ProMALex: Progressive Modular Adapters for Multi-Jurisdictional Legal Language Modeling

Santosh T.Y.S.S, Mohamed Hesham Elganayni
School of Computation, Information, and Technology;
Technical University of Munich, Germany

Abstract

This paper addresses the challenge of adapting language models to the jurisdiction-specific nature of legal corpora. Existing approaches—training separate models for each jurisdiction or using a single shared model—either fail to leverage common legal principles beneficial for low-resource settings or risk negative interference from conflicting jurisdictional interpretations. To overcome these limitations, we propose a parameter-efficient framework ProMALex, that first derives hierarchical relationships across jurisdictions and progressively inserts adapter modules across model layers based on jurisdictional similarity. This design allows modules in lower layers to be shared across jurisdictions, capturing common legal principles, while higher layers specialize through jurisdiction-specific adapters. Experimental results on two legal language modeling benchmarks demonstrate that ProMALex outperforms both fully shared and jurisdiction-specific models.

1 Introduction

Language models, with their unsupervised pre-training on vast datasets, excel in NLP tasks (Devlin, 2018; Raffel et al., 2020; Radford et al., 2019; Brown et al., 2020; Zhang et al., 2022) and have transformed the legal industry with applications like contract analysis and statutory interpretation (Dahl et al., 2024; Katz et al., 2024; Martínez, 2024; Blair-Stanek et al., 2023; Engel and Mcadams, 2024; Kang et al., 2023; Guha et al., 2024; Santosh et al., 2025). However, adapting these models to specialized legal tasks poses challenges (Katz et al., 2023; Santosh et al., 2024a), making continued pre-training on legal corpora a promising approach to enhance legal domain-specific understanding (Chalkidis et al., 2020, 2023; Niklaus et al., 2023b; Paul et al., 2023; Santosh et al., 2024c; Colombo et al., 2024b).

Yet, legal data introduces unique challenges due to its jurisdiction-dependent nature. Unlike domains such as medicine or science, where principles are globally consistent, legal frameworks—shaped by statutes, precedents, and enforcement mechanisms—vary significantly across jurisdictions. For example, the doctrine of strict liability has jurisdiction-specific interpretations. In the UK, it originates from Rylands v. Fletcher (1868), which established liability for inherently dangerous activities regardless of fault. In contrast, US law primarily applies strict liability to product liability, as seen in Katz v. The Queen (1989), where manufacturers are held accountable for defective products. Similarly, the right to privacy reflect shared principles but diverge in implementation: in the EU, privacy is codified through the GDPR, which sets stringent guidelines for data protection. Under the European Convention on Human Rights, Article 8 guarantees respect for private and family life. Meanwhile, in India, privacy was established as a fundamental right under Article 21 of the Constitution in Justice K.S. Puttaswamy v. Union of India (2017). A model trained without accounting for jurisdictional distinctions risks conflating these nuances, leading to misinterpretations of their scopes, legal thresholds, enforcement mechanisms, or sources of authority.

A naïve solution to address this is training separate models for each jurisdiction (Paul et al., 2023; Zheng et al., 2021), preserves jurisdictional distinctions, but fails to leverage shared legal principles, making it inefficient for low-resource jurisdictions. Conversely, a single shared model across jurisdictions, as done in prior works (Chalkidis et al., 2020; Niklaus et al., 2023b; Santosh et al., 2024c; Colombo et al., 2024b), enables knowledge sharing but risks negative interference, where conflicting interpretations undermine jurisdiction-specific understanding. For example, conflating GDPR frameworks with Indian privacy case law,

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22201–22217

leading to misinterpretations of legal thresholds or rights applicable to specific scenarios.

Achieving an optimal balance is therefore crucial: shared representations capture common legal principles, facilitating positive transfer across related jurisdictions, while jurisdiction-specific modules disentangle the nuanced reasoning and precedential variations. However, existing approaches often struggle to strike this balance, either overfitting to dominant jurisdictions or oversimplifying legal system heterogeneity. Additionally, fully fine-tuning large models is expensive, underscoring the need for parameter-efficient solutions that enable controlled knowledge sharing while preserving jurisdiction-specific specialization.

To address these challenges, we propose ProMALex (Progressive Module Adapters for Legal Language Modeling), a parameter-efficient framework leveraging LoRA (Hu et al., 2021) modules inserted into a frozen pre-trained model. ProMALex employs hierarchical agglomerative clustering on dataset embeddings to automatically derive jurisdictional relationships. This clustering structure and relatedness scores guides: (1) which jurisdictions should share parameters, (2) up to which layer parameters remain shared before splitting into jurisdiction-specific modules, and (3) how parameters are allocated for each jurisdiction after the split. Initially, ProMALex uses shared LoRA parameters across all jurisdictions at lower layers. As layers progress, the parameters are progressively split into jurisdiction-specific modules based on the derived jurisdictional relationships. This progressive strategy leverages shared foundational knowledge at lower layers with fine-grained specialization for jurisdictional nuances at higher layers.

We evaluate ProMALex through two legal language modeling experiments using case law documents from: (1) English-speaking jurisdictions, including the EU, ECHR, India, Canada, and the UK, and (2) a multilingual setup covering jurisdictions such as Switzerland, Germany, Denmark, and Belgium. ProMALex consistently outperforms both fully-shared models and jurisdiction-specific models, particularly in low-resource jurisdictions. It also exhibits superior robustness on legislative corpora, effectively grasping formalized legal principles in legislation from the nuanced, context-driven reasoning in case law.

2 Related Work

Legal-specific Pretraining Continuing pretraining on domain-specific corpora has shown consistent improvements in downstream task performance (Gururangan et al., 2020). This method has been widely adopted in legal NLP, resulting in models like LegalBERT (Chalkidis et al., 2020), CaseLawBERT (Zheng et al., 2021), and PoLBERT (Henderson et al., 2022), among others. Recent works such as LexLMs (Chalkidis et al., 2023) consolidate efforts by pretraining on multinational corpora like LeXFiles. Similarly, multilingual models, including LegalXLM (Niklaus et al., 2023b), extend these advancements to non-English jurisdictions.

While encoder-only architectures dominate, emerging approaches explore encoder-decoder and decoder-based architectures, leveraging instruction tuning and preference optimization for enhanced task-specific performance (Santosh et al., 2024c; Niklaus et al., 2024; Colombo et al., 2024b). Despite significant progress, existing approaches fail to account for jurisdictional variations, diluting critical distinctions. Our work addresses these limitations by focusing on efficient pretraining tailored for multi-jurisdictional corpora to preserve and leverage jurisdiction-dependent nuances. Refer to Appendix A for a comprehensive review of legal-specific models.

