A Novel Synthetic Dataset for Few-Shot Legal Relation Extraction in German

Shiva Banasaz Nouri[1], **Elena Leitner**[2], **Julian Moreno Schneider**[2], **Georg Rehm**[2,3]
[1] Technische Universität Berlin, Berlin, Germany
[2] Deutsches Forschungszentrum für Künstliche Intelligenz, Berlin, Germany
[3] Humboldt-Universität zu Berlin, Berlin, Germany
Corresponding author: georg.rehm@dfki.de

Abstract

The legal domain is particularly challenging for natural language processing due to the personal and sensitive information it contains. Despite significant advances, applying large language models (LLMs) to relation extraction (RE) in legal documents remains challenging, not only because of the task's complexity, but also due to privacy, compliance, and infrastructure constraints under regulations such as the EU AI Act. To address these challenges, we propose a novel synthetic dataset for German legal relation extraction, created using generative LLMs through a controlled, privacy-preserving, template-based pipeline. The dataset allows for reproducible and legally compliant experimentation. We benchmark it using two few-shot learning paradigms, a description-enhanced Model-Agnostic Meta-Learning (MAML) framework and Prototypical Networks with supervised contrastive loss and curriculum-aware prototype enrichment. Our results demonstrate that combining few-shot learning with structured semantic knowledge achieves robust and interpretable results, with the curriculum-aware Proto-Contrastive model reaching an F_1-score of 99.83%.

Keywords: Relation Extraction, Legal NLP, Few-Shot Learning, Meta-Learning, LLMs, Synthetic Data

1. Introduction

Relation Extraction (RE) is a core task in Natural Language Processing (NLP) that seeks to identify and classify semantic relations between entities in unstructured text. At its core, RE takes as input a sentence (or document) with marked entities and outputs a relation label that links them. By transforming raw text into structured relational information, RE underpins applications such as knowledge graph construction, legal search, and document understanding (Hogan et al., 2021). Over the past decade, RE methods have advanced from rule-based and feature-driven approaches to neural architectures and transformer-based encoders, achieving strong results in general-domain benchmarks such as TACRED (Zhang et al., 2017), SemEval, and FewRel (Han et al., 2018).

Nevertheless, RE in specialised domains remains a significant challenge. In particular, German legal texts exhibit complex syntax, long-distance dependencies, and extensive use of compounds and formal expressions (Glaser et al., 2021). At the same time, the lack of annotated legal corpora, exacerbated by privacy restrictions and the high cost of expert annotation (Ariai et al., 2025), prevents the application of supervised learning approaches. As a result, the development and evaluation of relation extraction models for legal German have been severely limited (Peikert et al., 2022).

This scarcity of labeled data underscores an open problem in legal NLP, how to enable robust relation extraction in German legal texts without relying on large annotated corpora. While pretrained language models offer strong generalisation capabilities (Devlin et al., 2019), they still depend on domain-relevant examples to capture the complex syntax and terminology of legal language. Moreover, privacy-sensitive legal data cannot be freely distributed, making reproducible evaluation nearly impossible without synthetic alternatives (Ghosh et al., 2023; Li et al., 2023).

In this paper, we address the question of how to perform effective relation extraction in German legal documents under low-resource conditions.

By leveraging few-shot learning methods such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) and Prototypical Networks with supervised contrastive loss (Snell et al., 2017; Khosla et al., 2020), we design models that generalise from only a handful of annotated examples. To mitigate data scarcity, we further develop a synthetic dataset pipeline using LLMs and curated entity lists, ensuring privacy while generating legally plausible examples (Schmidt et al., 2024; Ali et al., 2024).

The architecture is built around pretrained German BERT encoders (Aßenmacher et al., 2021; Darji et al., 2021), with extensions that fuse semantic relation descriptions into prototype representations. We also propose a curriculum-aware contrastive model that gradually increases semantic complexity during training (Wu et al., 2024).

Through a series of ablation studies our first solution was tested against baselines such as zero-shot embedding similarity (Reimers and Gurevych, 2019) and vanilla Prototypical Networks. These

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10579–10591
11–16 May 2026.

experiments underscored how semantic priors help anchor prototypes in domain meaning (Liu et al., 2022, 2020), contrastive loss enhances class separability (Khosla et al., 2020), and curriculum scheduling facilitates gradual adaptation to semantic complexity (Wu et al., 2024). As later confirmed in our evaluation (Section 5), these components jointly contributed to more stable convergence and improved generalisation across few-shot scenarios.

This paper makes two main contributions.

- We introduce a novel dataset for German legal Relation Extraction, generated through a controlled template-based pipeline, formatted in FewRel (Han et al., 2018) style, and released under an MIT license.[1]
- We present model innovations that combine meta-learning, contrastive prototypes, and curriculum-aware description fusion, demonstrating competitive performance in few-shot legal Relation Extraction.[2]

2. Related Work

Early RE systems were predominantly rule-based or pattern-driven, relying on handcrafted patterns, dependency paths, or linguistic resources (Mintz et al., 2009; Aydar et al., 2020), demonstrating high precision in restricted domains but lacking scalability and robustness. The next generation of feature-based machine learning approaches leveraged lexical, syntactic, and dependency features (Bachinger et al., 2024). While more flexible than purely rule-based methods, they still required heavy feature engineering and domain expertise.

