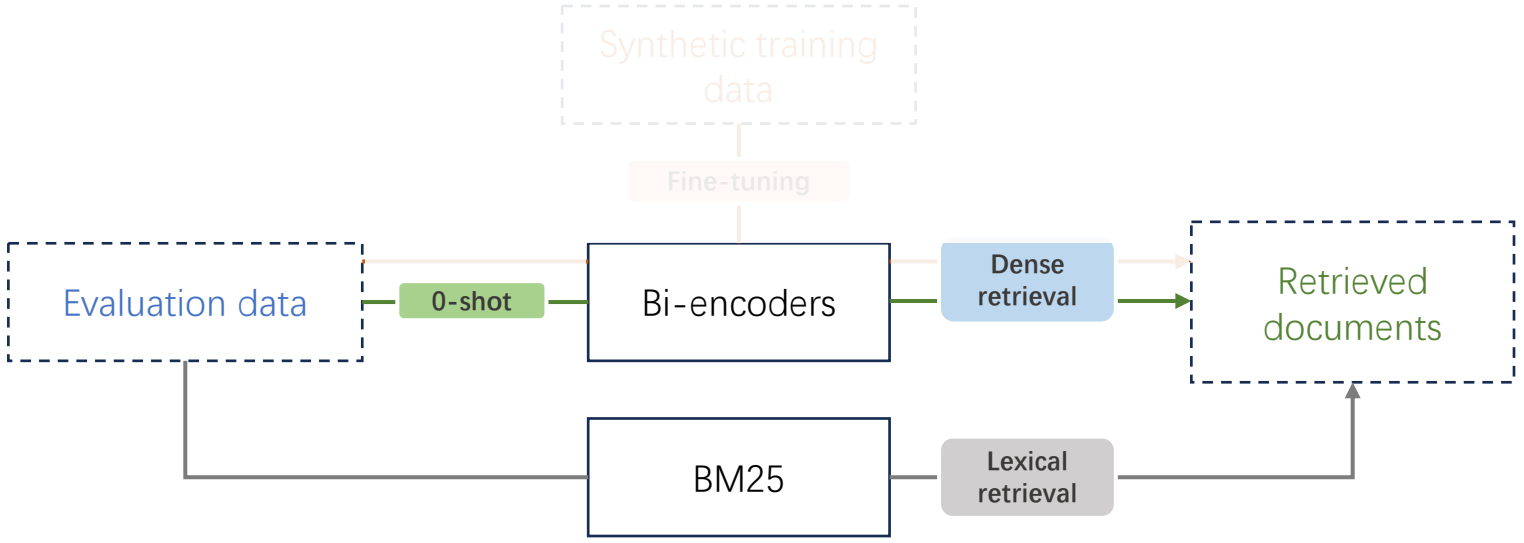
- Bi-encoders:
  - LaBSE (Feng et al., 2020)
  - bge-m3 (Chen et al., 2024)
  - gte-multilingual-base (Zhang et al., 2024)
  - Qwen3-Embedding-0.6B (Zhang et al., 2025)
- Baseline:
  - BM25 (Robertson and Zaragoza, 2009)



