Enhancing Legal Reasoning in Pre-trained Language Models via Knowledge Graph-Guided Multi-Task Pre-training

Anonymous ACL submission

Abstract

Legal language understanding requires reasoning over semi-structured regulatory documents, yet standard pre-training objectives do not explicitly target this structural capability. Because current legal language models rely solely on masked language modeling, they treat statutory text as flat token sequences and ignore the underlying hierarchical constraints. To bridge this gap, we present a multi-task pre-training approach for German traffic law (*StVO*). Our approach combines three complementary objectives: Next Sentence Prediction (NSP), Knowledge Graph (KG) Guided Clustering (CLU), and Knowledge Graph (KG) Aware Paragraph Classification (CLS). This approach explicitly encodes structured legal knowledge directly into the model weights. This encoding enables zero-shot legal reasoning at inference time and removes the need for discrete graph traversal. We evaluate our approach across nine Transformer-based encoders and 324 experimental runs. On synthetic *StVO* question answering, our models demonstrate consistent relative improvements between 10.3% and 155.0% over baselines. Furthermore, our architecture establishes new state-of-the-art (SOTA) results on four legal benchmarks. It also achieves consistent zero-shot transfer gains between 3.4% and 15.8% on Austrian driving regulations. These experimental results indicate that our training approach generalizes successfully beyond the initial training legal domain.

1 Introduction

Legal Natural Language Processing (NLP) models are considered a base component for Reliable Artificial Intelligence (rAI) within semi-structured resources such as statutory books and contractual texts (Yao et al., 2024). A persistent challenge involves the trade-off between predictive accuracy and interpretability; as SOTA Pre-trained Language Models (PLMs) achieve high performance, they remain largely unexplainable; conversely, rule-based methods offer transparency but suffer from flexibility (Zhong et al., 2020). Despite recent proposals targeting this balance (An and Zhang, 2024; Wang et al., 2024), the gap between empirical performance and verifiable legal reasoning persists as a critical bottleneck (Zhong et al., 2020).

Existing legal domain-adaptive pre-training approaches—including Lawformer (Xiao et al., 2021), SAILER (Li et al., 2023), and Legal-BERT (Chalkidis et al., 2020)—rely on implicit structural modeling through long-range attention, structural headers, or domain-specific MLM, respectively. None *explicitly* incorporates the formal semantic relationships and hierarchical constraints of the semi-structured legal texts during pre-training. To resolve this limitation, we introduce a training approach that explicitly injects knowledge graphs into the embedding space of PLMs.

In this work, we propose a multi-task domain-adaptive pre-training approach validated in the German Road Traffic Regulations (*StVO*) as representative for the semi-structured legal texts, encoding legal structure via three complementary objectives: Next Sentence Prediction (NSP) for document coherence, Knowledge Graph-guided Clustering (CLU) for relational structure, and KG-aware paragraph Classification (CLS) for task-aligned reasoning.

Our main contributions in this work are:

- We introduce a unified multi-task pre-training approach for semi-structured regulatory text that explicitly integrates KG structures into latent representations, unlike existing text-concatenation approaches.
- We employ TransE-derived legal paragraph embeddings as *soft* pre-training targets, enabling the transfer of relational legal knowledge without hard classification constraints and allowing

the model to learn a continuous range of legal concepts.

- A differentiable gated KG-aware attention training mechanism that performs real-time latent-space retrieval from a KG paragraph embeddings library, dynamically weighting external legal context against internal model representations. Unlike traditional retrieval-augmented approaches, which require discrete external database lookups during inference.
- A synthetic German traffic law QA dataset ($N = 12{,}554$) with paragraph-level citation grounding and strict paragraph-level splits to prevent data leakage, providing a new benchmark for citation-accurate legal reasoning.

2 Related Work

Legal Language Models. Domain-specific Transformer adaptations have shaped legal NLP. Chalkidis et al. (2020) established *Legal-BERT* via continued pre-training on English legal corpora. Xiao et al. (2021) introduced *Lawformer* for long-range dependencies. Li et al. (2023) proposed *SAILER* for structure-aware dense retrieval. These models rely on raw linguistic sequences and do not explicitly incorporate structured relational knowledge during pre-training. They remain structurally blind to underlying semantic dependencies. On the evaluation side, LexGLUE (Chalkidis et al., 2021) standardized assessment across seven English legal tasks, extended by Chalkidis et al. (2022) for fairness and Santosh et al. (2024) for summarization. LegalBench (Guha et al., 2023) provides legal English QA samples but lacks German regulatory coverage and paragraph citation sourcing.

Knowledge Graph-Enhanced Embeddings. The transition toward neuro-symbolic integration has driven structure-aware pre-training in legal NLP (Zhang et al., 2019; Li et al., 2023). Yao et al. (2019) proposed *KG-BERT*, treating triples as text sequences to leverage linguistic patterns for graph completion. Xie et al. (2016) and Alam et al. (2025) argue that standard embeddings overlook schematic information and logical axioms, suggesting ontology-enriched representations for deep semantic understanding. Tang et al. (2024) introduced *GraphGPT*, aligning structural graph patterns with natural language via dual-stage instruction tuning and Chain-of-Thought distillation for zero-shot relational reasoning. We build our approach on **TransE** (Bordes et al., 2013; Wang et al., 2014) specifically for its efficiency in preserving hierarchical topology as soft targets for the CLU objective. In preliminary experiments, we evaluated a diverse range of KG embeddings—including bilinear models (**DistMult**, (Yang et al., 2014)), complex-valued embeddings (**ComplEx**, (Trouillon et al., 2016)), and rotational models (**RotatE**, (Sun et al., 2019))—with **TransE** yielding the most robust performance on downstream tasks. This approach bypasses the computational costs typically associated with Transformer-based graph encoders (Yao et al., 2019).

Multi-Task and Knowledge Distillation. Typical **knowledge distillation** (Hinton et al., 2015) focuses on model compression; however, we adapt this paradigm to distill structured KG graph topology into *soft pre-training targets*. Furthermore, we leverage **Multi-Task Learning** (MTL) frameworks (Caruana, 1997; Kendall et al., 2018) to treat text-based modeling and graph-based understanding as synergistic objectives. Consequently, our work extends beyond traditional text augmentation by proposing an architecture in which model embeddings are intrinsically aligned with the structured logical hierarchies of a legal knowledge graph.

3 Methodology

Our methodology, depicted in Figure 1, follows a hierarchical progression across two converging pipelines.