With the advent of deep learning, neural models became dominant. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) can learn distributed representations directly from texts. With Transformer-based architectures (e. g., Devlin et al., 2019), state-of-the-art RE models are now typically based on LLMs that are fine-tuned with task-specific objectives, sometimes augmented with external knowledge or integrated into joint extraction frameworks (Zhao et al., 2024; Giorgi et al., 2019). Recently, LLMs have demonstrated strong in-context learning capabilities, effectively acting as few-shot learners that can perform RE and similar tasks from only a few examples without fine-tuning (Brown et al., 2020). However, deploying such models remains resource-intensive and raises practical challenges in specialised or regulated environments. In domains such as the legal and judicial sectors, ensuring data privacy, interpretability, and regulatory compliance (e. g., AI Act) poses barriers to integrating LLMs directly into production workflows (Ethikrat et al., 2023).

Traditional RE has often been implemented as a pipeline, with Named Entity Recognition (NER) followed by relation classification. As this approach suffers from error propagation, joint extraction methods were proposed to model entities and relations simultaneously (Miwa and Bansal, 2016; Giorgi et al., 2019). Joint models reduce error accumulation and capture interactions between entities and relations more effectively. However, joint architectures remain more complex and data-intensive.

While early RE research focused on the sentence level, real-world scenarios in the legal domain, require reasoning across sentences or paragraphs. This has motivated the development of document-level RE datasets such as DocRED (Yao et al., 2019), for which models need to handle long-distance dependencies and discourse-level context. Recent methods combine pretrained LLMs with graph-based reasoning or attention mechanisms to capture cross-sentence relations (Li et al., 2022; Popovic and Färber, 2022). Yet still, document-level RE remains significantly more difficult than sentence-level RE due to the sparsity of relation signals and the complexity of context modeling.

LLMs such as GPT-4[3] and LLaMA (Touvron et al., 2023) introduced a paradigm shift in RE: rather than requiring extensive fine-tuning, they can perform RE via in-context learning (ICL), where a small number of input-output examples are provided in the prompt (Brown et al., 2020; Wei et al., 2022). Follow-up studies explored strategies such as chain-of-thought prompting (Yu et al., 2023), multi-turn question answering (Wei et al., 2022), and prompt engineering tailored to the legal domain (Li and Yi, 2024). While demonstrating impressive zero- and few-shot generalisation capabilities, the performance of LLMs in specialised domains often lags behind domain-specific models.

Progress in RE has been closely coupled with the availability of annotated datasets. General-domain resources, e. g., TACRED (Zhang et al., 2017), SemEval (Gábor et al., 2018), and FewRel (Han et al., 2018), have been instrumental for benchmarking and developing new methods. For domain adaptation, e. g., CORE (Borchert et al., 2023) and SciERC (Zhang et al., 2024) have introduced relation classification tasks in business and scientific texts. However, these datasets are primarily in English and focus on sentence-level RE.

For German, annotated corpora remain scarce. Resources such as the GermEval shared tasks or the dataset by Leitner et al. (2020) provide cover-

[1] https://github.com/shivanouri/legal_relation_dataset

[2] https://github.com/shivanouri/legal-fewrel

[3] https://openai.com/index/gpt-4o-system-card/

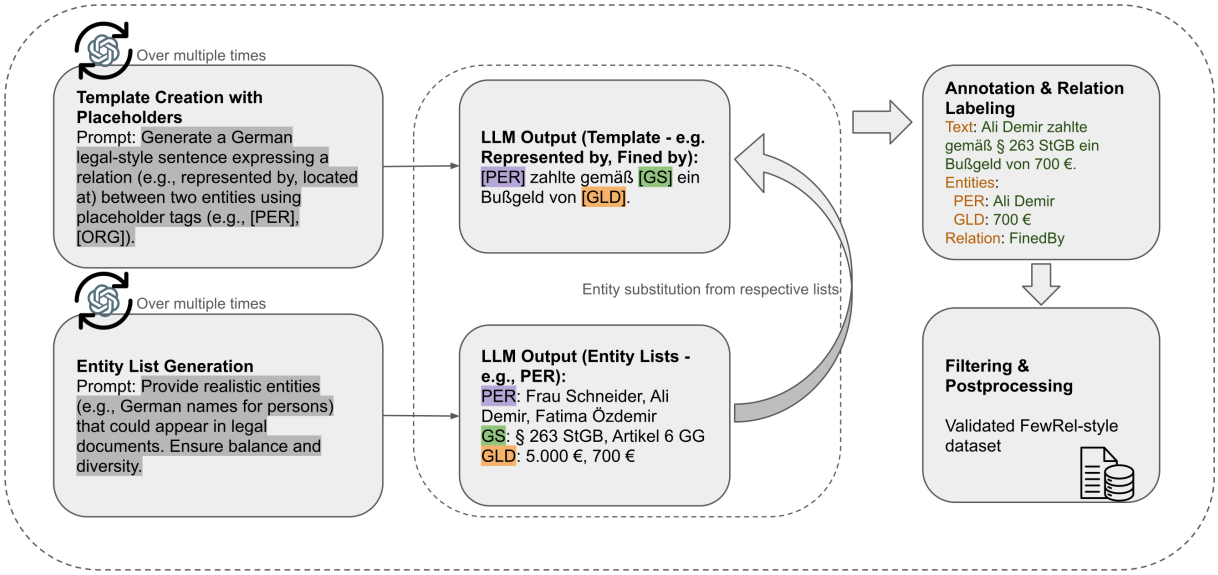


Figure 1: Synthetic dataset creation pipeline. Template sentences with placeholders and entity lists are generated using GPT-4o, combined via random substitution, and automatically annotated with entity spans and relation labels. The data is filtered, postprocessed, and converted into FewRel JSON.

age for legal NER, but relation-level annotations are largely absent. Existing German legal NLP efforts have focused on NER (Glaser et al., 2018; Peikert et al., 2022) or sentence classification (Darji et al., 2021), leaving RE underexplored. The lack of annotated corpora is further exacerbated by privacy regulations such as GDPR, which restrict the use of authentic legal case data.