Specializing LLMs to Domains Domain specialization techniques can be categorized into three approaches (Ling et al., 2023): external augmentation, which retrieves and incorporates relevant knowledge as context during inference (Long et al., 2023; Xu et al., 2023; Mao et al., 2024); prompt crafting, which uses domain-specific instructions or templates to guide model (Nachane et al., 2024; Heston and Khun, 2023; Yu et al., 2022; Jiang and Yang, 2023); and internal knowledge updates, which fine-tune the model on domain-specific text (Kraljevic et al., 2021; Rongali et al., 2020; Colombo et al., 2024b). While external augmentation and prompting are effective, they face challenges such as conflicts between retrieved and internal knowledge and insufficient recall of domain-specific details. Hence, we focus on fine-tuning using adapter-based techniques (e.g., Adapters, LoRA), which mitigate catastrophic forgetting and allow resource-efficient adaptation (Poth et al., 2023; Houlsby et al., 2019; Hu et al., 2021). Our approach is particularly tailored to multi-jurisdictional legal modeling, paralleling multi-task or multi-domain challenges.

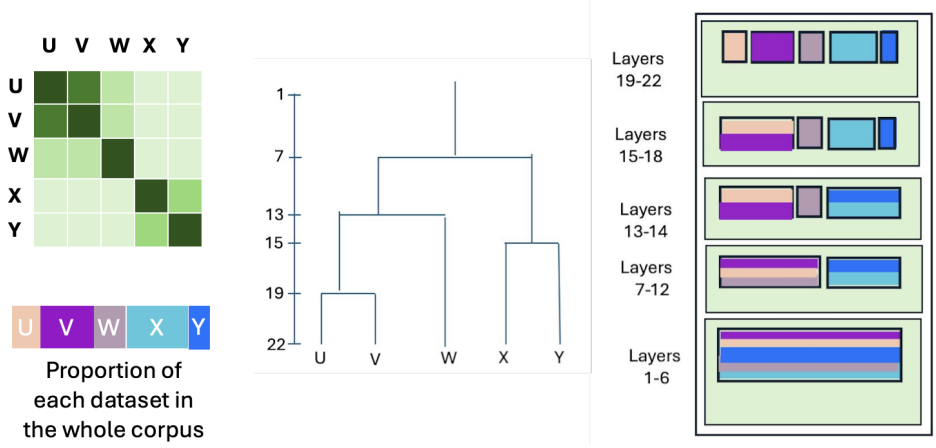


Figure 1: (Left) Pairwise similarities between jurisdictions and the proportion of each jurisdiction's dataset relative to the entire corpus. (Center) Hierarchical tree (dendrogram) derived from the similarity values, scaled from 1 to 22 (total number of model layers). (Right) Design of progressive adapters across model layers. The colors of each jurisdiction represent the adapter block that gets activated during both training and inference. The size of the adapter blocks indicates their rank, proportional to the size of the corresponding jurisdiction's dataset.

Existing multi-task learning methods fall into two main categories. End-to-end models either train all tasks jointly, often ignoring task-specific variations (Xin et al., 2024), or adopt a mixture-of-experts framework to learn dynamically allocating experts (Dou et al., 2024). Two-stage methods, on the other hand, train single-task modules and use routing mechanisms, fusion layers, or weighted combinations of individual task-specific models (Feng et al., 2024; Pfeiffer et al., 2020; Huang et al., 2023). While two-stage models offer modularity, they lack direct parameter sharing (Gururangan et al., 2021; Diao et al., 2023). End-to-end models, in contrast, allow parameter sharing but often fail to leverage metadata information for domain distinctions (Shen et al., 2023).

Our framework, ProMALex, builds on hierarchical approaches like Chronopoulou et al. (2021), which use predefined domain relationships to obtain hierarchical tree structure for selective parameter sharing, enabling positive transfer while minimizing interference. However, they apply a uniform hierarchical tree of adapters across all layers, over-constraining lower layers and under-specializing higher ones, limiting their ability to capture nuanced domain variations. ProMALex addresses this by adopting a progressive approach that leverages this hierarchy to identify optimal transition points for shifting from shared to jurisdiction-specific parameters, ensuring optimal sharing.

3 Our Approach: ProMALex

Our goal is to adapt a pre-trained base language model to a multi-jurisdictional legal corpus spanning N jurisdictions using parameter-efficient modules like LoRA with a fixed rank K per layer, enabling efficient adaptation without full retraining. The model is trained on a language modeling task to predict the next token based on context. A key assumption is the availability of jurisdictional information for each instance during both training—a common scenario in real-world deployments to maintain contextual jurisdictional relevance. To achieve this, we propose ProMALex, which employs a data-driven approach to identify relationships among jurisdictions to create a hierarchical structure that guides the progressive application of LoRA adapters across model layers.

3.1 Hierarchical Structure Creation

To identify relationships between jurisdictions, we generate embeddings by randomly sampling an equal number of instances from each jurisdiction's corpus. Using the pre-trained base language model, we compute contextual representations for these samples, aggregating them to produce a normalized mean embedding for each jurisdiction, which captures its characteristics (Vu et al., 2020; Aggarwal et al., 2024). We then apply Agglomerative Clustering, a bottom-up hierarchical algorithm, where each jurisdiction starts as its own cluster, and these clusters are successively merged based on cosine similarity, ensuring jurisdictions with

similar dataset characteristics are merged first. The resulting hierarchical structure is represented as a dendrogram (Fig. 1), where closely related jurisdictions remain clustered deeper in the hierarchy, while more distinct jurisdictions diverge at lower depths in the tree. This hierarchical structure forms the basis for the Progressive Adapter Design.

3.2 Progressive Adapter Design

The similarity scores obtained from the hierarchical clustering tree are scaled from 1 to the total number of layers in the pre-trained model. These scores and the hierarchical relationships define how jurisdictions share adapters at each layer, specifying up to which layer they use a common adapter before transitioning to specialized modules.

In the higher levels of the hierarchy, where jurisdictions exhibit high similarity, a single shared LoRA module is applied across all jurisdictions in the lower layers, enabling efficient learning of common patterns. As the model advances to higher layers, the clustering hierarchy dictates the introduction of specialized LoRA modules. Jurisdictions that become increasingly distinct are assigned independent LoRA modules, allowing the model to capture their unique characteristics. This progressive approach ensures that closely related jurisdictions share LoRA adapters over more layers, while those further apart split earlier, achieving a balance between efficient shared learning and precise specialization to capture nuanced distinctions. For example, as shown in Figure 1, jurisdictions {U,V,W,X,Y} share a common adapter up to Layer 7 and they split into two groups: {U,V,W} and {X,Y}. The sub-group {X,Y} further splits at Layer 15, while {U,V,W} splits into {U,V} and {W} at Layer 13. Finally, {U,V} separates into individual adapters for U and V at Layer 19.

To adhere to a fixed budget of rank K per layer, the available LoRA rank is divided proportionally at each split layer based on the dataset sizes of the jurisdictions in each group. Jurisdictions with larger datasets are allocated a greater share of the parameter budget, ensuring sufficient capacity to model their complexity. During training and inference, data activates the corresponding LoRA modules based on the jurisdiction label as it progresses through the model layers. The progressive adapter design enables the model to leverage shared knowledge in lower layers based on jurisdictional similarity while progressively introducing specialized adapters in higher layers to capture fine-grained distinctions for accurate and contextually relevant legal language modeling.