Experimental Setup

Zero-Shot Evaluation

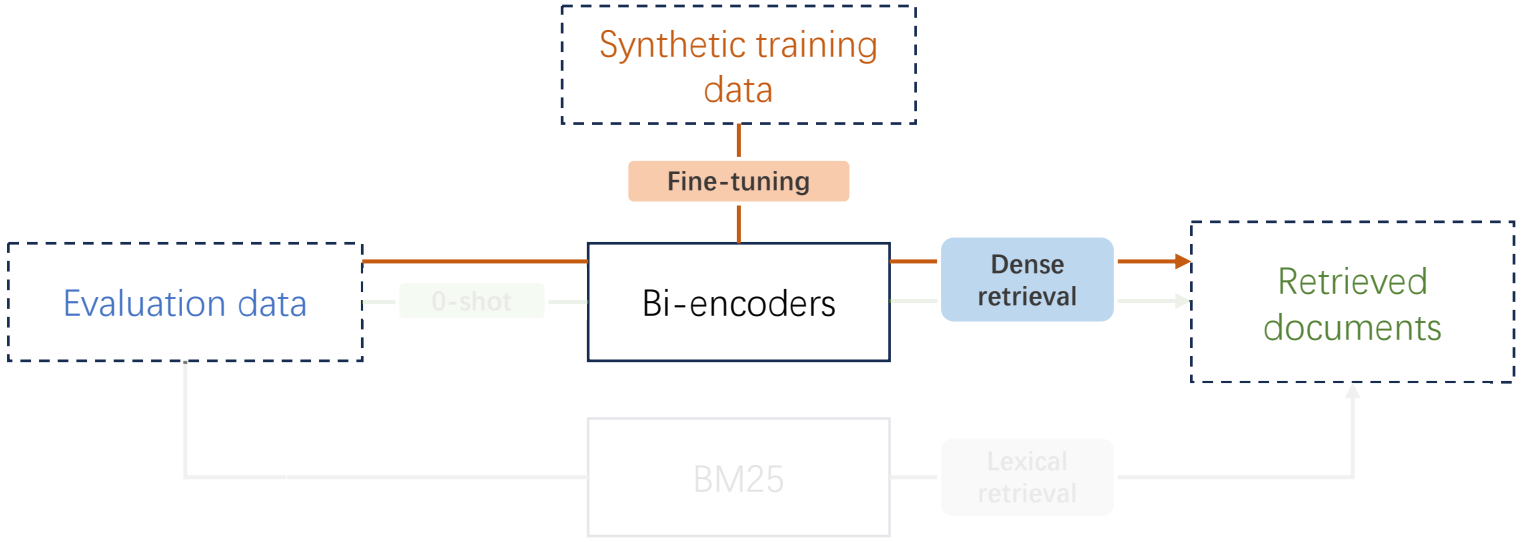
- Rank documents according to their **cosine similarity** to the query.
- Sparse Retrieval (BM25): **lexical baseline** and proxy for task-level difficulty.



Experimental Setup

Fine-tuning

- We fine-tune models using the **InfoNCE loss with in-batch negatives** (Oord et al., 2019).
- Models are fine-tuned **separately on both types of** synthetic data.