- **The KG Pipeline:** This workflow transforms raw *StVO* legal text into a formal **Knowledge Graph** via the legal pattern-based parser (Siddig et al., 2025). These relationships are mapped into a latent translation space using **TransE**, generating paragraph embeddings ($\mathbf{e}_p$) that capture the topological core of the legal code.
- **The LM Pipeline:** This workflow begins with a **pre-trained base encoder**. It then enters a **domain-adaptive** stage, where the model undergoes **KG-guided multi-task pre-training**. In this stage, the model internalizes structured legal logic before proceeding to final **Evaluations**.

By synthesizing these two pipelines, the architecture moves beyond simple surface-level text patterns, ensuring that the underlying regulatory struc-

ture of German traffic law (*StVO*) is fundamentally injected into the model's representations.

3.1 Problem Formulation

Given *StVO* legal document $\mathcal{D}$, we train representations that enable accurate legal reasoning on question-answering. Formally, for a legal question and a possible answer pair represented in a sequence of words (q, a) and its source legal paragraph p, we learn a function $f : (q, a, p) \rightarrow \{0, 1\}$ indicating whether the answer is legally correct given the paragraph. The training approach incorporates three complementary objectives in Section 3.4: (1) NSP for document coherence, (2) CLU for KG-guided semantic clustering, and (3) CLS for task-aligned binary classification.

3.2 TransE Paragraph Embeddings Generation

We parse *StVO* into a knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ via the parser (Siddig et al., 2025) exploiting the semi-structured hierarchy of the German traffic law (categories, paragraphs, sub-paragraphs, sentences, tables, sections, signs). The graph contains **554 nodes** and **945 edges** across five different relations: HAS_PARAGRAPH, HAS_SENTENCE, HAS_SECTION, HAS_SIGN, HAS_TABLE. We train TransE to obtain Paragraph-level embeddings $\mathbf{e} \in \mathbb{R}^{500}$ serve as *soft clustering targets*, preserving relational structure and enabling gradient flow through cosine similarity in the clustering training objective in Section 3.4.

3.3 Hierarchical Document Encoding

To handle the structural complexity and length of the legal sentences in the *StVO*, we adopt the idea of a two-level hierarchical architecture (Yang et al., 2016; Chalkidis et al., 2019). Each legal sentence is decomposed into an ordered set of segments $\{s_1, \ldots, s_K\}$ by our own heuristic structure-aware segmenter. First, a shared **Segment Encoder** $\mathcal{E}_{\text{base}}$ processes individual segments s_k to produce local embeddings: $\mathbf{h}_k = \mathcal{E}_{\text{base}}(s_k)$. Second, a **Document Encoder** $\mathcal{E}_{\text{doc}}$ aggregates the sequence $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$ to capture global dependencies, yielding the final document representation $\mathbf{z}_d = \text{MaxPool}(\mathcal{E}_{\text{doc}}(\mathbf{H}))$. This design preserves intra-document structural context while remaining computationally efficient for extensive statutory sequences.

3.4 Multi-Task Pre-Training Objectives

Task 1 — Next Sentence Prediction (NSP) we employ domain-adapted Next Sentence Prediction (NSP) (Devlin et al., 2019). Positive pairs are drawn from consecutive sentences within the same legal paragraph, while negative pairs are sampled from random cross-paragraph sequences. Given a pair of legal sentences (s_i, s_j) and a binary label $y \in \{0, 1\}$ indicating whether s_j is the actual successor of s_i, the objective is to minimize the binary cross-entropy loss:

$$\mathcal{L}_{\text{NSP}} = -\sum [y \log(P(y = 1|s_i, s_j)) + (1 - y) \log(1 - P(y = 1|s_i, s_j))] \quad (1)$$

where $P(y = 1|s_i, s_j)$ represents the probability predicted by the model that the sentences are consecutive.

Task 2 — KG-Guided Clustering (CLU). For each legal segment s, the hierarchical encoding produces a representation $\mathbf{z}_s \in \mathbb{R}^d$. TransE pre-training has already positioned every paragraph p at a learned coordinate $\mathbf{e}_p \in \mathbb{R}^{500}$ in a continuous space where structurally related paragraphs (e.g., § 8 and § 9 on right-of-way) cluster together. The CLU objective exploits this geometry as a supervision signal via margin-based contrastive learning (Schroff et al., 2015; Wang et al., 2014):

$$\mathcal{L}_{\text{CLU}} = \underbrace{1 - \cos(\mathbf{z}_s, \mathbf{e}_{c^+})}_{\text{pull toward source paragraph}} + \underbrace{\max(0, \cos(\mathbf{z}_s, \mathbf{e}_{c^-}) + m)}_{\text{push from unrelated paragraph}} \quad (2)$$

where $c^- \sim \text{Uniform}(\mathcal{P} \setminus \{c^+\})$ samples a random negative from the 59 *StVO* paragraphs. The *positive term* minimises the cosine distance between $\mathbf{z}_s$ and the KG embedding of its true source paragraph c^+, anchoring the encoder to the correct region. The *negative term* is a hinge loss that fires only when an unrelated paragraph c^- lies within margin m of $\mathbf{z}_s$, penalising insufficient separation while ignoring already well-separated negatives. Crucially, because $\mathbf{e}_{c^+}$ is a 500-dimensional continuous vector encoding structural relationships of the KG, the gradient carries relational geometry rather than discrete isolated targets.

Task 3 — KG-aware Paragraph Classification (CLS) Standard classification models rely

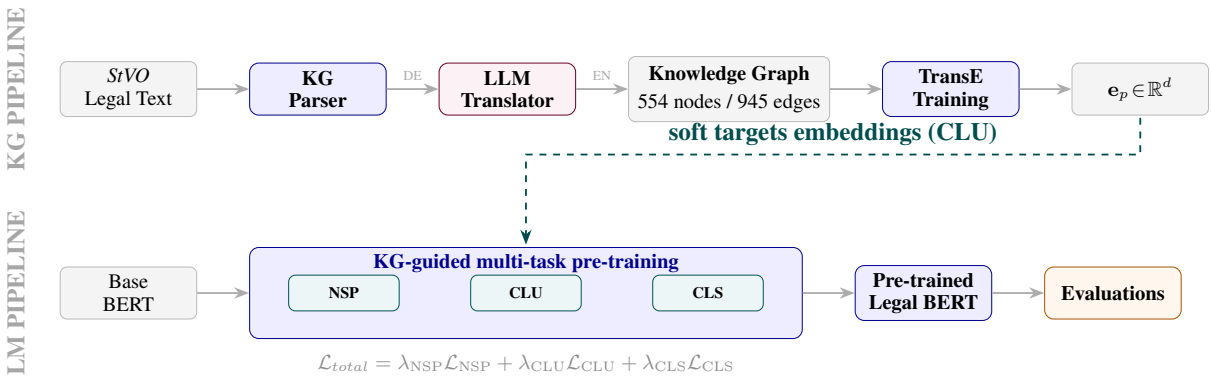


Figure 1: Architecture of the proposed KG-guided multi-task pre-training pipeline. German *StVO* legal text is parsed, translated via an LLM translator, and structurally aligned with a corresponding English-based knowledge graph to inject explicit relational constraints into model reasoning. A detailed visualization of the generated graph topology is provided in Appendix A.