4 Experiments

4.1 Datasets

We conduct experiments in two setups to evaluate the adaptability of legal language models: one using English legal corpora and the other using multilingual legal corpora from various jurisdictions. We use the Case law documents, representing judicial decisions issued by courts, for training and testing these models. Case law often contains rich reasoning, judicial interpretations, and practical applications of legislation, making it an ideal source for training models in understanding the nuances of legal language and context.

For the English setup, we include case law files from five jurisdictions: EU, ECHR, UK, India, and Canada, derived from the LexFiles corpus (Chalkidis et al., 2023). For the multilingual setup, we use case law corpora from four jurisdictions: Switzerland, Germany, Denmark, and Belgium, obtained from Multilegalpile corpus (Niklaus et al., 2023b). Specific details of these dataset sources are described in Appendix B.

Dataset sizes, including the number of documents and tokens for each jurisdiction in both setups, are summarized in Appendix Table 4 and 5.

4.2 Baselines & Implementation Details

For our experiments, we use GPT-2-XL (Radford et al., 2019) as the base model for the English setup and mGPT-XL (Shliazhko et al., 2022) for the multilingual setup. Both models are adapted to legal corpora using LoRA modules (Hu et al., 2021) with a rank budget of 64 per layer for parameter-efficient fine-tuning. Detailed experimental hyperparameters are provided in Appendix C.

We compare ProMALex against the following baselines: **(a) Shared**: A single LoRA (rank 64 per layer) is trained on the combined data from all jurisdictions, without explicit conditioning on jurisdiction labels during training or inference. This approach reflects the most common paradigm in prior works on legal language modeling (Chalkidis et al., 2023; Niklaus et al., 2023a; Colombo et al., 2024b; Santosh et al., 2024c). **(b) Mixture of Experts-LoRA (MoELoRA)** (Dou et al., 2024; Luo et al., 2024): Inspired by MoE (Fedus et al., 2022), this approach replaces Mutliheaded attention weights with LoRA experts and employs a router to select

	ECHR	EU	Ind	UK	Can	Avg.	Swi	Ger	Den	Bel	Avg.
Base	8.857	7.947	17.281	16.178	68.253	23.703	604.369	625.124	449.576	281.797	490.217
Shared	6.562	6.603	10.847	10.673	5.466	8.030	8.543	8.081	7.039	4.752	7.104
MoELoRA	7.564	7.561	13.142	12.915	6.019	9.440	10.158	9.565	8.239	5.389	8.338
Single-Uni	6.349	6.381	10.651	10.509	5.375	7.853	8.497	8.037	6.879	4.528	6.985
Single-NonUni	6.374	6.366	10.669	10.302	5.473	7.837	7.502	7.531	6.301	4.251	6.396
Two-Stage	6.479	6.372	10.783	10.314	5.678	7.925	8.506	8.047	7.012	4.738	7.076
HLoRA	6.425	6.374	10.584	10.393	5.213	7.798	6.935	7.481	6.286	4.228	6.233
ProMALex	**6.080**	**6.093**	**10.118**	**10.203**	**4.986**	**7.496**	**5.643**	**6.918**	**5.594**	**3.717**	**5.468**

Table 1: Perplexity scores across jurisdictions and their averages in both English and Multilingual setups (lower is better). ProMALex consistently outperforms baseline models, achieving lower perplexity scores in both settings.

appropriate experts. Both the LoRA experts and the router are optimized during training. Routing does not explicitly use jurisdiction labels, leaving the model to learn the optimal allocation of data to experts. To address the common issue of router convergence favoring a small subset of experts, we include a balancing loss (Shazeer et al., 2017) that equalizes expert utilization. In our experiments, we use five experts for the English setup and four for the multilingual setup. The total rank per layer is constrained to 64, distributed uniformly among the experts. **(c, d) Single LoRA (Uniform, Non-Uniform)**: separate LoRA experts are trained for each jurisdiction, and inference is performed exclusively using the expert corresponding to the specific jurisdiction. The total rank per layer is constrained to 64, with two variations for rank allocation: in the uniform case, the rank is evenly distributed across all experts, while in the non-uniform case, the rank is allocated proportionally based on the number of tokens in the training data for each jurisdiction. **(e) Two-stage MoELoRA**, inspired from Wu et al. (2024); Pfeiffer et al. (2020), this approach first trains single LoRA modules for each jurisdiction separately. In the second stage, the model dynamically learns how to combine the performance of multiple trained LoRAs through a routing network. Only the gating module is trained in the second stage, with the LoRA experts from the first stage frozen. For this setup, we use the non-uniform single LoRA experts in the first stage, as they have demonstrated better performance compared to the uniform variant. **(f) Hierarchical LoRA (HLoRA)** (Chronopoulou et al., 2021), similar to ours, derives a data-driven hierarchical clustering structure of different jurisdictions and designs hierarchical adapters that are applied uniformly across each transformer layer. To ensure a fair comparison with our method, the rank budget of 64 per layer is evenly distributed across all LoRA blocks at different depth levels within each layer. Jurisdiction labels are explicitly used during both training and inference to route data through predefined hierarchical paths within each layer.

In all router-based methods, the router is implemented as a two-layer MLP, which adds extra parameter overhead compared to ProMALex, Single LoRA, Shared, or HLoRA models.

4.3 Results

We report the perplexity scores for each jurisdiction in the case law corpus, along with the overall average, for the English and Multilingual setup in Table 1. All methods are implemented with a fixed additional compute budget constraint on top of the base model to ensure a fair comparison across them. The results show that a fully *shared* LoRA module across all jurisdictions outperforms *MoELoRA* in both setups across all jurisdictions, regardless of their dataset proportions. This suggests that *MoELoRA* struggles to effectively distinguish between jurisdictions, leading to suboptimal allocation of experts during training and inefficient activation during inference. This finding highlights the complexity of learning jurisdiction-specific nuances, in fine-grained domains, such as case law corpora, where topical overlap between jurisdictions is significant and demonstrates that predefining and allocating expert modules may be more effective than allowing the model to learn these distinctions implicitly.

Furthermore, a *single* LoRA module per jurisdiction (without any sharing) outperforms shared configurations, indicating that combining data from multiple jurisdictions makes it harder for the model to disentangle and learn jurisdiction-specific nuances. This challenge is particularly pronounced for low-resource jurisdictions in the multilingual setup, where overrepresented jurisdictions receive disproportionate parameter allocation, thereby ex-

acerbating underfitting in less-represented jurisdictions. These findings underscore the importance of a balanced parameter distribution, rather than implicit allocation by model through learning. Interestingly, *non-uniform single*-module allocation, where parameters are distributed proportionally based on dataset size, performs better than *uniform-single* allocation, suggesting an optimal parameter distribution in a multi-jurisdiction setup, balances resource allocation based on data availability. In the multilingual setup, where dataset distribution is skewed, assigning more parameters to highly populated corpora like Switzerland significantly improves perplexity. However, reducing parameter allocation for other jurisdictions such as Germany, Denmark, and Belgium also leads to improvement, highlighting that excessive parameter allocation to low-resource jurisdictions may cause overfitting and hinder generalization.