This scarcity has motivated synthetic data generation approaches. FewRel-style annotation formats have been adopted for few-shot evaluation (Han et al., 2018), and LLM-based methods have recently been applied to generate high-quality synthetic RE datasets (He et al., 2022; Schmidt et al., 2024). In legal NLP, synthetic augmentation is particularly promising as it avoids privacy violations while producing reproducible training resources (Ghosh et al., 2023; Li and Yi, 2024). Nonetheless, to our knowledge, no publicly available FewRel-style dataset exists for German legal RE, which constitutes a major gap that our work addresses.

3. Dataset Pipeline Creation

We constructed a synthetic dataset through a controlled template-based pipeline enhanced with LLMs. The pipeline aims to generate linguistically realistic and semantically precise training examples while preserving privacy and reproducibility. Figure 1 illustrates the main stages of the pipeline, from template generation and entity substitution to annotation, filtering, and conversion into a FewRel-compatible format.

We used GPT-4o to generate German legal-style sentences containing placeholders for entity types (e. g., `[PER]`, `[ORG]`, `[GS]`). The prompts used to generate these template sentences are provided in Appendix A. The templates focus exclusively on grammatical and relational structures. To minimise hallucination, only a small number (5-8) of templates were generated per prompt.

For each entity type, lists were compiled using a combination of LLM assistance and manual refinement. These include diverse names and references, ensuring demographic balance (e. g., gender, ethnic background), coverage of organisational types, and variety of legal statutes across criminal, civil, and administrative law. Placeholders were substituted with entity instances drawn randomly from the curated lists (each list containing about 40 different entities), yielding coherent and legally plausible sentences. Random substitution reduces lexical bias and encourages generalisation across varied contexts. Importantly, because all entities are injected only at this stage rather than generated directly by the LLM, the dataset is protected from memorisation or leakage of sensitive real-world legal cases. This guarantees that the resulting corpus is synthetic, privacy-preserving, and uncompromised. An example is shown in Figure 2, where abstract placeholders (e. g., [PER], [AN]) are replaced with realistic entity names to produce complete legal sentences. This process produces gold-standard annotations: entity spans, their types, and the relation label that corresponds to the template. The data is exported in FewRel-style JSON format, enabling direct use in episodic few-shot training. The dataset focuses on six relation types considered frequent in German legal texts, linking entities such as persons, organiza-

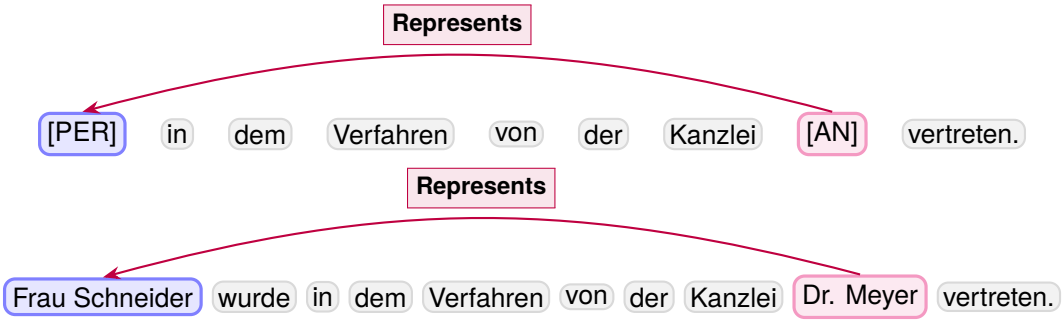


Figure 2: Visualisation of the relation *RepresentedBy*. Top: abstract template sentence with entity placeholders [PER] and [AN]. Bottom: instantiated legal sentence with concrete entities (*Frau Schneider* represented by *Kanzlei Dr. Meyer*). Arrows indicate the semantic relation between head and tail entities.

tions, laws, and monetary amounts (see Table 1).

The generated samples were filtered to remove structurally ambiguous or malformed sentences. Where necessary, templates were rephrased to introduce minor linguistic variation while preserving legal plausibility. Each final sentence contains one valid relation instance, ensuring consistency.

Additionally, we introduce a scoring metric combining sentence length, rarity of tokens, and presence of legal constructions (e. g., subordinate clauses, modal verbs, fixed legal phrases). Sentences with scores above a threshold were labeled as "hard," while others were "easy". Approx. 20% of the corpus falls into the hard category, providing a balanced evaluation setup for few-shot models.

Challenges in LLM-based Data Generation

During the initial experiments with GPT-4o for data generation, several challenges emerged. Our first attempt provided the LLM with the target relation, potential head and tail entities, and asked it to generate corresponding examples. However, this often resulted in swapped head-tail positions, hallucinated relations not present in the prompt, or inconsistent label assignment.

Generating data directly in JSON format also proved ineffective, as the model tended to focus excessively on maintaining structure rather than preserving semantic coherence in the relations. When we attempted to include real-world examples (from Leitner et al., 2020) as few-shot demonstrations, the model initially produced some valid samples, but after only a few generations it began to introduce grammatical errors, inconsistent syntax, and even invented relations without entities. This approach also proved computationally inefficient, as it required a large number of tokens per prompt without delivering proportionally higher-quality outputs. Incorporating inline entity tags (e. g., <PER> . . . </PER>) within the text caused further instability, as the model frequently altered or omitted tags, resulting in inconsistent annotation formats.

Ultimately, the most reliable approach was to decompose the generation task into smaller, highly focused steps, i. e., first generating templates with placeholders and then separately inserting entities. This workflow reduced format drift, improved linguistic consistency, and significantly increased the overall reliability of the generated dataset.