solely on the self-contained encoder representation $\mathbf{z}_s$. We relax this constraint by introducing a **dynamic paragraph attention** training mechanism that consults the *StVO* paragraphs at decision time. Using $\mathbf{z}_s$ as a query, the model computes an attention distribution α over the full set of KG paragraph embeddings $\{\mathbf{e}_j\}$, yielding a KG-context vector $\mathbf{c}_{\text{KG}}$:

$$\alpha_j = \text{softmax}\left(\frac{\mathbf{z}_s^\top \mathbf{e}_j}{\sqrt{d_h}}\right), \quad \mathbf{c}_{\text{KG}} = \sum \alpha_j \mathbf{e}_j \quad (3)$$

To integrate this knowledge, a **gated fusion** step computes $\tilde{\mathbf{z}}_s = g \odot \mathbf{c}_{\text{KG}} + (1-g) \odot \mathbf{z}_s$, where the gate $g = \sigma([\mathbf{z}_s; \mathbf{c}_{\text{KG}}]\mathbf{W}_g)$ adaptively regulates the influence of the KG paragraph embeddings by weighing the semantic compatibility between the input text and the retrieved KG context. Unlike prior work that integrates KG knowledge as static pre-training constraints (Peters et al., 2019) or via token-level entity linking (Zhang et al., 2019), our mechanism operates at the *paragraph level,* routes KG knowledge dynamically per input instance, and gates its influence to zero when the retrieved context is uninformative. This transfers the classification decisions directly to the formal *StVO* during training without requiring graph traversal at inference. The final loss is described as follows:

$$\begin{aligned} \hat{y} &= \sigma(\mathbf{W}_c \tilde{\mathbf{z}}_s + b_c) \\ \mathcal{L}_{\text{CLS}} &= -y \log \hat{y} - (1-y)\log(1-\hat{y}) \end{aligned} \quad (4)$$

The individual objectives target different aspects of the model's capabilities. However, we unify them through a multi-task learning approach. The total objective is defined as:

$$\mathcal{L}_{total} = \lambda_{\text{NSP}}\mathcal{L}_{\text{NSP}} + \lambda_{\text{CLU}}\mathcal{L}_{\text{CLU}} + \lambda_{\text{CLS}}\mathcal{L}_{\text{CLS}} \quad \text{where } \sum \lambda = 1.0 \quad (5)$$

4 Experimental Settings

To evaluate the effectiveness of our approach, we design our experiments around three main questions.

RQ1: To what extent does the semi-structured nature of *StVO* facilitate the transfer of question-answering legal reasoning across languages, and how do domain-specific linguistic features impact this transfer?

RQ2: What are the limits of cross-domain generalizability for models trained on semi-structured *StVO* when applied to foreign driving regulations with varying degrees of overlap?

RQ3: How effectively does *StVO* legal knowledge transfer to other law-related tasks beyond traffic regulations?

4.1 Datasets

German Road Traffic Regulations (*StVO*). The German Road Traffic Regulations (*Straßenverkehrsordnung*;*StVO*) (Bundesministerium für Digitales und Verkehr, 2013) form the core corpus for our KG pre-training approach. The *StVO* is a highly semi-structured, self-contained statute governing public road traffic and operations in Germany. Rather than treating it as flat text, we exploit its hierarchical layout to construct a formal KG (see the visual topology in Appendix A).

Synthetic *StVO* QA Benchmark. Existing legal QA benchmarks (Guha et al., 2023) focus on case law, and lack citation grounding. We use LLM-based generation (Jang et al., 2023; Dai et al., 2023) to produce 12,554 labeled balanced QA pairs. For each *StVO* sentence s_i, we prompt Mistral-7B-Instruct (Jiang et al., 2023) and Llama-3.1-8B (Grattafiori et al., 2024) to generate open-ended questions with paired correct (y_{pos}, source-grounded) and incorrect (y_{neg}, plausible but contradictory) answers. Dataset statistics and representative example are provided in Table 1 and Box 4.1, respectively. To ensure legal reliability, generated pairs pass a six-stage validation pipeline that checks structural and field constraints, rejects non-reasoning yes/no questions, verifies domain-specific terminology (e.g., *Vorfahrt*, right of way), removes trivial paraphrases via lexical-overlap analysis, and detects placeholders; only pairs scoring $Q \geq 0.6$ on a composite metric are retained. Following Lewis et al. (2021), we prevent data leakage via **paragraph-level splitting**. By reserving 17% of paragraph identifiers exclusively for testing, we ensure the model cannot exploit content memorization. This yields a robust split of **10,245 train / 2,309 test** pairs with zero paragraph overlap (Table 1).

Box 4.1: Generated QA Pair — § 8 Right of way — *StVO*

Given the legal sentence from paragraph § 8 (*StVO*): *"At intersections without traffic signs or lights, vehicles coming from the right have right-of-way"*, the LLM generates: **Correct pair** (y_{pos}):

- Q: *"Which vehicle has priority at an unmarked intersection?"*
- A: *"At intersections without traffic signs or lights, vehicles approaching from the right have right-of-way."*

Incorrect pair (y_{neg}):

- Q: *"Who must yield at an intersection without signals?"*
- A: *"The vehicle traveling at lower speed must always yield to faster-moving traffic."* [Incorrect: contradicts the right-hand priority rule]

In-Domain Traffic Regulation Benchmarks. To assess out-of-domain generalization, we utilize two open-source road safety datasets: 890 Irish theory questions (Kelly, 2016) and 1,954 Austrian *ÖAMTC Führerschein-Test* questions (Sumer, 2013). We cast each multiple-choice item as a binary correctness judgment against individual options. After excluding image-dependent samples to ensure a purely linguistic evaluation, the resulting benchmarks comprise 2,500 test samples (from the remaining textual 625 questions) for the Irish set and 3,852 samples (from the remaining textual 963 questions) for the Austrian set.