Two-stage MoELoRA, which builds on non-uniform allocation by introducing a learned routing mechanism to activate specific experts per input, underperforms compared to the single-module model. This suggests that the model struggles to learn an effective strategy for expert activation and knowledge sharing across jurisdictions, ultimately diminishing its performance. The addition of learned routing introduces complexity, but without an effective method to leverage it, the model fails to achieve the desired outcome. *HLoRA*, which predefines module activation paths based on jurisdictional relationships derived from embeddings, improves over both *shared* and *single* models. This confirms our earlier observation that both complete specialization and complete sharing of parameters are suboptimal, with predefined hierarchical sharing offering a more effective alternative. However, HLoRA applies a static hierarchical tree across all layers, which may constrain lower layers too rigidly while under-specializing higher layers. This limitation restricts the model's ability to capture nuanced jurisdiction variations and disrupts the optimal allocation of parameters when too many modules are used at certain layers.

In contrast, *ProMALex* further improves over HLoRA by introducing a progressive hierarchical approach. Rather than enforcing a fixed hierarchical tree across all layers, ProMALex dynamically determines transition points where layers shift from shared to jurisdiction-specific parameters. This flexibility allows for more adaptive specialization and ensures that parameters are optimally allocated at each split, improving overall model performance. By progressively disentangling shared and jurisdiction-specific knowledge at different layers, ProMALex captures fine-grained distinctions across jurisdictions, allowing it to outperform all prior methods in both setups.

Across both setups, the performance gap is larger in the multilingual setup compared to English, which can be attributed to the base model's pretraining knowledge. To analyze this, we report perplexity scores on the base model without any fine-tuning, capturing its inherent jurisdictional knowledge from pretraining. The results suggest that the English model already possesses significant knowledge of these jurisdictions due to its pretraining corpus, whereas the multilingual model has much higher perplexity scores, indicating less inherent familiarity with specific jurisdictions. This highlights the challenges of multilingual models in acquiring and generalizing domain-specific knowledge across varied legal systems, further reinforcing the effectiveness of ProMALex's progressive approach.

4.4 Discussion & Analysis

4.4.1 Ablation on ProMALex

We analyze two key design components of ProMALex: (i) deriving similarity values between jurisdictions to determine the hierarchical structure for progressive adapter design and the layers at which splits occur, and (ii) the strategy for parameter allocation between jurisdictions after splitting from a shared module.

For the first component, we compare ProMALex's Dataset Embedding Similarity (DES) (Section 3.1) with two alternative similarity measures: (i) Vocab Overlap Ratio (VOR): It quantifies jurisdictional similarity at the lexical level by calculating the vocabulary overlap using Jaccard Similarity Index (in %) between the top most frequent words (excluding stopwords) in each corpus (Gururangan et al., 2020). VOR has been used in prior works on legal case summarization (Santosh et al., 2024b) to assess cross-jurisdictional similarity. (ii) Model Parameter Similarity (MPS): This method is based on the idea that additional parameters, learned through parameter-efficient fine-tuning on a specific jurisdiction, capture jurisdiction-specific knowledge. Previous studies (Feng, 2023; Zhou et al., 2022; Aggarwal et al., 2024) have shown that these additional parameters can serve as effective embeddings for quantifying jurisdictional similarity. To

	ECHR	EU	Ind	UK	Can	Avg.	Swi	Ger	Den	Bel	Avg.
DES (ProMALex)	**6.080**	**6.093**	**10.118**	**10.203**	**4.986**	**7.496**	**5.643**	**6.918**	**5.594**	**3.717**	**5.468**
MPS	6.163	6.202	10.317	10.295	5.061	7.608	5.651	7.029	5.732	4.069	5.620
VOR	6.232	6.234	10.382	10.279	5.107	7.647	5.674	7.191	5.931	4.091	5.722
Uniform	6.346	6.316	10.445	10.298	5.171	7.715	5.988	7.432	6.399	4.301	6.030
Non-Uniform	6.304	6.282	10.418	10.268	5.168	7.688	5.898	7.229	6.334	4.139	5.900

Table 2: Ablation Study on ProMALex Components: Perplexity Scores Across Jurisdictions and Averages in English and Multilingual Setups (lower is better). The first group investigtes different methods for determining relationships between jurisdictions, while the second explores parameter allocation strategies during splitting.

implement MPS, we fine-tune a model on each corpus (similar to our single-uniform model) and compute the similarity between the learned model parameters. These similarity values are then used to guide the hierarchical structure and determine layer transitions, following a similar approach as in ProMALex and the parameters after split are allocated based on the proportion of each dataset.

We report the perplexity scores for each jurisdiction and their averages in both setups in Table 2, and visualize the hierarchical structures with layer transitions in the Appendix D. All methods outperform the prior baselines in Table 1, demonstrating the effectiveness of the modular progressive design for optimal sharing and specialization across jurisdictions. Among the methods, VOR is the least effective, indicating that vocabulary overlap alone does not provide a reliable or well-calibrated measure for determining jurisdictional similarity or transition points. Additionally, VOR determines similarity independently of the base model, failing to leverage the model's inherent knowledge, which may lead to suboptimal results for a given base model. MPS improves over VOR but lags behind DES. Despite the additional complexity of fine-tuning the model to obtain parameters that capture jurisdictional similarity, MPS does not outperform the simpler DES approach, which directly utilizes the existing base model for similarity determination. The main reason for this difference is that while MPS utilizes the additional learned parameters on top of base model to encode jurisdiction-specific knowledge, it may not be optimally configured to initiate the hierarchy design for effective complementarity during training. In contrast, DES leverages the base model to derive similarities, facilitating the construction of a well-calibrated adapter design hierarchy. This approach helps disentangle the base model's embedding space, allowing for more effective sharing and specialization across jurisdictions. These findings suggest that utilizing the base model's learned parametric space to determine jurisdictional similarities can lead to more efficient model adaptation. Exploring methods that leverage the base model's parametric space for similarity estimation presents a promising direction for future research.

To study the second design component—parameter allocation—we use a simple VOR-based hierarchical structure and compare two allocation strategies: (i) non-uniform allocation, where parameters are assigned in proportion to dataset size after each split, and (ii) uniform allocation, where parameters are evenly distributed across all jurisdictions after a split. Instead of relying on similarity values, we perform splits at equal intervals across layers, determined by the total tree depth. From the second group in Table 2, we observe that non-uniform allocation consistently improves performance across all jurisdictions, leading to better resource utilization. In contrast, uniform allocation may lead to overfitting in low-resource jurisdictions and underfitting in high-resource ones. Interestingly, dataset size emerges as a strong indicator for parameter allocation in this setup, though this may not generalize if datasets contain redundancy. This finding warrants further investigation into allocation strategies that consider dataset diversity by analyzing the density of the parametric embedding space.