4. Few-Shot Relation Extraction

Relation extraction in German legal texts faces two persistent challenges: 1. the scarcity of annotated corpora due to privacy constraints and high annotation costs, and 2. the linguistic complexity of the domain, including very long sentences, compound nouns, and context-dependent relations. These conditions make conventional supervised learning approaches impractical.

Few-shot learning offers an alternative by enabling models to generalise from a few labeled examples. Instead of requiring large-scale annotated datasets, models are trained to rapidly adapt to new relation types under minimal supervision. This paradigm is relevant for the legal domain, where new relation categories may appear across cases and annotated data is expensive to produce.

The dataset includes six relation types that serve as the classification targets in our few-shot experiments (see Table 1). We focus on two complementary few-shot paradigms: **Model-Agnostic Meta-Learning** (Finn et al., 2017), representing optimisation-based meta-learning, and **Prototypical Networks with Contrastive Loss** (Snell et al., 2017; Khosla et al., 2020), representing metric-based learning. Comparing these approaches allows us to investigate whether *fast adaptation through gradient-based updates* (MAML) or *robust embedding spaces with prototype separation* (Proto-Contrastive) is more effective for legal relation extraction.

Relation	Description	Entity Types (Head → Tail)
FiledLawsuitAgainst	One party files a legal action against another.	PER/ORG/UN → PER/ORG/UN
AccusedOf	A person or organization is accused of violating a law or regulation.	PER/ORG/UN → GS/VO
RepresentedBy	A person or organization is represented by a lawyer or another entity.	PER/ORG/UN → AN/ORG
OccurredOn	A legal event or contract occurred on a specific date.	VT/RS → DAT
FineToBePaid	A person or organization must pay a monetary fine.	PER/ORG/UN → GLD
LocatedAt	An entity is located in a city or country.	PER/ORG/UN/GRT → ST/LD

Table 1: Overview of relation types defined in the synthetic German legal relation extraction dataset, including their semantic descriptions and corresponding entity pairs.

4.1. Model 1: MAML for RE

MAML (Finn et al., 2017) is an optimisation-based meta-learning framework designed to enable rapid adaptation to new tasks from only a few examples. Instead of training a model for a single task, MAML optimises its parameters across a distribution of tasks, producing an initialisation that can be fine-tuned with minimal gradient updates.

In the context of RE, each episode corresponds to an N-way K-shot classification task over legal relation types. Given a support set S_i and query set Q_i for task T_i, the model parameters θ are adapted via gradient descent in the inner loop:

$$\theta_i' = \theta - \alpha \nabla_\theta L_{S_i}(f_\theta), \tag{1}$$

where α is the inner learning rate. The adapted parameters θ_i' are then evaluated on the query set, and the outer loop updates the initialisation θ to improve generalisation across tasks:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_i L_{Q_i}(f_{\theta_i'}), \tag{2}$$

with β denoting the meta learning rate.

To better align with the semantic subtleties of legal relations, we extend MAML with **description-aware prototypes** (Borchert et al., 2024). Human-readable textual descriptions of each relation are encoded with a pretrained BERT model and fused with the support-based prototypes:

$$\tilde{p}_r = \lambda \cdot p_r^{\text{support}} + (1 - \lambda) \cdot p_r^{\text{desc}}, \tag{3}$$

where $\lambda \in [0, 1]$ balances observed examples and semantic priors. This enrichment grounds prototypes in domain knowledge, guiding the inner-loop adaptation toward more meaningful parameter updates.

The key advantage of MAML is its *ability to quickly adapt* to unseen relation types with minimal supervision. However, this comes at the cost of higher computational complexity due to two-loop optimisation and reduced interpretability compared to prototype-based approaches.

4.2. Model 2: Prototypical Networks with Contrastive Loss and Curriculum-Aware Extension

Prototypical Networks (Snell et al., 2017) represent a metric-based approach to few-shot learning. Each relation class is characterised by a prototype, computed as the mean embedding of its support examples:

$$p_r = \frac{1}{K} \sum_{i=1}^{K} f(x_i^r), \tag{4}$$

where $f(\cdot)$ denotes the encoder (BERT-based in our case) and K is the number of support examples. Query samples (Q) are classified according to their distance to the prototypes, typically via negative squared Euclidean similarity.

In practice, the similarity function $\text{sim}(\cdot)$ is defined flexibly and can be instantiated using different measures during training-most commonly either negative squared Euclidean distance or cosine (dot-product) similarity, depending on the model configuration.

To improve discriminability in legal RE, we integrate a **supervised contrastive loss** (Khosla et al., 2020). For a query embedding q_j and its true prototype p_{y_j}, the contrastive component encourages proximity to the correct class while pushing away other classes:

$$L_{\text{contrast}} = -\frac{1}{|Q|} \sum_{j=1}^{|Q|} \log \frac{\exp(\text{sim}(q_j, p_{y_j})/\tau)}{\sum_{r \in C} \exp(\text{sim}(q_j, p_r)/\tau)}, \tag{5}$$

where τ is a temperature parameter and C is the set of relation classes in the episode. The final loss combines cross-entropy and contrastive terms:

$$L = L_{\text{CE}} + \gamma \cdot L_{\text{contrast}}, \tag{6}$$

with γ controlling the contrastive weight.

Building on this, we introduce a curriculum-aware prototype enrichment that leverages human-readable relation descriptions (Wang et al., 2022;

Yang et al., 2020). Each relation r is associated with a set of textual descriptions D_r. These are encoded into embeddings d_r and combined with support-based prototypes:

$$\tilde{p}_r = \lambda \cdot p_r + (1 - \lambda) \cdot d_r, \tag{7}$$

where λ adapts over training phases. Early phases emphasise semantic priors from descriptions, while later phases rely more heavily on support instances. To structure learning, we adopt a curriculum schedule that gradually exposes increasingly complex descriptions (e. g., easiest $\rightarrow$ medium $\rightarrow$ hardest).