Statistic	Train	Test
QA pairs	10,245	2,309
Unique paragraphs	49	10
Positive samples	50.1%	49.4%
Negative samples	49.9%	50.6%

Table 1: Synthetic *StVO* QA dataset statistics.

LexGLUE Benchmark. Chalkidis et al. (2021) provides seven English legal NLP tasks, enabling comparison with SOTA models. By benchmarking on LexGLUE, we probe the extent to which our KG-guided pre-training approach generalizes beyond statutory context.

4.2 Models and Training

We evaluate nine transformer-based encoders: **1) five legal-domain** (`bert-base-eurlex`, `bert-base-echr`, `bert-base-contracts`, `custom-legalbert` (Chalkidis et al., 2020), and `Italian-Legal-BERT` (Vitiugin et al., 2022)); **2) three general-purpose** (`google-bert-base` (Devlin et al., 2019), `legal-bert-base`, `legal-bert-small` (Chalkidis et al., 2020)); and **3) one non-Latin cross-lingual** (`Legal-heBERT` (Chriqui and Yahav, 2021)).

We explore **36** λ **configurations** over $\{0.0, 0.1, \ldots, 0.9\}$ per task under the unit-sum constraint ($\lambda_{\text{NSP}} + \lambda_{\text{CLU}} + \lambda_{\text{CLS}} = 1$), yielding **324 experimental runs** in total (9 models $\times$ 36 configs). The full configuration grid is described in Appendix B.

4.3 Evaluation Metrics

To provide a comprehensive assessment of the model performance, we report standard classification metrics: Macro-F_1 (m-F_1), Micro-F_1 (μ-F_1) (Buckland and Gey, 1994), Precision (P) and Recall (R) (Powers, 2020). To account for potential

class imbalances in our synthesized data, we include Balanced Accuracy (Brodersen et al., 2010). Finally, we report the absolute gain in m-F_1 over the respective base models to quantify the direct impact of our training approach.

5 Results and Discussion

5.1 Impact of KG-aware Pre-Training on *StVO* Legal Reasoning

This part evaluates how incorporating structured KG training directly influences the legal reasoning capabilities of PLMs. Multi-task pre-training yields large, consistent gains, but its magnitude depends sharply on the base model's domain and capacity (Table 2).

All five legal-domain encoders exceed 0.8 m-F_1 after pre-training, and the gain is largely decoupled from base performance. `bert-base-eurlex` rises from the weakest base (0.338) to 0.862 (the largest absolute gain, $+0.524$), while `bert-base-echr` reaches the overall peak (0.864). This decoupling indicates the improvement stems from the KG-guided objectives rather than from a strong starting point. Counterintuitively, the general-purpose `google-bert-base` gains more ($+0.205$, to 0.630) than the broad, multi-field `legal-bert-base` ($+0.114$ to 0.569). We attribute this to *domain-specific interference* where multiple legal English tasks conflict with the *StVO* domain. On the other hand, the single-subfield specialists like echr and eurlex align with it cleanly.

The two weakest cases fail for distinct reasons. `legal-bert-small` plateaus at 0.537 ($+0.050$), consistent with a capacity bottleneck that we examine across configurations in Section 5.5. `Legal-heBERT` instead suffers from a severe precision-recall imbalance ($P = 1.000$, $R = 0.096$), reflecting a fundamental linguistic and structural incompatibility between the Hebrew pre-training and the English target. This shows up as monotonic decay under task scaling (Figure 3). Consequently, `Legal-heBERT` cannot exploit the KG training objectives at all, and any gains it shows reflect the English CLS head alone rather than KG pre-training.

Overall, legal-domain models outperform general-purpose ones after KG pre-training, indicating that domain knowledge *synergizes* with the objectives rather than substituting for them. This is subject to sufficient capacity and language compatibility.

5.2 Cross-Lingual Transfer (RQ1)

To isolate cross-lingual transfer of legal knowledge, we train the NSP and CLU objectives exclusively on non-translated *German StVO* text, while the CLS head is trained on a subset of the *English* synthesized QA pairs. We then evaluate zero-shot on the English test set. Because NSP and CLU never see English, any zero-shot gain reflects German-learned structure surfacing through the English classification head (Table 3).

Five models achieve positive zero-shot gains ($+0.094$ to $+0.499$). The knowledge learned from German NSP+CLU is recoverable in English. `bert-base-eurlex` transfers most strongly ($0.338 \rightarrow 0.837$, $+0.499$) and barely improves with further fine-tuning ($+0.032$), indicating that the legal knowledge already injected through German NSP and CLU is sufficient on its own, leaving little for target-language fine-tuning to add. Three models transfer poorly zero-shot (`bert-base-echr`, `custom-legalbert`, `legal-bert-small`; down to -0.148) but recover fully when fine-tuned, with `legal-bert-small` posting the largest FT gain ($+0.532$, to 0.871). All models converge to a narrow band (0.862–0.878), indicating German pre-training encodes the required legal knowledge regardless of transfer mode.

5.3 In-Domain Traffic-Law Transfer (RQ2)

In Figure 2 all tested models improve zero-shot on the Austrian benchmark ($+0.041$ to $+0.158$), confirming transfer within the same regulatory family. Irish gains are smaller and less consistent ($+0.002$ to $+0.082$), and `Legal-heBERT` is the only model with negative transfer (-0.049), consistent with its limitations in Section 5.1. Austrian gains exceed Irish gains for nearly every model because the Austrian *ÖAMTC* regulations share greater regulatory overlap with *StVO*.

StVO pre-training generalizes within the European traffic-regulation family, but the gains are modulated more by linguistic distance (Table 3) than by traffic law distance (Figure 2). Together, these results indicate that cross-linguistic distance is a stronger barrier to transfer than cross-regulatory distance within European traffic law.

