Advanced Information Retrieval

Fine-Tuning and Transferability in Legal Information Retrieval

Group Number: 27

Raphael Habichler* Mark Sesko† Paul Brandstätter‡

November 11, 2025

1 Abstract

This project tests whether embedding models fine-tuned on one legal system work well on others. We fine-tune Qwen3-Embedding-8B on German legal data and test its performance on Austrian, German, and Chinese legal documents.

2 Idea and Goal

Research Question: Can a legal embedding model fine-tuned on German law effectively retrieve relevant documents in other jurisdictions (Austrian and Chinese law)?

We want to build legal search systems that work across different countries without needing large training datasets for each one. This solves the problem of searching and comparing laws across different legal systems.

3 Main Task

We fine-tune the Qwen3-Embedding-8B model [4] on German legal data and test how well it works on Austrian and Chinese law.

4 Dataset and Processing

4.1 Data Sources

We use three legal datasets. Austrian Law Data is synthetic. We create it from Austrian legal documents and generate queries using LLMs since no labeled pairs exist. German

*student-id: 12419578, Role: Data Processing & Fine-tuning
†student-id: 12114879, Role: Evaluation & Analysis
‡student-id: 12212566, Role: Dataset Preparation & Documentation

Law Data comes from GerDaLIR [3] and a German legal corpus [1] with pre-labeled query-document pairs for training and testing. Chinese Law Data uses LeCaRDv2 [2] with existing annotations.

4.2 Processing

We normalize text, segment documents, and prepare query-document pairs. For Austrian data, we generate queries from documents using LLMs.

5 Methods and Models

We use Qwen3-Embedding-8B [4] and compare two setups. The baseline tests the pre-trained model directly on all three datasets. The fine-tuned version trains the model on German data first, then tests it on all three datasets.

5.1 Fine-tuning Process

We train the model on German legal query-document pairs from GerDaLIR. The training uses cosine similarity loss, which increases similarity for relevant pairs and decreases it for irrelevant ones. Queries and documents are encoded separately, then pooled into fixed-size vectors. The goal is to make similar legal texts have similar vectors. We then test this trained model on Austrian and Chinese data to see how well it transfers across different legal systems and languages.

6 Evaluation

We use three standard metrics to measure performance. Precision@k shows what fraction of the top-k results are relevant. Recall@k shows what fraction of all relevant documents appear in the top-k results. nDCG@k measures ranking quality by giving more weight to relevant documents ranked higher. We compare the baseline and fine-tuned models on all three datasets to measure how well the model transfers.

7 Workflow

Figure 1 shows our research pipeline from data collection to evaluation.

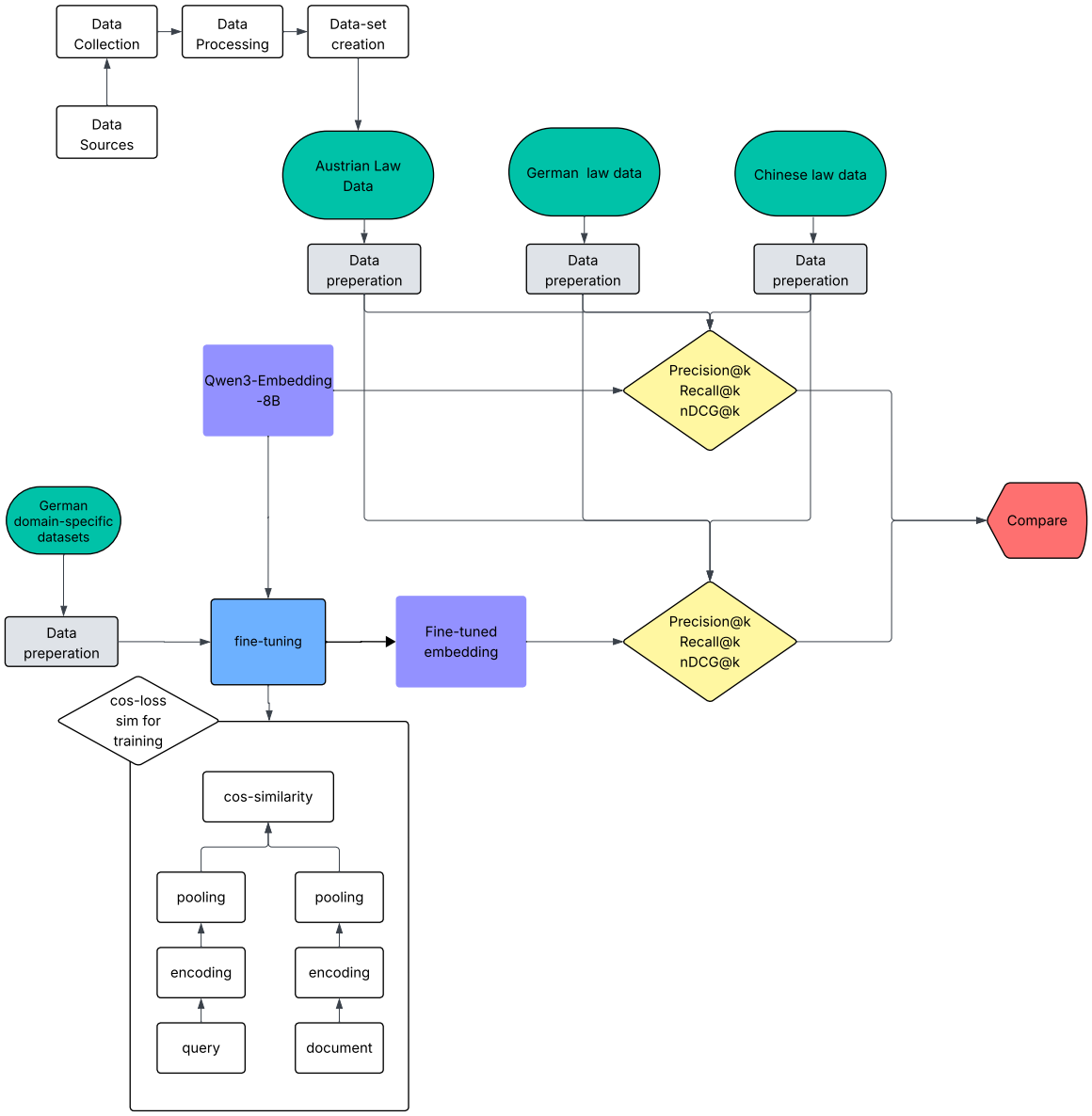

Figure 1: Project workflow and methodology

References

[1] Christoph Hoppe et al. "Towards Intelligent Legal Advisors for Document Retrieval and Question-Answering in German Legal Documents". In: *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 2021, pp. 29–32. DOI: 10.1109/AIKE52691.2021.00011.

[2] Haitao Li et al. *LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset*. 2023. arXiv: 2310.17609 [cs.CL].

[3] Marco Wrzalik and Dirk Krechel. "GerDaLIR: A German Dataset for Legal Information Retrieval". In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 123–128. URL: https://aclanthology.org/2021.nllp-1.13.

[4] Yanzhao Zhang et al. "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models". In: *arXiv preprint arXiv:2506.05176* (2025).