This *fusion of statistical and symbolic knowledge* stabilises training, reduces confusion between semantically similar relations, and grounds prototypes in domain semantics. The resulting model is both *efficient at inference* (no fine-tuning required) and *interpretable* through enriched prototypes.

The two paradigms explored in this work, MAML and Prototypical Networks with Contrastive Loss, address few-shot RE from fundamentally different perspectives. MAML learns a parameter initialisation that can be rapidly fine-tuned for new relation types, which makes it particularly appealing in principle for situations where *fast adaptation* to unseen classes is required. In contrast, Prototypical Networks emphasise the structure of the embedding space, classifying queries by their distance to class prototypes. When enhanced with supervised contrastive loss and curriculum-aware prototype enrichment, this approach is designed to achieve *robust separation* between semantically similar relations.

Rather than directly testing domain-specific use cases such as adaptive learning speed or semantic overlap, our evaluation compares both methods within a standardised episodic setup (N-way K-shot), allowing a fair comparison of their behavior under identical few-shot conditions. This design highlights their relative performance in the legal RE context without claiming a full analysis of model strengths and limitations. Conceptually, the comparison emphasises the contrast between *optimisation-based adaptation* (MAML) and *metric-based embedding separation* (Proto-Contrastive), providing initial evidence on how these complementary strategies transfer to the German legal domain.

5. Experiments

5.1. Experimental Setup

All experiments were conducted using the synthetic German legal RE dataset described in Section 3. The dataset was split into training, validation, and test partitions with balanced coverage across the six relation types. Following the FewRel protocol, models were evaluated under an episodic N-way K-shot setup. Each episode consists of $N = 6$ relation classes, $K \in \{1, 3, 5\}$ labeled support examples per class, and $Q = 5$ query examples. Performance was measured on 100 randomly sampled test episodes per configuration, reporting micro-averaged precision, recall, and F_1-score.

Models were implemented in PyTorch using Hugging Face Transformers, with training conducted on a single NVIDIA A100 GPU. We employed a general-domain German BERT (dbmdz/bert-base-german-uncased) and a legal-domain BERT encoder (elenanereiss/bert-german-ler). Hyperparameters were tuned on the validation split:

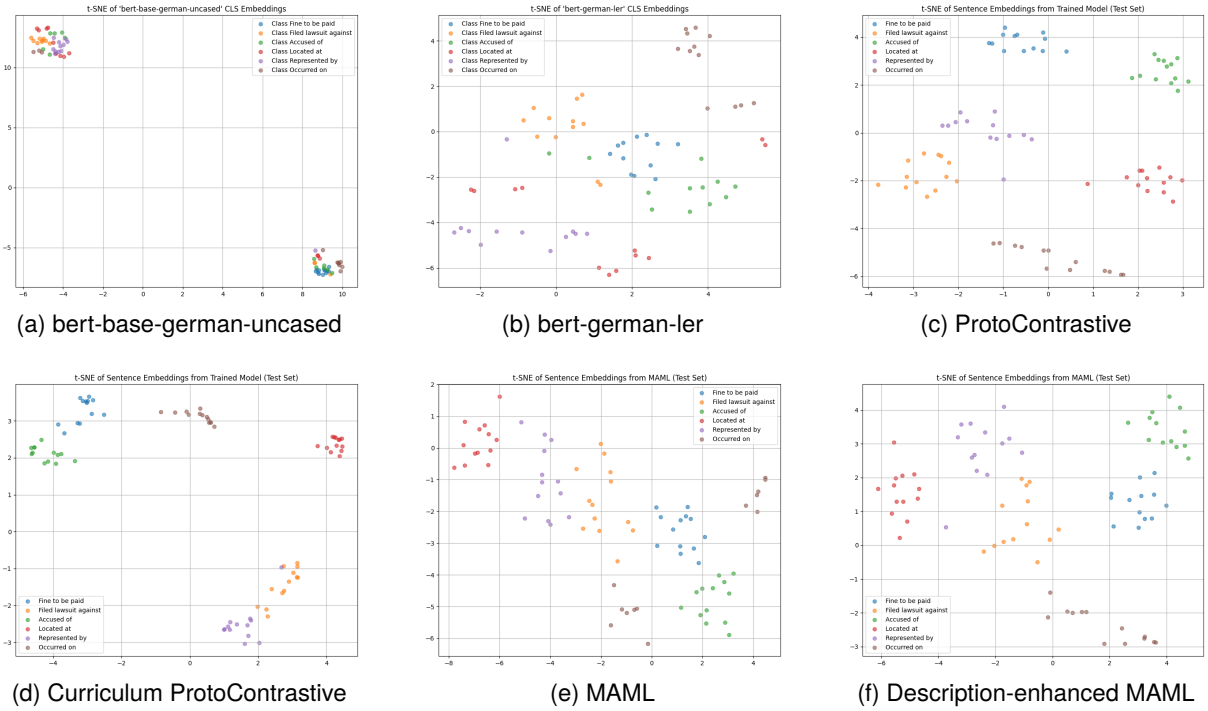
- **Optimizer:** Adam with learning rate $1e{-}5$ and gradient clipping at 5.0.
- **Episode batch size:** 2 episodes per batch.
- **Iterations:** 300–360 training steps with validation every 20 steps.
- **Evaluation metric:** Micro-averaged F_1 across episodes.