5.4 Out-of-Domain Legal Transfer (RQ3)

Here, we probe whether *StVO* pre-training transfers beyond traffic law to seven English legal tasks

Model	Base m-F_1	*Multi-task Pre-training on StVO (Zero-shot evaluations)* PT m-F_1	PT Gain	PT μ-F_1	PT Bal. Acc.	PT P	PT R	*Fine-tuning* FT m-F_1	FT Gain
Top Performers (m-$F_1 > 0.8$)									
bert-base-echr	0.566	**0.864**	+0.298	0.865	0.864	0.937	0.779	0.912	+0.048
bert-base-eurlex	0.338	**0.862**	+0.524	0.863	0.864	0.812	0.939	0.909	+0.047
bert-base-contracts	0.475	**0.845**	+0.370	0.846	0.848	0.784	0.950	0.905	+0.060
custom-legalbert	0.467	**0.820**	+0.353	0.820	0.820	0.812	0.828	0.900	+0.080
Italian-Legal-BERT	0.514	**0.813**	+0.299	0.816	0.815	0.905	0.701	0.893	+0.080
Moderate Performers									
google-bert-base	0.425	**0.630**	+0.205	0.662	0.666	0.597	0.970	0.919	+0.289
legal-bert-base	0.456	**0.569**	+0.114	0.613	0.617	0.565	0.945	0.888	+0.319
Limited Improvement									
legal-bert-small	0.486	**0.537**	+0.050	0.544	0.545	0.530	0.671	0.892	+0.355
Legal-heBERT	0.336	**0.435**	+0.099	0.554	0.548	1.000	0.096	0.870	+0.435

Table 2: Model performance on the synthetic *StVO* Q&A benchmark. **Base**: m-F_1 with no *StVO* knowledge; **PT**: after multi-task pre-training; **FT**: after fine-tuning on a few downstream samples from the synthetic *StVO* Q&A benchmark. PT and FT Gain report absolute m-F_1 improvement over Base and PT scores, respectively.

Model	Base m-F_1	*Multi-task Pre-training on StVO (Zero-shot evaluations)* PT m-F_1	PT Gain	PT μ-F_1	PT Bal. Acc.	PT P	PT R	*Fine-tuning* FT m-F_1	FT Gain
Successful Cross-Lingual Transfer									
bert-base-eurlex	0.338	**0.837**	+0.499	0.837	0.837	0.832	0.839	0.869	+0.032
google-bert-base	0.425	**0.773**	+0.347	0.775	0.777	0.719	0.893	0.871	+0.099
bert-base-contracts	0.475	**0.694**	+0.219	0.701	0.699	0.776	0.555	0.878	+0.184
legal-bert-base	0.456	**0.599**	+0.144	0.628	0.632	0.579	0.909	0.876	+0.276
Italian-Legal-BERT	0.514	**0.608**	+0.094	0.635	0.632	0.764	0.377	0.876	+0.268
Failed Zero-Shot, Recovered via Fine-Tuning									
bert-base-echr	**0.566**	0.563	−0.003	0.606	0.610	0.561	0.933	0.867	+0.304
custom-legalbert	**0.467**	0.377	−0.090	0.517	0.511	0.671	0.043	0.862	+0.485
legal-bert-small	**0.486**	0.339	−0.148	0.411	0.415	0.443	0.749	0.871	+**0.532**

Table 3: The objectives (NSP+CLU) are pre-trained on non-translated German *StVO*. Models are evaluated on the synthetic *StVO* Q&A benchmark. **Base**: m-F_1 before pre-training. **PT** columns report zero-shot performance after pre-training. **PT Gain** = PT m-F_1− Base. **FT** columns report performance after fine-tuning on a few downstream samples from the synthetic *StVO* Q&A. **FT Gain** = FT − PT. Bold marks the higher of Base and PT m-F_1 per model.

(Table 4). Despite training exclusively on German traffic regulations, our models set new SOTA on four metrics across four tasks—ECtHR-A, ECtHR-B, UNFAIR-ToS, and CaseHOLD— indicating that the approach encodes legal reasoning transferable beyond the *StVO* domain. See Appendix D for the complete task evaluations.

5.5 Task synergy and interference

We ablate the three objectives across all models and 36 λ configurations to isolate each one's contribution, as shown in Figure 3. The full NSP+CLU+CLS configuration outperforms all partial task subsets for 7 of the 9 evaluated models. For `bert-base-echr`, the integration of all three objectives achieves the peak performance, elevating m-F_1 from 0.587 (CLS only) to 0.864 (ALL), a gain of +0.277 over the strongest standalone baseline. Aggregated results identify CLS as the dominant objective. Its removal precipitates a mean m-F_1 drop of 0.277. `bert-base-contracts` in Figure 3 presents a notable exception to the observed synergy: the NSP+CLS configuration (0.845) outperforms the full ALL suite (0.797). We attribute this performance degradation to two primary factors: (a) a semantic domain mismatch between the legal entities of the *StVO* and the established legal primitives in US contract law, and (b) gradient interference within the CLU objective, where random negative sampling potentially pairs semantically related paragraphs as negatives, thereby introducing some noise that might desta-

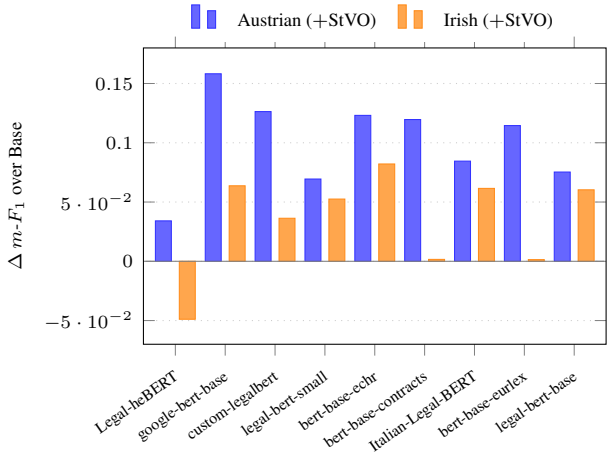


Figure 2: m-F_1 gain of StVO multi-task pre-training over each base model on Austrian and Irish zero-shot driving exams. Austrian gains are consistently larger, reflecting greater regulatory overlap with German *StVO*. Legal-heBERT is the only model showing negative transfer on Irish (-0.049).

bilize the optimization of this specific architecture. The performance degradation observed for `legal-bert-small` in Figure 3 suggests a capacity bottleneck, where the model's limited parameter space fails to capture the synergy between complex training objectives. The results conclude that the full configuration or dual-objective subset outperforms any single objective. The CLS task acts as the dominant contributor because it aligns directly with the downstream task structure.