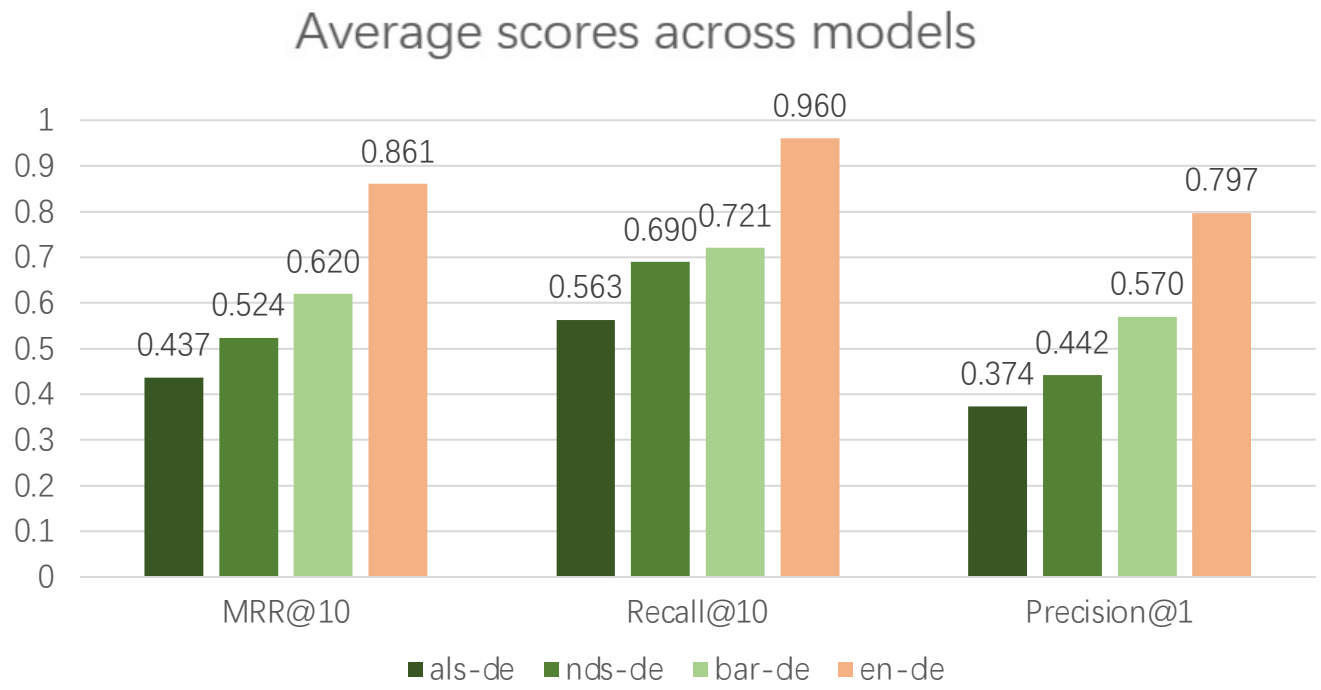


Agenda

- Introduction
- Evaluation Protocol
- Experimental Setup
- Results
- Conclusion

Results

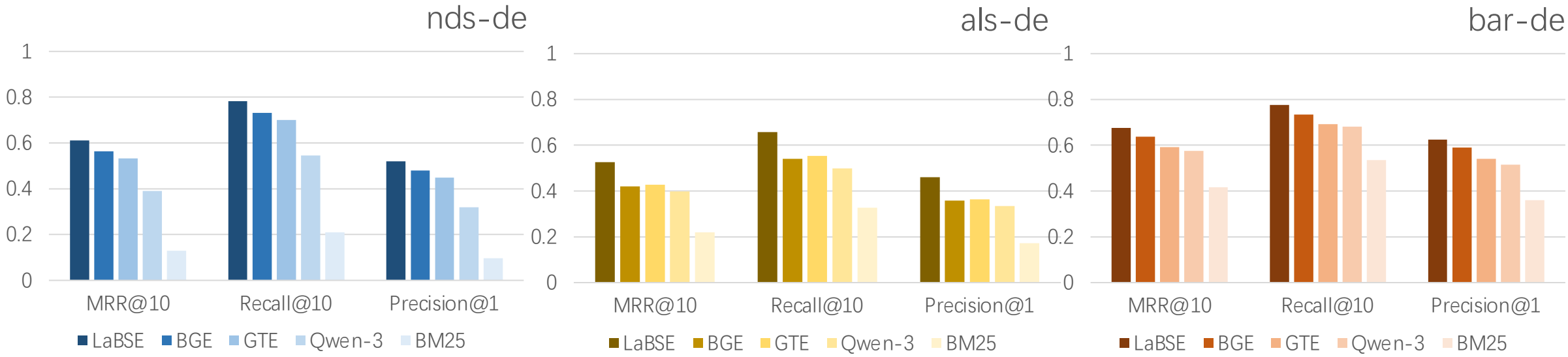
Zero-shot Results



- Current models score on average **0.2-0.4 higher (Precision@1)** in aligning **standard and high-resource languages** than dialects.