Comparing Non-uniform to VOR in Table 2, the key difference lies in how transition layers are determined: VOR uses similarity values, whereas Non-uniform relies on equal spacing. We observe that VOR-based allocation consistently outperforms Non-uniform allocation across all jurisdictions. This suggests that leveraging similarity measures enables more calibrated and adaptive transition points, leading to better specialization, in contrast, to equal spacing between layers, reinforc-

	ECHR	EU	India	UK	Canada	Average
Base	4.976	6.215	12.370	7.105	9.286	7.991
Shared	4.659	5.148	7.207	4.581	4.347	5.188
MoELoRA	4.487	5.288	9.444	4.783	4.404	5.681
Single-Uniform	4.413	5.195	10.427	4.585	4.262	5.776
Single-NonUniform	4.282	5.193	10.310	4.588	4.238	5.722
Two-stage MoELoRA	4.629	5.164	9.896	4.942	**4.163**	5.759
HLoRA	4.502	5.213	9.613	4.923	4.224	5.695
ProMALex	**4.093**	**5.134**	**6.824**	**4.476**	4.296	**4.965**

Table 3: Perplexity scores across five English jurisdictions and their average on the legislation corpus, assessing generalization through case law codification (lower is better). ProMALex achieves the lowest perplexity scores, outperforming all baseline models.

ing the utility of similarity-driven approaches in guiding progressive module design.

4.4.2 Jurisdiction Specificity

In addition to evaluating the models on the case law corpus, we test them on a jurisdiction-specific legislation corpus. This setup evaluates the models to understand jurisdictional distinctions—such as legal scopes, thresholds, and enforcement mechanisms—from the complex, context-driven reasoning in case law and map them to the formalized structures of legislation. This process, known as legal codification, represents a reverse knowledge transfer scenario, where the models generalize from the nuanced applications of law in case decisions to the standardized legal norms in legislation. This aligns with real-world use cases, such as systems that convert implicit legal knowledge from case law into explicit legal principles, facilitating automated codification and legislative drafting.

We obtain the legislation corpus for EU, UK, and Canada from LexFiles (Chalkidis et al., 2023), ECHR from Santosh et al. (2023) and India from CivicTech India. Specific details of these dataset sources are described in Appendix B.

We report the perplexity scores on the legislation corpus across each jurisdiction, in Table 3. On average, the *shared* model outperforms *MoELoRA*, with the reverse trend observed in specific jurisdictions, such as ECHR. Interestingly, the single model performs worse than the shared model, which contrasts with the trends seen on the case law corpus (Tab. 1). While the single model shows improvements in ECHR and Canada, it experiences a significant decline when applied to the Indian corpus. Two-stage model improves performance on the Indian, EU and Canada corpus but results in a decline in other jurisdictions, leading to an overall slightly worse performance compared to the shared. While HLoRA improves upon the single variants, it still lags behind the shared models. ProMALex improves across jurisdictions, showcasing its robust generalizability and jurisdiction specificity.

These results highlight the importance of further investigation into jurisdiction-specific characteristics. For instance, the Indian corpus exhibits a marked decline in performance with single models compared to shared models. This suggests that, in certain jurisdictions, the legislation corpus may not closely align with the case law corpus, particularly when specific legal rules are not explicitly addressed in case law or lack sufficient frequency to form clear patterns. This discrepancy may arise from factors such as memorization effects, the nature of the base model's pre-training data (as indicated by its perplexity scores), and the varying depth of legislative principles covered within case law. Understanding these dynamics is crucial to enhance codification abilities of these models.

A detailed case study is provided in App. E.

5 Conclusion

In this paper, we propose ProMALex, an effective framework for adapting language models to multi-jurisdictional legal corpora. ProMALex leverages hierarchical relationships between jurisdictions, derived from dataset embeddings of a base model, to design a progressive adapter structure. This structure enables sharing adapters for similar jurisdictions in lower layers, while gradually splitting into jurisdiction-specific modules in higher layers. Similarity scores guide the transition layers, and dataset proportions inform parameter allocation across jurisdictions. Experimental results demonstrate that ProMALex effectively disentangles the parameter space, achieving a balance between jurisdiction-specific specialization and cross-jurisdiction inter-

action. Additionally, ProMALex outperforms previous methods in the challenging task of distilling legal principles from case law to structured legislation, effectively mitigating jurisdictional conflicts during pre-training. Future work will focus on refining parameter allocation strategies by considering dataset diversity and enhancing the understanding of similarities across different legal corpora.

Limitations

One key limitation of this study is the diversity of the corpora used in our experiments. While we employed two setups—one in English covering five jurisdictions and another multilingual setup covering four different jurisdictions—the analysis could be extended to cover a broader range of languages and legal systems. Increasing the linguistic and cultural diversity of the corpora will help better assess the robustness of ProMALex across a variety of legal traditions. Moreover, a higher granularity in labeling legal fields within these systems would contribute to a more nuanced understanding of jurisdictional differences and help refine the model's ability to handle a wider array of legal contexts. Additionally, we have used the term "jurisdiction" in a broad sense to represent national legal frameworks, but in countries like India, where individual states have distinct jurisdictional constraints, there are interesting opportunities for future work. Incorporating state-specific variations within a country could provide further granularity and enrich the model's ability to handle intra-national legal complexities.

In our experiments, we used GPT-2 as the baseline model for the language modeling tasks. Future work could extend this analysis to other models, to investigate the effects of memorization in pre-training data. This could provide insights into how memorization influences adaptation and similarity computation in the progressive module design, shedding light on potential improvements in handling long-term legal dependencies and fine-tuning efficiency. While ProMALex has been evaluated on legal language modeling tasks, further work is needed to evaluate its performance on open-ended generation tasks where the model would need to generate legal text that adheres to specific jurisdictional requirements. This would involve assessing not only the model's ability to produce contextually appropriate outputs, but also its capacity to ensure legal accuracy in jurisdiction-specific scenarios. Such evaluations would be critical to determine the practical applicability of ProMALex in real-world legal tasks where nuanced legal reasoning and adherence to jurisdictional-specific norms are paramount.

Ethics Statement

The datasets used in this study are publicly available and sourced from prior works. While some of these datasets are not anonymized and contain real names, we do not anticipate direct harm resulting from our experiments. Nevertheless, the use of such sensitive data requires careful consideration of privacy, fairness, and ethical implications.

We do not advocate for the unchecked deployment of large language models in legal domains without addressing significant challenges. These challenges arise primarily from the pre-training of LLMs on large, diverse datasets, which can reinforce stereotypical patterns embedded in historical data. Legal corpora often reflect long-standing biases, which, when used for training, may perpetuate harmful societal stereotypes. Texts from legal systems may embed historical inequalities related to gender, race, or socio-economic status, and LLMs trained on such data risk inadvertently amplifying these biases, undermining fairness and inclusivity in legal decision-making.