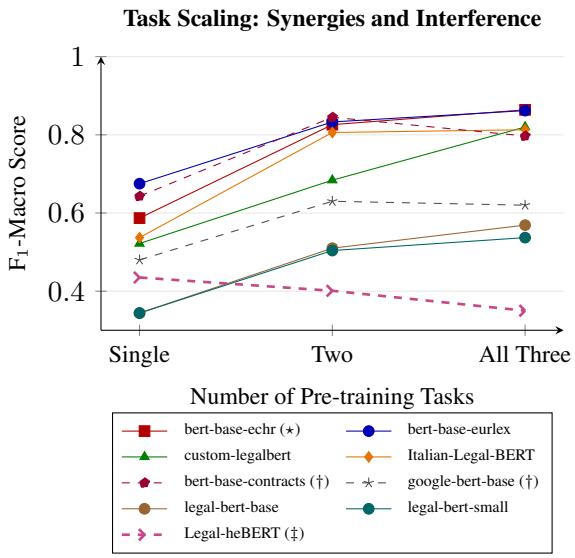


Figure 3: Task scaling across nine models. **bert-base-echr** (⋆) peaks at full-task synergy (0.864). (†) models show negative CLU interference when added to NSP+CLS. **Legal-heBERT** (‡) degrades monotonically.

Model	ECtHR-A		ECtHR-B		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	F_1
SOTA Published Baselines							
TFIDF-SVM	62.6	48.9	73.0	63.8	94.7	75.0	22.4
BERT	71.2	63.6	79.7	73.4	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	95.5	80.3	72.6
Longformer	69.9	64.7	79.4	71.7	95.5	80.9	71.9
Legal-BERT	70.0	64.0	80.4	**74.7**	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	96.0	82.3	75.4
Ours (StVO multi-task)							
BERT-EurLex+StVO	70.8	**65.7**†	79.7	73.8	95.7	79.4	71.5
BERT-ECHR+StVO	**72.7**†	61.2	**81.0**†	73.9	95.8	82.5	71.9
Custom-LegalBERT+StVO	71.2	61.6	78.6	70.6	**96.8**†	**86.0**†	**77.0**†

Table 4: LexGLUE results on the four tasks where *StVO* pre-training achieves new SOTA (full seven-task results in Appendix D). Scores are F_1 (%): macro (m) and micro (μ). **Bold** marks the best score per column; † marks a new SOTA over the published baselines of Chalkidis et al. (2021). CaseHOLD is reported as a single F_1 (μ=m).

6 Conclusion and Future Work

This paper presented a knowledge graph-guided multi-task pre-training approach for *StVO* legal question answering, encoding structured regulatory knowledge through three complementary objectives. Our main finding is that explicitly encoding legal structure via knowledge-graph-guided pre-training yields compounding benefits beyond domain-adaptive single-objective training.

Across the evaluated models, legal-domain models consistently outperform general-purpose ones, confirming that legal-domain knowledge leverages the objectives rather than being replaced by them. Model capacity and linguistic distance can nonetheless override this advantage. Our ablation identifies CLS as the dominant objective, with NSP and CLU serving as complementary structural tasks. The full configuration benefits most models, and the exceptions show that this synergy depends on semantic compatibility between the knowledge graph and the base model. The approach also transfers beyond its source domain. German structural pre-training transfers cross-lingually to English. It further yields gains on Austrian and Irish driving regulations and establishes new SOTA results on several English legal benchmarks. Together, these results indicate that knowledge-graph-guided pre-training on a single regulatory domain can bridge to structurally similar legal reasoning tasks across languages.

Future work will replace random CLU negative sampling with hard-negative sampling and extend the approach to other European regulatory corpora.

Limitations

Domain and language scope. Our approach is evaluated only on German traffic regulations (*StVO*). Generalization to other legal systems, languages, and regulatory domains remains to be validated empirically.

Synthetic *StVO* Benchmark. The 12,554 QA pairs are LLM-generated and may not reflect authentic examination distributions, particularly with respect to ambiguity and multi-hop reasoning. Because the training and evaluation pairs share the same generators, the reported gains should be read as evidence on synthetic legal QA rather than on authentic exams.

CLU negative sampling. Random negative sampling can pair semantically related paragraphs as negatives, introducing interference for out-of-domain bases such as `bert-base-contracts`. Possible alternatives such as hard-negative sampling is left to future work.

Knowledge-graph dependence. The KG is parsed from an LLM-translated English of the *StVO*, and the CLU and CLS objectives depend on TransE embeddings derived from it. This pipeline introduces translation and embedding noise that we do not quantify. The CLS attention also scales over the full set of paragraph embeddings. This is tractable for the 59-paragraph *StVO* graph, but its cost for larger knowledge graphs is untested.

Ethical Considerations

The models presented here are research prototypes trained on synthetic question-answer pairs derived from the German *StVO*. They are not intended to provide legally binding advice, and their outputs must not substitute for consultation with a qualified legal professional. Deployment in safety-critical settings such as automated driver assessment without human oversight is strongly discouraged. The pre-training and evaluation corpora are LLM-generated and passed through a quality-filtering pipeline. Filtering does not guarantee correctness. The synthetic QA pairs may still contain factual inaccuracies or misrepresent legal provisions. Because no held-out subset was validated by human legal experts, the true legal reasoning depth of the data remains uncertain. The models are trained exclusively on German traffic law. Cross-traffic transfer to Austrian and Irish driving regulations is evaluated zero-shot and shows variable performance. Users should not apply these models outside the German traffic-law context without further validation. The base encoders are pre-trained on existing legal corpora such as EUR-LEX and may encode biases present in those sources. Continued pre-training on *StVO* does not remove such biases, and their downstream effects on legal reasoning have not been systematically audited in this work.

References

Mehwish Alam, Frank van Harmelen, and Maribel Acosta. 2025. Towards semantically enriched embeddings for knowledge graph completion. *Neurosymbolic Artificial Intelligence*, 1:NAI–240731.

Dezhi An and Yuda Zhang. 2024. Research on interpretability charge prediction method based on legal provisions incorporated. *Proceedings of the 2024 3rd International Conference on Cyber Security, Artificial Intelligence and Digital Economy*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.

Bundesministerium für Digitales und Verkehr. 2013. Straßenverkehrs-ordnung vom 6. märz 2013 (bgbl. i s. 367), die zuletzt durch artikel 4 der verordnung vom 30. januar 2026 (bgbl. 2026 i nr. 32) geändert worden ist. Federal Ministry of Justice Portal, https://www.gesetze-im-internet.de/stvo_2013/. Accessed: 2026-05-19.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of

law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. 2022. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*.