Results

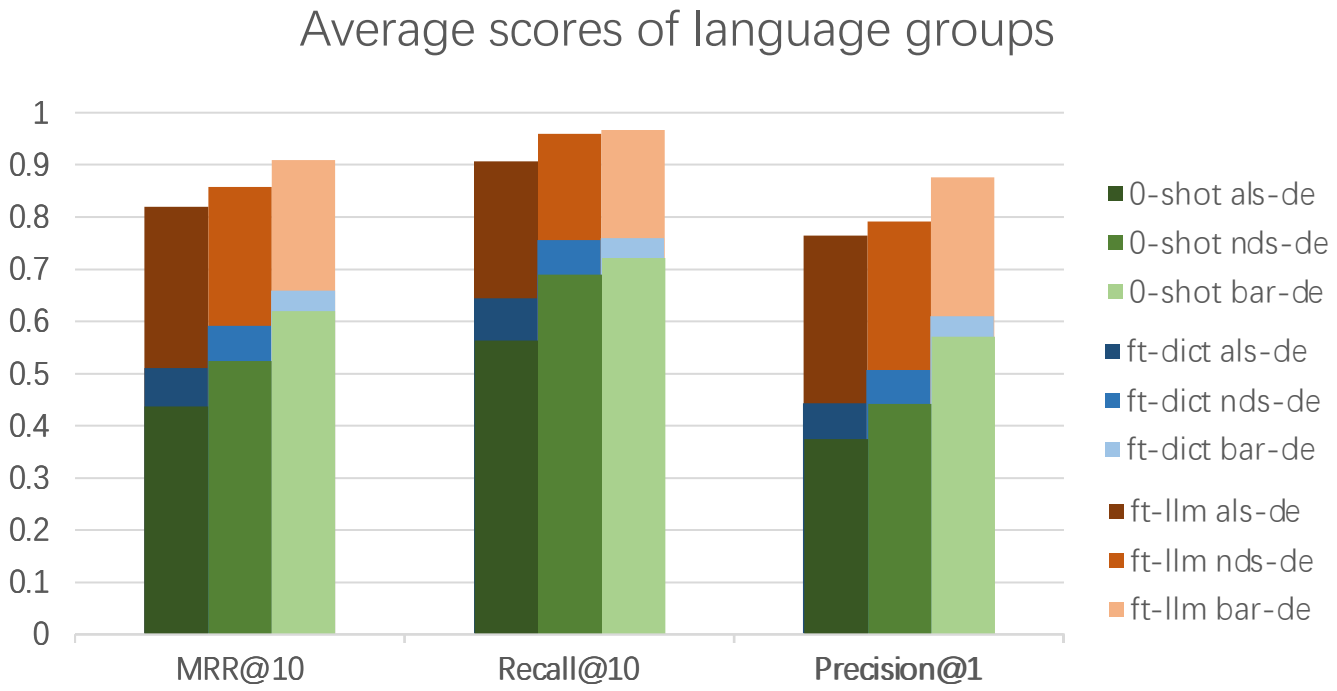
Zero-shot Results



- All **dense retrieval models** substantially outperform BM25.
- **LaBSE** achieves the highest overall performance.
- All models perform best on **bar-de**.
- Lexical overlap **directly relates** to retrieval difficulty.

