Fine-Tuning and Transferability in Legal Information Retrieval

Advanced Information Retrieval — Group 27

Raphael Habichler • Mark Sesko • Paul Brandstätter

Repository: https://github.com/rhabichl/Advanced-Information-Retrieval

Adapter: https://huggingface.co/krapfi/Qwen3-Embedding-8B-Ger-Legal

Introduction

- We study embedding-based legal retrieval across jurisdictions.
- Research question: Can a model fine-tuned on German law improve retrieval in Austrian and Chinese law?
- Goal: build legal search that transfers well without collecting large labeled datasets for every country.

Data

- We evaluate on three datasets:
  - German: GerDaLIRSmall (test)
  - Austrian: synthetic query-document pairs created from Austrian legal texts
  - Chinese: LeCaRDv2 (test)
- German and Chinese have relevance labels (can have multiple relevant documents per query).
- Austrian setup is closer to one-to-one matching (each query has one source document).

Methods

- Model: Qwen3-Embedding-8B
- Baseline pipeline:
  - Download datasets
  - Embed queries and documents
  - Rank by dot-product similarity
- Evaluation metrics: Recall@k, Precision@k, nDCG@k for $k \in \{1, 5, 10, 20, 50, 100\}$
- Fine-tuning:
  - Train a LoRA adapter (PEFT) on German query-document pairs (GerDaLIR)
  - Re-run the same evaluation with the adapter loaded

Results (German)

- Fine-tuning improves retrieval on German data.

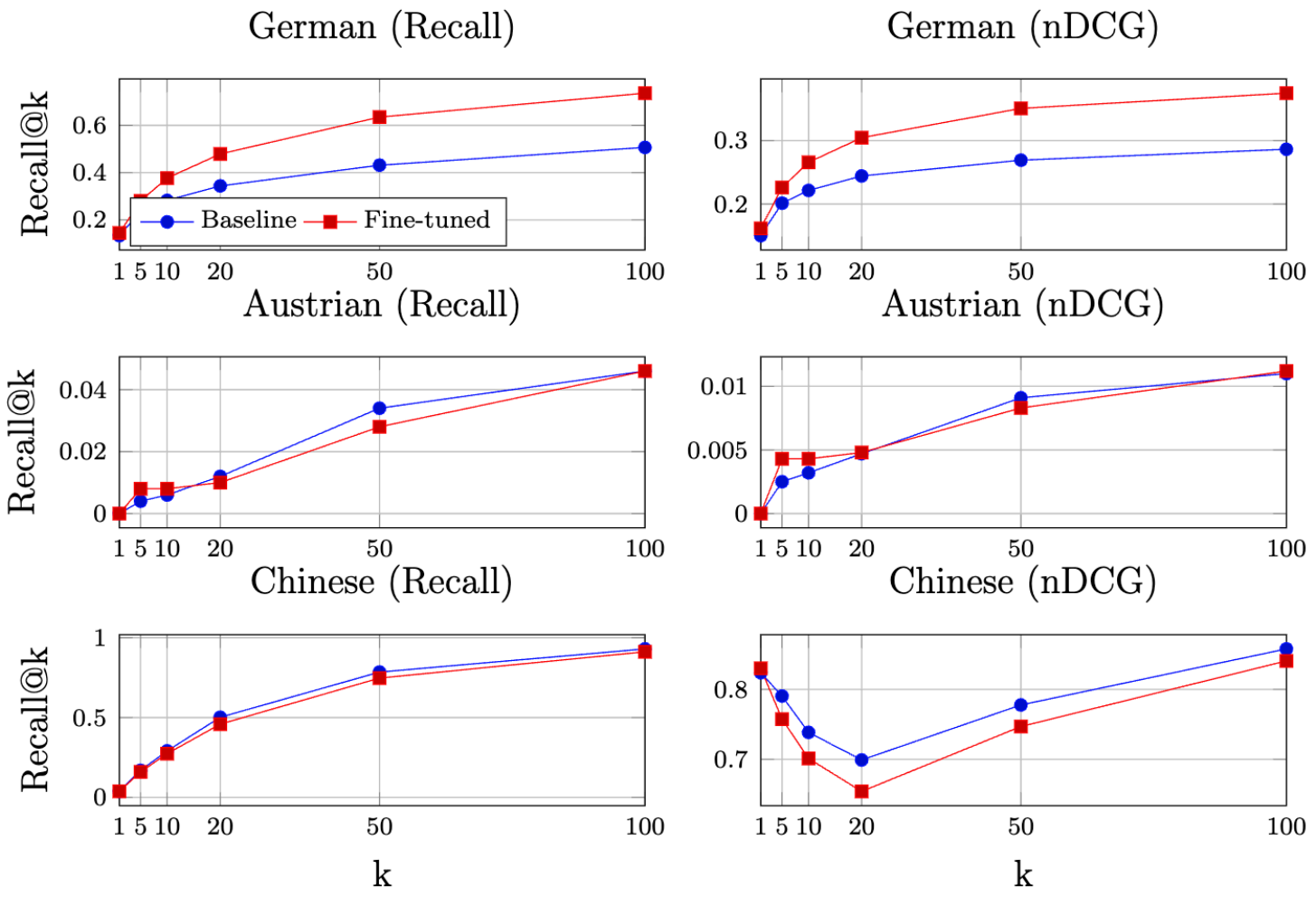
Metric	@1	@5	@10	@20	@50	@100
Recall (base)	0.1330	0.2343	0.2826	0.3436	0.4313	0.5066
Recall (FT)	0.1445	0.2803	0.3767	0.4788	0.6350	0.7361
nDCG (base)	0.1500	0.2013	0.2214	0.2442	0.2691	0.2863
nDCG (FT)	0.1615	0.2260	0.2657	0.3046	0.3508	0.3748

Results (Austrian + Chinese)

- No clear transfer to Austrian and Chinese.
- Key examples:

Dataset	Metric	Base	Fine-tuned
Austrian	Recall@100	0.0460	0.0460
Austrian	nDCG@100	0.0110	0.0112
Chinese	Recall@100	0.9299	0.9118
Chinese	nDCG@100	0.8579	0.8407

Visualization



Interpretation

- German fine-tuning improves German retrieval because training and evaluation match.
- For Austrian and Chinese, we do not see clear transfer.
- Possible reasons:
  - Domain shift (different legal system)
  - Language shift (Chinese)
  - Synthetic Austrian queries may not match real search behavior

Conclusion (and limitations)

- Main takeaway: fine-tuning on German helps German retrieval, but does not automatically transfer to Austrian or Chinese.
- Limitations:
  - Synthetic Austrian dataset
  - Random query sampling in evaluation
  - Truncation to a fixed text length
  - Single fine-tuning configuration

Thank you

Questions?