Avihay Chriqui and Inbal Yahav. 2021. Hebert: Pre-training a bidirectional transformer for hebrew and sentiment analysis. *arXiv preprint arXiv:2102.01909*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Aug-gpt: Leveraging chatgpt for text data augmentation. *Preprint*, arXiv:2302.13007.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin M. K. Peters, Brandon Waldon, Daniel N. Rockmore, Diego A. Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *ArXiv*, abs/2308.11462.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2023. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *Preprint*, arXiv:2204.14211.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Adam Kelly. 2016. Theory Test Questions: A List of Sample Questions and Answers for the Irish Driver Theory Test. GitHub Repository. Accessed: 2024-03-23.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Preprint*, arXiv:2102.07033.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1035–1044.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 43–54.

David M. W. Powers. 2020. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

TYSS Santosh, Cornelius Weiss, and Matthias Grabmair. 2024. Lexsumm and lext5: Benchmarking and modeling legal summarization tasks in english. *arXiv preprint arXiv:2410.09527*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Ibrahim Siddig, Sviatoslav Tugeev, Munir Georges, and AImotion Bavaria. 2025. Pattern-based parsing of german traffic regulations (stvo) for legal knowledge graph construction (kgc). *ESSV25*.

Dominik Sumer. 2013. Driving License Trainer: Node.js application which helps you learning for your driving license test. GitHub Repository. Austrian driving license data based on ÖAMTC tests.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

Fedor Vitiugin, Aitor García-Pablos, and Montse Cuadros. 2022. Italian-legal-bert: A pre-trained transformer-based model for italian legal language. In *Proceedings of the 8th Italian Conference on Computational Linguistics (CLiC-it 2022)*.

Xuran Wang, Xinguang Zhang, Vanessa Hoo, Zhouhang Shao, and Xuguang Zhang. 2024. Legalreasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration. *IEEE Access*, 12:166843–166854.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1591–1601.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pages 108–112.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1441–1451.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

A Knowledge Graph Topology and Visualization

Figure 4 illustrates a portion of the knowledge graph (KG) that was created using the suggested parsing approach (Siddig et al., 2025) and derived from the Straßenverkehrs-Ordnung (*StVO*). This section presents the topological layout of the extracted knowledge base to provide structural transparency regarding the legal ontologies used during multi-task pre-training. The German *Straßenverkehrsordnung* (*StVO*) corpus was translated and systematically parsed into a legal knowledge graph (KG) structure. The resulting network comprises 554 discrete nodes and 945 directed edges. Nodes represent specific regulatory elements, such as paragraphs (§). Edges define the directional dependencies, nested hierarchies, and cross-references that bind these paragraphs together.

B Hyperparameter Configuration

This appendix reports the complete hyperparameter settings for the training stages of our pipeline: (B.1) TransE knowledge graph embedding and (B.2) KG-guided multi-task pre-training. All the aforementioned stages use seed `42` and a single NVIDIA H100 with 80 GB and CUDA 12.1.

B.1 TransE Embedding Training

We train TransE on the German *StVO* triple set with 554 nodes, 945 edges, and 5 relations to produce the 500-dimensional paragraph embeddings ($\mathbf{e}_p$) used as CLU soft targets (Section 3.4). The model is optimized with a margin-ranking loss over uniformly corrupted head/tail negatives. The best checkpoint is selected by validation loss.

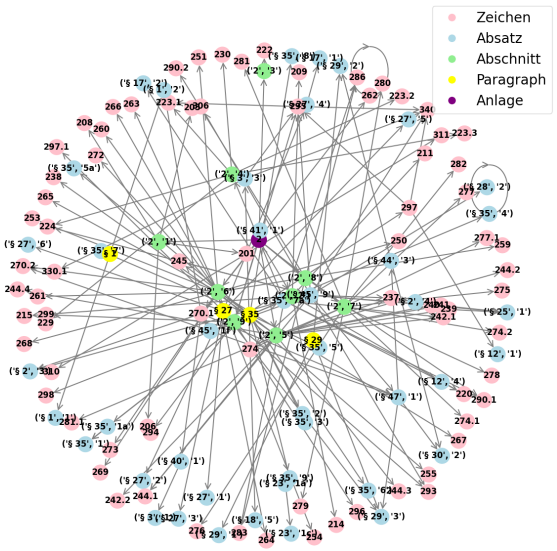


Figure 4: Partial visual topology of the generated *StVO* knowledge graph (554 nodes, 945 edges). Every node has tuples that correspond to its unique IDs, and the node content incorporates the related legal knowledge

B.2 KG-Guided Multi-Task Pre-Training

All base encoders use the masked-modeling configuration, combining NSP with CLU and CLS. The three loss weights satisfy $\lambda_{\text{NSP}} + \lambda_{\text{CLU}} + \lambda_{\text{CLS}} = 1$; we sweep the grid $\{0.0, 0.1, \ldots, 0.9\}$ subject to this unit-sum constraint. λ_{CLU} is determined as $1 - (\lambda_{\text{NSP}} + \lambda_{\text{CLS}})$, yielding 36 valid configurations per model (Section 4.2).

C Full Ablation Results

Table 7 reports the complete task-contribution ablation across all models and seven task configurations, supporting the analysis in Section 5.5. Figure 5 visualizes the per-model task contributions (Figure 5a) and the average importance of each objective (Figure 5b). Removing CLS causes the largest mean m-F_1 drop (0.277), far exceeding the drops from removing NSP (0.105) or CLU (0.056). Paired t-tests across the models confirm that the two primary objectives contribute significantly: CLS [$t(8)$=4.35, p<0.01] and NSP [$t(8)$=2.52, p<0.05]. The KG-guided CLU objective yields additional gains for top-performing models such as `bert-base-echr` (Table 2), but its aggregate effect is not statistically significant [$t(8)$=1.76, p=0.11]. This indicates that CLU's contribution is architecture-dependent and can introduce task interference in some general-purpose base models (Table 7).

Hyperparameter	Value
Scoring function	$-\lVert \mathbf{e}_h + \mathbf{r} - \mathbf{e}_t \rVert_2$
Embedding dimension d	500
Distance norm p	2 (L2)
Margin γ	1.0
Loss	Margin ranking
Negative sampling	Uniform (head + tail)
Optimizer	Adam
Learning rate	1×10^{-3}
Weight decay	1×10^{-6}
LR scheduler	ReduceLROnPlateau
L_2 regularization	0.01
Dropout	0.1
Batch size	256
Epochs	50
Gradient clipping	1.0 (max-norm)
Initialization	Xavier uniform, renorm to 1
Train / val / test split	80 / 10 / 10
Model selection	Best validation loss

Table 5: TransE hyperparameters for the *StVO KG* paragraph embedding.