Results

Fine-tuning Results

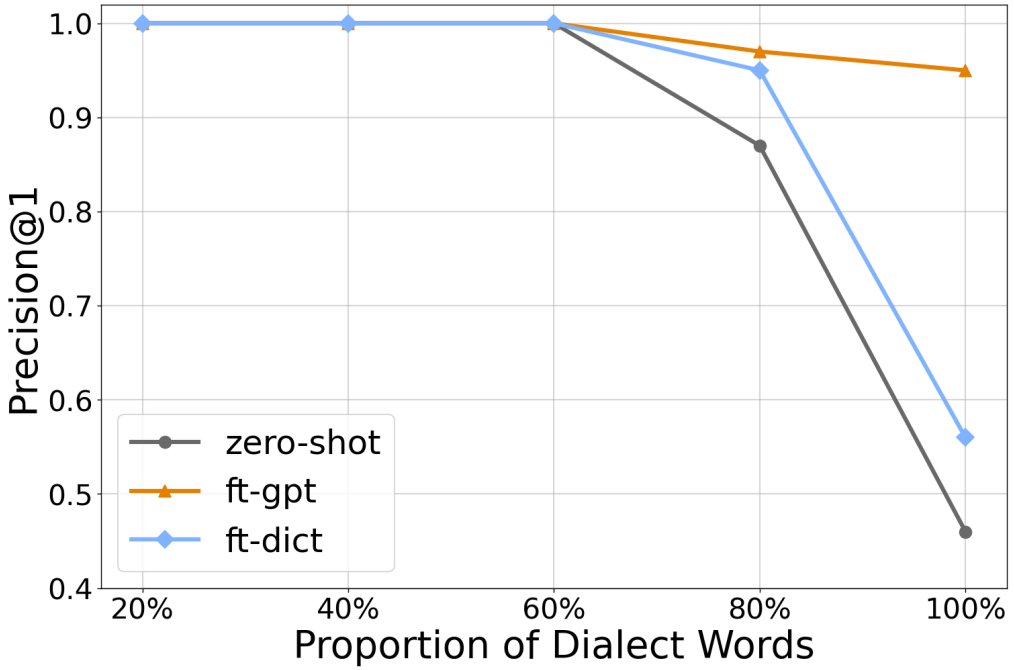


- Our fine-tuning method **gains higher average scores (up to 0.4 for als-de on Precision@1)** compared to zero-shot retrieval across models.
- Fine-tuning on LLM-generated translations outperforms dictionary-based translations (with **up to 0.32 higher Precision scores for als-de).**

Results

Ablation Study: Robustness to Dialect Mixing

- 20%-60% dialect ratio → performance remains high.
- >60% ratio → performance drops.
- Sufficient token overlap can compensate for a model's lack of dialect understanding.



Results obtained with LaBSE. Trends are consistent across all bi-encoders (see paper).

Agenda

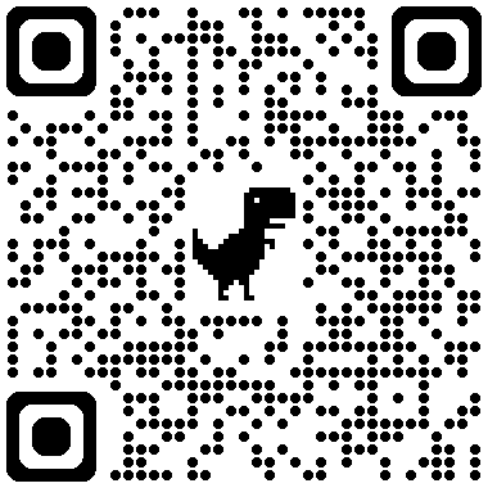
- Introduction
- Evaluation Protocol
- Experimental Setup
- Results
- Conclusion

Conclusion

- Gap between retrieval in standard languages (en-de) and dialects (dial-de).
- Fine-tuning on synthetic data consistently improves results, especially for LLM-generated translations.
- Retrieval effectiveness starts to drop when the proportion of dialect words exceeds a critical threshold.

Thank you for your attention!

Code and Data



https://github.com/mainlp/dialect-bitext-mining

References

- 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. In Findings of the Association for Computational Linguistics: ACL 2024, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Wan-hua Her and Udo Kruschwitz. 2024. Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study. In Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, pages 155–167, Torino, Italia. ELRA and ICCL.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361, Online. Association for Computational Linguistics.
- Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval.

References

- Robert Litschko, Oliver Kraus, Verena Blaschke, and Barbara Plank. 2025. Cross-Dialect Information Retrieval: Information Access in Low-Resource and High-Variance Languages. In Proceedings of the 31st International Conference on Computational Linguistics, pages 10158–10171, Abu Dhabi, UAE. Association for Computational Linguistics.
- Robert Litschko, Verena Blaschke, Diana Burkhardt, Barbara Plank, and Diego Frassinelli. 2025. Make Every Letter Count: Building Dialect Variation Dictionaries from Monolingual Corpora. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14157–14174, Suzhou, China. Association for Computational Linguistics.
- Robertson, Stephen, and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval 3.4 (2009): 333-389.
- Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In Proceedings of the Fifth Conference on Machine Translation, pages 1174–1182, Online. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

References

- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models.