As baselines, we implemented 1. a zero-shot embedding similarity model using cosine similarity over sentence embeddings, and 2. a vanilla Prototypical Network without contrastive loss. These baselines illustrate the inherent difficulty of relation extraction in the legal domain and provide a reference point for evaluating our few-shot extensions. The embedding similarity models achieved only 21.0%, 18.0%, and 5.0% F_1 for bert-german-ler, gbert-legal-ner, and bert-base-german-uncased, respectively, underscoring the limited zero-shot generalisation of pretrained models in this specialised domain. In contrast, the *vanilla Prototypical Network* reached a substantially higher F_1 score of 89.66% in the 5-shot setting, establishing a strong lower bound for the meta-learning and contrastive variants proposed in this work.

5.2. MAML

We evaluated the model-agnostic meta-learning framework under 1-, 3-, and 5-shot conditions. The inner loop learning rate α was set to 0.01, with the number of gradient update steps varied between 1, 3, and 5. We further tested the integration of semantic relation descriptions through description-aware prototypes, controlled by a fusion weight $\alpha \in \{0.5, 1.0\}$.

Results show that MAML with Legal BERT consistently outperforms the general-domain encoder, with the best configuration (3-shot, $\alpha = 0.5$) achieving an F_1 of 97.41%. Interestingly, increasing the number of inner-loop updates did not improve results: a single update step yielded slightly better stability (97.17% F_1) than 3 or 5 updates, suggesting that excessive adaptation risks overfitting in low-resource scenarios. Description fusion improved



(a) bert-base-german-uncased (b) bert-german-ler (c) ProtoContrastive

(d) Curriculum ProtoContrastive (e) MAML (f) Description-enhanced MAML

Figure 3: Visualisation of sentence embeddings (via t-SNE) generated by different encoder-model combinations, illustrating how few-shot learning methods and domain-specific encoders influence the clustering and separation of relation classes.

performance in most cases, confirming that semantic priors help guide adaptation to unseen relations.

5.3. Prototypical Networks with Contrastive Loss

The Prototypical Network baseline provided a strong foundation, achieving up to 91.9% accuracy under 3-shot conditions. To improve discriminability, we extended it with a supervised contrastive loss. This Proto-Contrastive model benefited greatly from domain-adapted encoders: with Legal BERT, it reached an F_1 of 97.75% at 3-shot, outperforming both vanilla prototypes and MAML under the same conditions. Even in the 1-shot scenario, Proto-Contrastive with Legal BERT attained 94.75% F_1, demonstrating high data efficiency.

We further experimented with a curriculum-aware extension that fused prototypes with relation description embeddings of increasing complexity. This approach proved particularly beneficial when using general-domain BERT, where curriculum-guided semantic grounding compensated for weaker domain adaptation. The best configuration (5-shot, general-domain BERT) achieved an F_1 of 99.83%, surpassing all other models, including those using Legal BERT.

5.4. Discussion

The comparative evaluation suggests complementary tendencies between optimisation- and metric-based approaches. MAML demonstrates the ability to adapt with few gradient updates, particularly when guided by description fusion, although this advantage is theoretical rather than explicitly measured. Its two-loop optimisation, however, increases computational cost and can lead to instability under low-resource conditions.

In contrast, Prototypical Networks with supervised contrastive loss offer simpler training and inference, as no fine-tuning is required at test time. Their performance is consistently strong, particularly with domain-specific encoders. When combined with curriculum-aware description enrichment, Proto-Contrastive not only bridges the gap to legal-domain encoders but can even surpass them. This suggests that structured semantic guidance is a powerful complement to domain pretraining.

To better understand how different encoders and few-shot models structure the representation space, we visualise the learned embeddings using t-SNE on test-set instances (Figure 3). The embedding spaces produced by MAML, Proto-Contrastive, and Curriculum-aware Proto-Contrastive models exhibit clear class-wise clustering, indicating that semantic separation improves progressively with the introduction of contrastive and curriculum-based learning.

Overall, the experiments demonstrate that few-shot RE in German legal texts benefits from both semantic priors and domain adaptation. MAML offers an adaptable optimisation perspective within the few-shot paradigm, yet the metric-based Proto-Contrastive models achieve a more favorable balance between efficiency, stability, and accuracy. The prototype enrichment offers a promising direction for improving interpretability in future analyses.

Model / Setting	**Value (%)**
Prototypical Network – General BERT	89.66
MAML – General BERT	96.50
MAML – Legal BERT ($\alpha = 0.5$)	97.41
ProtoContrastive – General BERT	93.25
ProtoContrastive – Legal BERT	97.00
ProtoContrastiveCurriculum – General	**99.83**
ProtoContrastiveCurriculum – Legal	97.33

Table 2: Comparison of models with few-shot techniques (F_1, 6-way-5-shot learning).

6. Conclusions

This article addressed the challenge of relation extraction in German legal texts under low-resource conditions. We introduced a synthetic dataset generation pipeline that combines LLM-driven template expansion with curated entity substitution, producing privacy-preserving and reproducible training data in FewRel-compatible format. Despite its modest size of 250 sentences, the dataset fills a crucial gap in German legal NLP, where annotated corpora remain scarce due to privacy and annotation costs. The pipeline can easily be used to generate more sentences for the six relations or for other relations.

On the modeling side, we evaluated two complementary few-shot paradigms: optimisation-based meta-learning with MAML and metric-based Prototypical Networks extended with supervised contrastive loss and curriculum-aware prototype enrichment. Our experiments demonstrated that 1. domain-specific encoders such as Legal BERT substantially improve performance, 2. semantic priors from relation descriptions guide models toward more robust generalisation, and 3. curriculum-aware Proto-Contrastive models can even outperform domain-adapted encoders, achieving near-perfect few-shot performance.

These results show that integrating structured knowledge with few-shot learning is a promising direction for legal NLP. The dataset, code, and models will be released as open source material to support reproducibility and further research.