Hyperparameter	Value
Modeling type	Masked (NSP core)
KG embedding source	TransE, $d = 500$
Optimizer	AdamW
Learning rate	5×10^{-5}
LR scheduler	Linear w/ warmup
warmup ratio	0.10
Epochs	20
Batch size (per task)	16 (NSP, CLU, CLS)
Gradient accumulation	2
Effective batch size	32
Max sequence length	128 (NSP / CLU / CLS)
Max segments per legal sentence	10
CLS label space	2 (correct / incorrect)
λ grid	$\{0.0, \ldots, 0.9\}$, unit-sum
Configurations per model	36

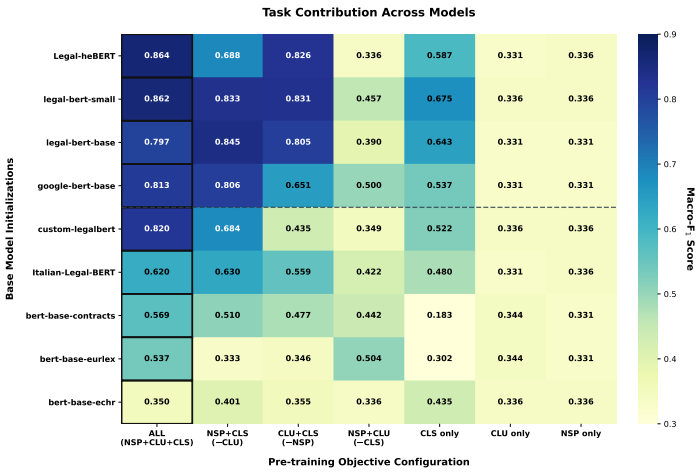
Table 6: Multi-task pre-training hyperparameters, shared across all base encoders.

D LexGLUE Full Results

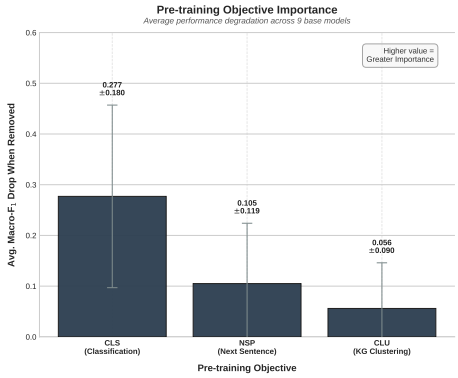
Table 8 reports the complete LexGLUE benchmark results for StVO-pretrained models against published baselines (Chalkidis et al., 2021) across all seven tasks, covering both μ-F_1 and m-F_1 averaged metrics. Our models establish new SOTA on four metrics (ECtHR-A,ECtHR-A/B,UNFAIR-ToS, and CaseHOLD). See Section 5.4 for the main analysis.

Base Model	ALL	NSP+CLS	CLU+CLS	NSP+CLU	CLS only	CLU only	NSP only	Δ (w/o CLS)
bert-base-echr	**0.864**	0.688	0.826	0.336	0.587	0.331	0.336	−0.528
bert-base-eurlex	**0.862**	0.833	0.831	0.457	0.675	0.336	0.336	−0.405
bert-base-contracts	0.797	**0.845**	0.805	0.390	0.643	0.331	0.331	−0.407
italian-legal-bert	**0.813**	0.806	0.651	0.500	0.537	0.331	0.331	−0.313
custom-legalbert	**0.820**	0.684	0.435	0.349	0.522	0.336	0.336	−0.471
google-bert-base	0.620	**0.630**	0.559	0.422	0.480	0.331	0.336	−0.198
legal-bert-base	**0.569**	0.510	0.477	0.442	0.183	0.344	0.331	−0.127
legal-bert-small	**0.537**	0.333	0.346	0.504	0.302	0.344	0.331	−0.033
Legal-heBERT	0.350	0.401	0.355	0.336	**0.435**	0.336	0.336	−0.014
Average	**0.692**	0.637	0.587	0.415	0.485	0.335	0.333	−0.278
Std Dev	0.172	0.164	0.187	0.063	0.147	0.005	0.003	0.183

Table 7: Complete ablation: 9 models × 7 task configurations. Bold = best per model. Δ (w/o CLS) = ALL − NSP+CLU.



(a) F_1-Macro per model across all 7 NSP/CLU/CLS configurations. Highlighted ALL column = full multi-task objective. Dashed line separates high- and low-performing models.



(b) Average m-F_1 drop when each objective is removed across all models. Error bars = std. dev.

Figure 5: Ablation study: CLS dominates (avg. drop = 0.277). `bert-base-contracts` and `google-bert-base` show negative CLU interference, confirming task synergy is architecture dependent.

Model	ECtHR-A		ECtHR-B		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ/m-F_1
Published Baselines (Chalkidis et al., 2021)													
TFIDF-SVM	62.6	48.9	73.0	63.8	74.0	64.4	63.4	47.9	87.0	81.4	94.7	75.0	22.4
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	**57.9**	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	**72.1**	57.4	88.2	**83.1**	95.5	80.3	72.6
Longformer	69.9	64.7	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
Legal-BERT	70.0	64.0	80.4	**74.7**	76.4	**66.5**	72.1	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	**76.6**	65.9	70.7	56.6	**88.3**	83.0	96.0	82.3	75.4
Our StVO Multi-Task Models (RQ3)													
BERT-EurLex+StVO	70.8	**65.7**†	79.7	73.8	70.8	61.1	70.3	46.6	87.8	82.1	95.7	79.4	71.5
BERT-ECHR+StVO	**72.7**†	61.2	**81.0**†	73.9	71.5	62.7	69.3	44.9	87.8	81.1	95.8	82.5	71.9
Custom-LegalBERT+StVO	71.2	61.6	78.6	70.6	68.9	48.1	67.7	43.5	87.2	81.2	**96.8**†	**86.0**†	**77.0**†

Table 8: Full LexGLUE results across all seven tasks: our *StVO*-pretrained models vs. the published baselines of Chalkidis et al. (2021). Scores are F_1 (%), micro (μ) and macro (m); CaseHOLD is a single F_1 (μ=m). **Bold** marks the best score per column; † marks a new SOTA over the published baselines. Our models set new SOTA on four tasks (ECtHR-A, ECtHR-B, UNFAIR-ToS, CaseHOLD). They show negative transfer on SCOTUS, EUR-LEX, and LEDGAR.