Another significant concern is hallucinated content. LLMs, especially when applied to legal tasks, may generate plausible-sounding but factually incorrect or legally inaccurate outputs. This can lead to considerable harm, including providing incorrect legal advice or judgments. Furthermore, fairness in legal AI is critical, as biases in the underlying data may cause models to perform poorly for certain groups or jurisdictions, leading to inequitable outcomes. For example, models may struggle with low-resource legal systems or fail to account for the distinct legal frameworks of different jurisdictions, resulting in inequitable outcomes.

This work represents technical progress in addressing jurisdiction-specific nuances in legal language models, with a focus on reducing bias and mitigating hallucinated content. By improving the model's ability to handle jurisdictional differences, we aim to promote more equitable outcomes across diverse legal systems. While this paper is a step toward mitigating these challenges, we acknowledge that much work remains. Continued research is necessary, and we stress the importance of trans-

parency, accountability, and ethical consideration in the development and deployment of AI systems in legal domain.

References

Divyanshu Aggarwal, Sankarshan Damle, Navin Goyal, Satya Lokam, and Sunayana Sitaram. 2024. Exploring continual fine-tuning for enhancing language ability in large language model. *arXiv preprint arXiv:2410.16006*.

Muhammad Al-qurishi, Sarah Alqaseemi, and Riad Souissi. 2022. Aralegal-bert: A pretrained language model for arabic legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 338–344.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Alexandra Chronopoulou, Matthew E Peters, and Jesse Dodge. 2021. Efficient hierarchical domain adaptation for pretrained language models. *arXiv preprint arXiv:2112.08786*.

Victor Hugo Ciurlino. 2021. Bertbr: a pretrained language model for law texts.

CivicTech India. Indian law penal code - json. https://github.com/civictech-India/Indian-Law-Penal-Code-Json. Accessed: 2025-01-18.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Filipe Coimbra Pereira de Melo, Gabriel Hautreux, Etienne Malaboeuf, Johanne Charpentier, Dominic Culver, and Michael Desa. 2024a. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024b. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models journal of legal analysis (forthcoming).

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pretrained language models memories. *arXiv preprint arXiv:2306.05406*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. 2024. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. Juribert: A masked-language model adaptation for french legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101.

Christoph Engel and Richard H Mcadams. 2024. Asking gpt for the ordinary meaning of statutory terms. *MPI Collective Goods Discussion Paper*, (2024/5).

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Lingyun Feng. 2023. Learning to predict task transferability via soft prompt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8829–8844.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-loras: An efficient multitask tuning for large language models. *arXiv preprint arXiv:2403.03432*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2021. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish legalese language model and corpora. *arXiv preprint arXiv:2110.12201*.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Thomas F Heston and Charya Khun. 2023. Prompt engineering in medical education. *International Medical Education*, 2(3):198–205.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? *arXiv preprint arXiv:2310.14880*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.

Daniele Licari and Giovanni Comandè. 2022. Italian-legal-bert: A pre-trained transformer language model for italian law. In *CEUR Workshop Proceedings (Ed.), The Knowledge Management for Law Workshop (KM4LAW)*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Quanyu Long, Wenya Wang, and Sinno Jialin Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. *arXiv preprint arXiv:2311.11551*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. Ragstudio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735.

Eric Martínez. 2024. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, pages 1–24.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurbert: A romanian bert model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv preprint arXiv:2403.04890*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023a. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023b. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Joel Niklaus, Lucia Zheng, Arya D McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E Ho, Garrett Honke, Percy Liang, and Christopher Manning. 2024. Flawn-t5: An empirical examination of effective instruction-tuning data mixtures for legal reasoning. *arXiv preprint arXiv:2404.02127*.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: a case study on indian law. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 187–196.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. *arXiv preprint arXiv:2311.11077*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Subendhu Rongali, Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Continual domain-tuning for pretrained language models. *arXiv preprint arXiv:2004.02288*.

TYS Santosh, Mahmoud Aly, Oana Ichim, and Matthias Grabmair. 2025. Lexgenie: Automated generation of structured reports for european court of human rights case law. *arXiv preprint arXiv:2503.03266*.

TYS Santosh, Kevin D Ashley, Katie Atkinson, and Matthias Grabmair. 2024a. Towards supporting legal argumentation with nlp: Is more data really all you need? *arXiv preprint arXiv:2406.10974*.

TYS Santosh, Oana Ichim, and Matthias Grabmair. 2023. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. *arXiv preprint arXiv:2302.00609*.

TYS Santosh, Vatsal Venkatkrishna, Saptarshi Ghosh, and Matthias Grabmair. 2024b. Beyond borders: Investigating cross-jurisdiction transfer in legal case summarization. *arXiv preprint arXiv:2403.19317*.

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

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Yikang Shen, Zheyu Zhang, Tianyou Cao, Shawn Tan, Zhenfang Chen, and Chuang Gan. 2023. Moduleformer: Learning modular large language models from uncurated data. *arXiv preprint arXiv:2306.04640*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Andrea Tagarelli and Andrea Simeri. 2022. Lamberta: Law article mining based on bert architecture for the italian civil code. In *Proc. 18th Italian Research Conference on Digital Libraries*, volume 3160.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

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.

Chunlei Xin, Yaojie Lu, Hongyu Lin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, and Le Sun. 2024. Beyond full fine-tuning: Harnessing the power of lora for multi-task instruction tuning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2307–2317.

Benfeng Xu, Chunxu Zhao, Wenbin Jiang, Pengfei Zhu, Songtai Dai, Chao Pang, Zhuo Sun, Shuohuan Wang, and Yu Sun. 2023. Retrieval-augmented domain adaptation of language models. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 54–64.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. Efficiently tuned parameters are task embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5007–5014.

A Related Work

Continuing pretraining models on domain-specific corpora has consistently demonstrated improvements in downstream task performance (Gururangan et al., 2020). This approach has been adopted in legal NLP, resulting in models such as LegalBERT (Chalkidis et al., 2020), trained on English legal texts from the US, EU, and UK; CaseLawBERT (Zheng et al., 2021), specialized in US federal and state court cases; PoLBERT (Henderson et al., 2022), on the Pile of Law corpus spanning US, Canadian, and EU legal documents; and InLegalBERT (Paul et al., 2023) on Indian legal texts. Most recently, LexLMs (Chalkidis et al., 2023) consolidated these efforts by pretraining on the multinational LeXFiles corpus, encompassing six English-speaking jurisdictions.

Beyond English, several multilingual jurisdiction-specific models have emerged, such as JuriBERT for French legal texts (Douka et al., 2021), JurBERT for Romanian (Masala et al., 2021), LamBERTa and ItalianLegalBERT for Italian (Tagarelli and Simeri, 2022; Licari and Comandè, 2022), RoBERTalex for Spanish (Gutiérrez-Fandiño et al., 2021), Lawformer for Chinese (Xiao et al., 2021), AraLegalBERT for Arabic (Al-qurishi et al., 2022), LCUBE for Korean (Hwang et al., 2022), and LegalBERT-pt and BERTBR for Portuguese (Ciurlino, 2021). Recently, LegalXLM (Niklaus et al., 2023b) introduced a multilingual legal model trained on the MultiLegalPile corpus, covering 24 languages and 17 jurisdictions.