Future work will focus on extending the system to multi-relation sentences and validating performance on authentic case law to assess robustness under naturally noisy conditions, in contrast to the current dataset where every sentence expresses a valid relation and does not include a "no relation" (NOTA) class. Moreover, the synthetic sentences are largely noise-free and grammatically well-formed, which may contribute to the near-perfect performance observed in controlled evaluation. Another possible direction is to extend the approach to *document-level relation extraction*, allowing relations that span multiple sentences or sections of a legal document to be identified. This would provide a more comprehensive representation of legal contexts and improve the model's ability to capture long-range dependencies in complex texts.

7. Ethical Considerations

Legal texts often contain sensitive personal data, which raises substantial privacy and compliance concerns, especially under regulations such as GDPR. To mitigate these risks, we deliberately avoid using authentic case data in our experiments. Instead, we construct a fully synthetic dataset through template-driven generation combined with curated entity substitution. All entities are injected at the substitution stage, ensuring that no real individuals, organisations, or legal cases are reproduced. This approach guarantees that the dataset is privacy-preserving and resistant to memorisation artifacts from LLMs.

Bias mitigation was an additional goal. Curated entity lists were constructed to include diverse personal names, organisational types, and statutory references, reflecting demographic variety and reducing systematic biases. Random substitution further prevents over-representation of specific entities or patterns, supporting fairer generalisation.

Despite these safeguards, limitations remain. Synthetic data cannot fully capture the ambiguity and complexity of authentic legal language, and performance on real-world corpora may differ. Furthermore, the current setup assumes a single relation per sentence and has not been validated with legal professionals. We encourage future work to involve expert-in-the-loop evaluation and to explore multi-relation and document-level extraction.

All dataset generation code and models will be released as open resources to support transparency, reproducibility, and responsible reuse.

8. Bibliographical References

Manzoor Ali, Muhammad Sohail Nisar, Muhammad Saleem, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2024. Enhancing relation

extraction through augmented data: Large language models unleashed. In *Natural Language Processing and Information Systems (NLDB)*, volume 14327 of *Lecture Notes in Computer Science*, pages 68–78. Springer.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges.

Mehmet Aydar, Ozge Bozal, and Furkan Ozbay. 2020. Neural relation extraction: A survey. Technical report, arXiv preprint arXiv:2007.04247.

Matthias Aßenmacher, Alessandra Corvonato, and Christian Heumann. 2021. Re-evaluating germeval17 using german pre-trained language models. In *Proceedings of the 6th Swiss Text Analytics Conference (SwissText)*, volume 2957, page –, Winterthur, Switzerland (Online).

Sarah T. Bachinger, Leila Feddoul, Marianne Jana Mauch, and Birgitta König-Ries. 2024. Extracting legal norm analysis categories from german law texts with large language models. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, dg.o '24, page 481–493, New York, NY, USA. Association for Computing Machinery.

Philipp Borchert, Jochen De Weerdt, Kristof Coussement, Arno De Caigny, and Marie-Francine Moens. 2023. Core: A few-shot company relation classification dataset for robust domain adaptation. *arXiv preprint arXiv:2310.12024*.

Philipp Borchert, Jochen De Weerdt, and Marie-Francine Moens. 2024. Efficient information extraction in few-shot relation classification through contrastive representation learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 638–646, Mexico City, Mexico. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2021. Exploring semantic similarity between german legal texts and referred laws. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 37–50. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deutscher Ethikrat et al. 2023. Mensch und maschine–herausforderungen durch künstliche intelligenz. *Stellungnahme. Berlin*, 14:2023.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135. JMLR.org.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Sreyan Ghosh, Chandra K. Evuru, Sonal Kumar, S. Ramaneswaran, Sakshi S., Utkarsh Tyagi, and Dinesh Manocha. 2023. Dale: Generative data augmentation for low-resource legal nlp. In *Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL)*.

John Giorgi, Xindi Wang, Nicola Sahar, Won Young Shin, Gary D. Bader, and Bo Wang. 2019. End-to-end named entity recognition and relation extraction using pre-trained language models. *arXiv preprint arXiv:1912.13415*.

Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021. Sentence boundary detection in german legal documents. In *ICAART 2021 - Proceedings of the 13th International Conference on Agents*

and Artificial Intelligence, ICAART 2021 - Proceedings of the 13th International Conference on Agents and Artificial Intelligence, pages 812–829. SciTePress. Publisher Copyright: © 2021 by SCITEPRESS - Science and Technology Publications, Lda.; 13th International Conference on Agents and Artificial Intelligence, ICAART 2021 ; Conference date: 04-02-2021 Through 06-02-2021.

Ingo Glaser, Bernhard Waltl, and Florian Matthes. 2018. Named entity recognition, extraction, and linking in german legal contracts. *Jusletter IT*, (February). Publisher Copyright: © 2018 Editions Weblaw. All rights reserved.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4803–4809, Brussels, Belgium.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics (TACL)*, 10:507–523.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Comput. Surv.*, 54(4).

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *CoRR*, abs/2004.11362.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Lishuang Li, Ruiyuan Lian, Hongbin Lu, and Jingyao Tang. 2022. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2098–2107, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shiye Li and Li Yi. 2024. A few-shot entity relation extraction method in the legal domain based on large language models. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '24, page 580–586, New York, NY, USA. Association for Computing Machinery.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. 2020. Part-aware prototype network for few-shot semantic segmentation. In *Computer Vision – ECCV 2020*, pages 142–158, Cham. Springer International Publishing.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1105–1116.

Silvio Peikert, Celia Birle, Jamal Al Qundus, Le Duyen Sandra Vu, and Adrian Paschke. 2022. Extracting references from german legal texts using named entity recognition. In *International Conference on Legal Knowledge and Information Systems 2022*.