Most of these rely on BERT-style encoder-only architectures. Recent works have explored other architectures, including encoder-decoder models such as LexT5 and FLawN-T5 (Santosh et al., 2024c; Niklaus et al., 2024), pretrained on English and multilingual legal corpora, respectively, decoder-based models such as SaulLM (Colombo et al., 2024b) focusing on English legal text. Instruction tuning has been employed to enhance responsiveness to task-specific prompts in SaulLM and FlawnT5 (Niklaus et al., 2024; Colombo et al., 2024b,a), while preference training techniques like DPO have been employed in SaulLM to align outputs with desired responses using synthetic legal scenarios (Colombo et al., 2024a).

B Datasets

Caselaw Corpus: UK case law is obtained from court decisions available on the British and Irish Legal Information Institute. EU court decisions, primarily issued by the Court of Justice of the European Union, are sourced from EUR-Lex. Canadian decisions originate from the dataset published by Henderson et al. (2022). ECHR case law, ruled by the European Court of Human Rights, is accessed via the HUDOC database. Indian Supreme Court cases are sourced from Indian Kanoon, as provided by Malik et al. (2021). Switzerland's trilingual caselaw corpus (German, French, and Italian) is obtain from from Entscheid Suche. Germany case law in German is from Open Legal Data. Denmark case law in Danish, sourced from DDSC. Belgian case law, available in French, are retrieved from

Jurportal.

We split each document in corpus into 2048 token based chunks. Dataset statistics are provided in Table 4 and 5 for English and Multilingual setups. **Legislation Corpus:** UK legislation is accessed via the UK National Archives, while Canadian acts and regulations are sourced from the official legislation portal of Canada. EU regulations, directives, and decisions are obtained from EUR-Lex. ECHR legislation is based on the European Convention on Human Rights, accessed through the ECHR website, as sourced by Santosh et al. (2023). Indian legislation is derived from the India Code website and sourced from CivicTech India. Dataset statistics are provided in Table 6.

C Implementation Details

We utilize the Hugging Face Transformers library (Wolf et al., 2020) for our experiments. For the English setup, we use GPT-2-XL (Radford et al., 2019) with 1.5 billion parameters, consisting of 48 Transformer blocks and a context window of 1024 tokens. For the multilingual setup, we employ mGPT-XL (Shliazhko et al., 2022) with 1.3 billion parameters, 48 Transformer blocks, and a context window of 2048 tokens. LoRA (Hu et al., 2021) modules are implemented using the PEFT library.

All experiments are conducted on a cluster equipped with four NVIDIA A100 GPUs. The models are trained using the AdamW optimizer (Loshchilov, 2017) with a learning rate of 2×10^{-5}, weight decay of 0.01, and a cosine learning rate schedule. A 10% warmup phase is employed, and mixed-precision training is used to enhance computational efficiency. We set the per-device batch size to 4 and use gradient accumulation across multiple GPUs to achieve an effective batch size, training for 1 epoch to mitigate overfitting on duplicated data. During evaluation, the first half of the tokens (512 for English and 1024 for Multilingual) are used as context, and perplexity is computed over the subsequent half. LoRA matrices with a rank of 64 are added per layer in the self-attention projections (c_attn and c_proj) of GPT-2. The rank is split according to the specific strategies described for each baseline.

D Visualizations of Hierarchical Trees (Dendrograms)

Figures 2 and 3 visualize the hierarchical trees produced using different methods (DES, VOR, MPS) in the English and Multilingual Setup respectively. The similarity scores obtained are scaled to number of layers, indicating the transition points in the progressive adapter design. Parameter allocation after the split is also indicated by specifying the rank of LoRA modules in each corresponding layer.

Based on the results in Table 2, the underperformance of VOR can be attributed to non-calibrated similarity values, a trend also observed in Figures 2 and 3. In contrast, the MPS method produces structures that differ substantially from DES, highlighting the need for further study into how these similarity constraints impact performance. Understanding these differences requires legal expertise, as interpreting the reasoning behind these structures is not trivial. While these insights are valuable for enhancing the alignment of produced structures with legal backing, determining the appropriate similarity structures is a complex task, as legal interpretations often involve nuanced, context-dependent reasoning that is difficult to quantify.

E Case Study

We present a case study to illustrate the effectiveness of ProMALex using judgments from the ECHR corpus. Consider the following excerpts:

Excerpt A *The Court reiterates that an interference with the exercise of freedom of peaceful assembly does not need to amount to an outright ban, whether legal or de facto, ... 55. An interference will constitute a breach of Article 11 unless it was prescribed by law, pursued one or more of the legitimate aims set out in paragraph 2 of that provision and was necessary in a democratic society for the achievement of those aims. . . . . 57. One of the requirements flowing from the expression prescribed by law is foreseeability. Thus, a norm cannot be regarded as a law unless it is formulated with sufficient precision to enable individuals to regulate their conduct; they must be able – if need be with appropriate advice – to foresee, to a degree that is reasonable in the circumstances, the consequences which a given action may entail.*

Analysis On comparing the perplexity values for the above excerpt between the **shared** model and **ProMALex**, we observed significant declines for phrases such as "prescribed by law," "necessary in a democratic society," and "foreseeability." These terms, while appearing general, carry jurisdiction-specific interpretations within the context of the ECHR. For instance, "prescribed by law" in ECHR

	ECHR	EU	India	UK	Canada	Average
			Train			
Documents	10.53K	24.33K	28.82K	39.04K	8.49K	111.20K
Chunks	156.15K	305.52K	189.06K	675.95K	71.01K	1.40M
Tokens	159.90M	312.85M	193.60M	692.18M	72.71M	1431.24M
			Test			
Documents	1.00K	3.17K	3.00K	4.00K	1.40K	12.57K
Chunks	15.31K	60.73K	18.88K	82.85K	11.64K	189.40K
Tokens	15.67M	62.19M	19.33M	84.84M	11.92M	193.95M

Table 4: Dataset statistics of case law corpora for five jurisdictions (ECHR, EU, India, UK, and Canada) in the English setup.

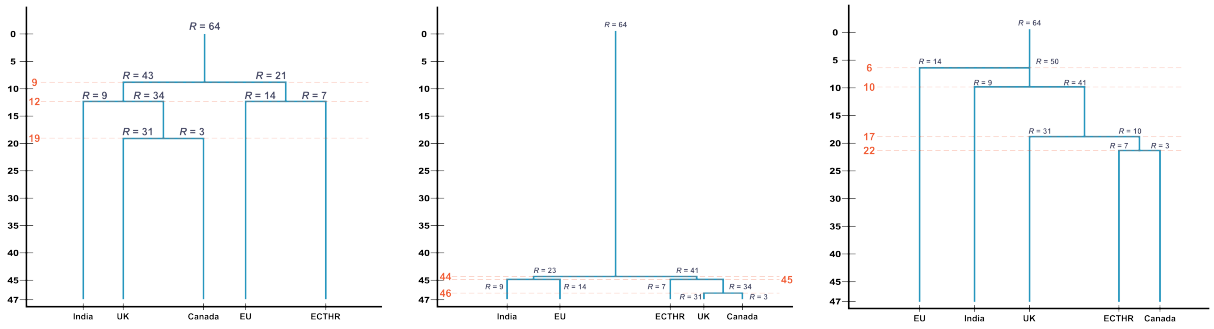
	Switzerland	Germany	Denmark	Belgium	Average
		Train			
Documents	492.80K	161.34K	3.55K	1.78K	659.47K
Chunks	1.90M	559.94K	23.26K	13.56K	2.50M
Tokens	3890.40M	1146.77M	47.63M	27.77M	5112.57M
		Test			
Documents	123.20K	40.34K	0.89K	0.45K	164.87K
Chunks	474.82K	139.00K	5.84K	3.53K	623.19K
Tokens	972.43M	284.66M	11.96M	7.23M	1276.29M

Table 5: Dataset statistics of case law corpora for four jurisdictions (Switzerland, Germany, Denmark and Belgium) in the Multilingual setup.

	ECHR	EU	India	UK	Canada	Average
Documents	33	6.83K	1.93K	4.00K	0.50K	13.29K
Chunks	0.15K	88.16K	2.01K	32.46K	4.01K	126.79K
Tokens	153.60K	90.28M	2.06M	33.23M	4.11M	129.84M

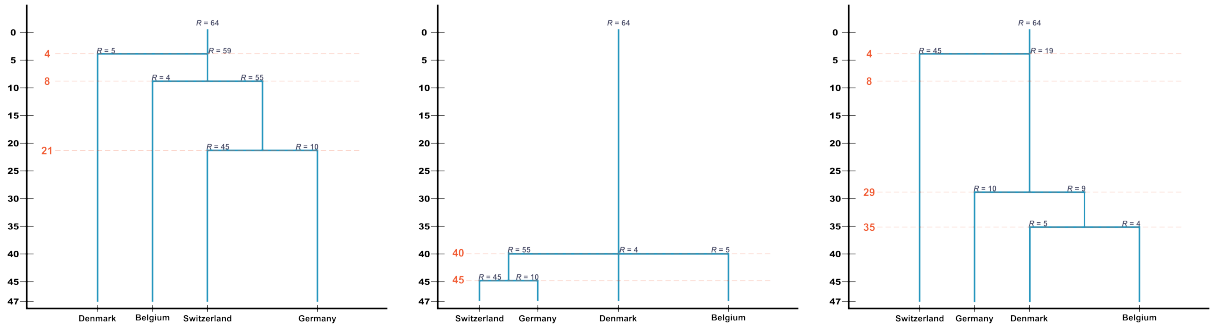
Table 6: Dataset statistics of the legislation corpora for five jurisdictions (ECHR, EU, India, UK, and Canada) in the English setup.

jurisprudence requires not only the existence of a legal basis but also adherence to principles of accessibility, foreseeability, and compatibility with the rule of law, emphasizing the protection of individuals from arbitrary state action. Similarly, "necessary in a democratic society" embodies the ECHR's proportionality test, balancing the protection of fundamental rights with legitimate public interest—a principle that varies significantly across legal systems. The term "foreseeability," while potentially appearing in other jurisdictions, is particularly nuanced in ECHR case law. It demands that laws be formulated with sufficient precision to enable individuals to foresee the legal consequences of their actions, a standard that reflects the Court's commitment to legal clarity and the rule of law. In other jurisdictions, "foreseeability" might focus more narrowly on criminal liability or contractual obligations, underscoring the importance of understanding its contextual application within ECHR. ProMALex demonstrates its ability to internalize and apply these nuanced principles effectively. ProMALex accurately captures the contextual subtleties essential for interpreting and reasoning within ECHR jurisprudence, underscoring its robust understanding of jurisdictional specifity without conflating with other jurisdictions in a multi-jurisdictional legal language modeling,



(a) DES (b) VOR (c) MPS

Figure 2: Dendrograms produced using different methods (DES, VOR, MPS) applied to the five jurisdictions in the English Setup. The vertical axis represents the scaled similarity scores, with respect to the number of layers in the base model, highlighting the transition points of progressive adapters. Parameter allocation after the split is also indicated by specifying the rank of LoRA modules in each corresponding layer at transition points.



(a) DES (b) VOR (c) MPS

Figure 3: Dendrograms produced using different methods (DES, VOR, MPS) applied to the four jurisdictions in the Multilingual Setup. The vertical axis represents the scaled similarity scores, with respect to the number of layers in the base model, highlighting the transition points of progressive adapters. Parameter allocation after the split is also indicated by specifying the rank of LoRA modules in each corresponding layer at transition points.

where precise interpretation is critical.

Excerpt B . . . . . *It follows that the applicants' initial arrest and detention were compatible with Article 5 § 1 (c) of the Convention, that this part of the application is manifestly ill-founded within the meaning of Article 35 § 3 of the Convention, and that it must be rejected within the meaning of Article 35 § 4. 112. An approach entailing the sort of global examination under the two limbs of Article 5 § 1 (c) which had been carried out in line with Steel and Others (cited above) was also followed the same year in McBride v. the United Kingdom ((dec.), no. 27786/95, 5 July 2001), ending in the conclusion that the complaint was manifestly ill=founded. 113. What transpires from the above is that the Court's approach in Steel and Others in 1998 (cited above) and the two follow-up decisions from 2001 (Nicol and Selvanayagam and McBride, both cited above) does not seem consistent with the Ciulla dictum from 1989, later repeated in several cases from Jėčius from 2000 to Ostendorf from 2013, to the effect that Article 5 § 1 (c) only covers deprivation of liberty in connection with criminal proceedings. Nor does it* . . . . . .

Analysis B For this excerpt, ProMALex showed a marked reduction in perplexity for phrases such as "manifestly ill-founded," "Ciulla dictum," and references to landmark cases like Steel and Others (1998), McBride (2001), and Ostendorf (2013). These phrases encapsulate complex procedural and doctrinal shifts within ECHR jurisprudence, requiring a nuanced understanding of case law and procedural principles. For example, "manifestly ill-founded" is a legal term specific to the Convention system, denoting complaints that fail to meet admissibility standards under Article 35 § 3. Similarly, references to Ciulla (1989) and subsequent rulings like Jėčius (2000) and Ostendorf

(2013) reflect evolving interpretations of Article 5 § 1 (c), particularly the tension between preventive detention and deprivation of liberty within criminal proceedings. By robustly associating legal terminology with relevant precedents and procedural principles, ProMALex excels in understanding the nuanced, jurisdiction-specific contexts of ECHR legal texts. This highlights ProMALex's potential to facilitate domain-specific reasoning and effective legal language modeling.

These case studies underscore the strength of ProMALex in handling the jurisdiction-specific nuances required in multi-jurisdictional legal language models.