Nicholas Popovic and Michael Färber. 2022. Few-shot document-level relation extraction. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5733–5746, Seattle, United States. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024. Prompting-based synthetic data generation for few-shot question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13168–13178, Torino, Italia. ELRA and ICCL.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mengru Wang, Jianming Zheng, and Honghui Chen. 2022. Taxonomy-aware prototypical network for few-shot relation extraction. *Mathematics*, 10(22).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hui Wu, Yuting He, Yidong Chen, Yu Bai, and Xiaodong Shi. 2024. Improving few-shot relation extraction through semantics-guided learning. *Neural Networks*, 169:453–461.

Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance prototypical network with text descriptions for few-shot relation classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2273–2276, New York, NY, USA. Association for Computing Machinery.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11).

A. Prompt Specification for Template-Based Synthetic Data Generation

Use Placeholder Entities: Generate Formal Legal Sentences in German:

Write sentences in formal, legal German, using a style that reflects the tone and structure of actual legal documents. The sentences should describe specific relationships between legal entities using placeholder names.

Represent different entity types with placeholders to keep the sentences generic and adaptable. Use the following placeholders:

- *Personen* (Individuals): *PER*
- *Anwälte* (Lawyers): *AN*
- *Organisationen* (Organizations): *ORG*
- *Unternehmen* (Companies): *UN*
- *Gerichte* (Courts): *GRT*
- *Länder* (Countries): *LD*
- *Städte* (Cities): *ST*
- *Gesetze* (Laws): *GS*
- *Verordnungen* (Regulations): *VO*
- *Verträge* (Contracts): *VT*
- *Rechtsprechungen* (Judicial Rulings): *RS*
- *Datum* (Dates): *DAT* for any specific date
- *Geldbetrag* (Monetary Amounts): *GLD* for any fine or monetary amount

Relations and Sentence Examples: Create sentences for each of the following relations. Use placeholders correctly and write sentences that resemble real legal statements.

- *Located at (Ansässig in)* Entities: *PER/UN/ORG/GRT - LD/ST* Prompt: *"Erstellen Sie einen Satz, in dem PER in der Stadt ST ansässig ist."* Output: *"PER ist in ST ansässig und übt dort seine berufliche Tätigkeit aus."*
- *Fine to be paid (Bußgeld zu zahlen)* Entities: *PER/UN/ORG - GLD* Prompt: *"Schreiben Sie einen Satz, in dem UN zur Zahlung eines Bußgeldes von GLD verurteilt wird."* Output: *"UN wurde zur Zahlung eines Bußgeldes von GLD verurteilt, da gegen geltende Umweltauflagen verstoßen wurde."*
- *Accused of (Angeklagt wegen)* Entities: *PER/UN/ORG - GS/VO* Prompt: *"Erstellen Sie einen Satz, in dem PER gemäß GS angeklagt ist."* Output: *"PER wird gemäß GS der vorsätzlichen Missachtung gesetzlicher Vorschriften beschuldigt."*
- *Represented by (Vertreten durch)* Entities: *PER/UN/ORG - AN/PER/UN* Prompt: *"Erstellen Sie einen Satz, in dem ORG durch AN vertreten wird."* Output: *"ORG wird durch AN in allen rechtlichen Angelegenheiten vertreten."*
- *Filed lawsuit against (Klage eingereicht gegen)* Entities: *PER/UN/ORG - PER/UN/ORG* Prompt: *"Erstellen Sie einen Satz, in dem UN eine Klage gegen UN beim Gericht GRT einreicht."* Output: *"UN hat eine Klage gegen UN beim GRT eingereicht und verlangt Schadensersatz für Vertragsverletzungen."*
- *Occurred on (Ereignete sich am)* Entities: *VT/RS - DAT* Prompt: *"Erstellen Sie einen Satz, in dem das Ereignis VT am DAT stattfand."* Output: *"VT ereignete sich am DAT und führte zu einer rechtlichen Auseinandersetzung."*

Sentence Structure and Tone:

Use passive language and conditional clauses common in legal texts (e.g., *falls*, *sofern*, *aufgrund von*). Incorporate complex sentence structures, such as subordinate or conditional clauses.

Legal Context and Terminology:

Use legal terms and concepts commonly found in German law, such as *Bußgeld*, *Verordnung*, *Klage*, and references to sections (e.g., *gemäß § 21 BGB*). Ensure that all relationships and terms are coherent with legal scenarios.

Example Generation Guide:

When prompted, generate a German sentence using the placeholders and relation. Respond only with the generated sentence in German, following these instructions.

Response Format:

Return the output as JSON containing `text`, `entities`, and `relations` fields. Each entity must include `text` (the placeholder), `type`, and character span. Each relation must include `type`, `head`, and `tail`.

Example:

```
{
  "text": "Der Beispieltext auf
     Deutsch mit Entitaeten.",
  "entities": [
```

```
    { "text": "entity1", "type":
      "Entity_type", "start": 0,
       "end": 5 },
    { "text": "entity2", "type":
      "Entity_type", "start": 10
      , "end": 15 }
  ],
  "relations": [
    { "type": "Relation_type", "
      head": "entity1", "tail":
      "entity2" }
  ]
}
```

Create multiple sentence variations for each relationship type, using slightly different structures and vocabulary to increase dataset diversity. Encourage sentences that reflect unique scenarios, such as joint lawsuits, fines for multiple violations, or cases involving multiple jurisdictions. Review and Post-Processing:

Generate sentences that can be easily reviewed and adapted for use in a structured dataset, with each sentence containing the necessary placeholders to be replaced programmatically.