Unlocking Legal Knowledge: A Multilingual Dataset for Judicial Summarization in Switzerland

Anonymous ACL submission

Abstract

Legal research is a time-consuming task that most lawyers face on a daily basis. A large part of legal research entails looking up relevant caselaw and bringing it in relation to the case at hand. Lawyers heavily rely on summaries (also called headnotes) to find the right cases quickly. However, not all decisions are annotated with headnotes and writing them is time-consuming. Automated headnote creation has the potential to make hundreds of thousands of decisions more accessible for legal research in Switzerland alone. To address this, we introduce the Swiss Landmark Decisions Summarization (SLDS) dataset, a cross-lingual resource with 20K landmark rulings from the Swiss Federal Supreme Court, each with headnotes in German, French, and Italian. We fine-tune models from the Qwen2.5, Llama 3.2, and Phi-3.5 families and compare them to larger proprietary models, including GPT-4o and Claude 3.5 Sonnet, and DeepSeek R1. While fine-tuned models achieve high lexical similarity, proprietary models excel in legal accuracy and coherence, as shown by an LLM-as-a-Judge evaluation.Our evaluation reveals that while fine-tuned models achieve strong lexical similarity, proprietary models generate more legally accurate and structured headnotes. Surprisingly, reasoning models do not significantly outperform general-purpose LLMs, indicating that structured factual accuracy is more crucial than deep logical reasoning in judicial summarization. To advance research in cross-lingual legal summarization, we release SLDS under a CC BY 4.0 license.

1 Introduction

A significant part of legal work involves research, where lawyers must find similar cases and navigate numerous judicial decisions, especially when interpreting laws with room for debate. Due to the time-intensive nature of this task, they usually rely on judgment summaries. However, creating these summaries is labor intensive and requires the expertise of judges and clerks, who are already burdened with a large caseload (Bieri, 2015) and time pressure (Ludewig and Lallave, 2013).

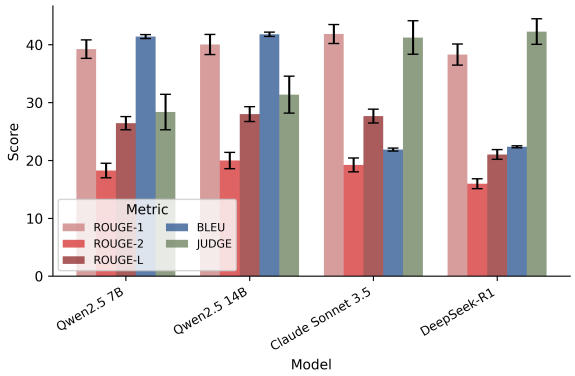


Figure 1: Results of two fine-tuned LLMs of the Qwen2.5 model family and two large pre-trained models evaluated on the test set of SLDS. While fine-tuning dominates outcomes in terms oflexical metrics, the smaller fine-tuned models do not yet reach the same output quality as their larger pre-trained counterparts, as indicated by the LLM-as-a-Juge (Zheng et al., 2023) score.

To alleviate this increasing need for efficient ways to navigate large amounts of legal documents, legal document summarization has become a critical area of interest in NLP (Jain et al., 2021). Over the years, researchers have made significant strides in both extractive and abstractive summarization of legal texts. Earlier work focused on extracting key sentences to create concise summaries (Grover et al., 2004; Hachey and Grover, 2006; Kim et al., 2013; Bhattacharya et al., 2021), while recent advancements have turned toward abstractive methods, which generate condensed paraphrases of the most important information in a document (Shukla et al., 2022; Niklaus and Giofré, 2022; Moro et al., 2023; Jain et al., 2024; Niklaus et al., 2024).

Datasets with legal documents and their corresponding summaries have been instrumental in enabling these advancements, yet they primarily focus on monolingual corpora or multiple jurisdictions. Therefore, existing datasets do not ade-

quately address the unique challenges posed by multilingual jurisdictions, such as Switzerland, where legal decisions are written in multiple languages and need to be summarized consistently. This gap is particularly relevant because many legal NLP tools and models are trained on English-centric datasets, which may not reliably generalize to cross-lingual environments.

We introduce the Swiss Landmark Decision Summarization (SLDS), a large-scale multilingual dataset of Swiss Supreme Court cases in German, French, and Italian, featuring headnotes that summarize key legal points and laws. By focusing on these concise legal digests, SLDS facilitates cross-lingual legal summarization research and supports the development of tools for professionals working across language barriers. The dataset is publicly available under a CC BY 4.0 license.[1]

Contributions Our contributions are two-fold:

1. **SLDS Dataset Release:** We introduce and publicly release the SLDS dataset, a large-scale, cross-lingual legal resource. It comprises 20K rulings from the Swiss Federal Supreme Court (SFSC) in German, French, or Italian, each accompanied by summaries in all three languages—resulting in 60K data rows. By making SLDS openly available, we aim to support and encourage multilingual legal NLP research.
2. **Comprehensive Benchmarking:** We fine-tune multiple models from the Qwen, Llama, and Phi families—including five Qwen variants, Llama 3.2 3B, and Phi-3.5-mini — and compare their performance to proprietary models (GPT-4o, Claude 3.5 Sonnet, and o3-mini) as well as the pre-trained DeepSeek R1 in a one-shot setting. Our evaluation, combining conventional summarization metrics with an LLM-as-a-Judge approach, highlights the trade-offs between fine-tuning and prompting while revealing the limitations of standard metrics in capturing the nuances of legal summarization.

2 Related Work

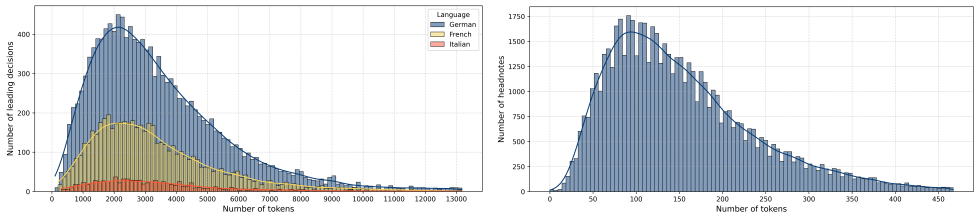
Recent research on legal text summarization has increasingly focused on abstractive summarization, leading to the creation of datasets tailored for fine-tuning pre-trained language models. Among monolingual datasets, two major English corpora stand out. BillSum (Kornilova and Eidelman, 2019) consists of 22K U.S. congressional and state bills and applies extractive models, including a BERT classifier and an ensemble method, to generate summaries using Maximal Marginal Relevance. The dataset also enables transfer learning for summarization across federal and state laws. Multi-LexSum (Shen et al., 2022) focuses on long civil rights lawsuits (75K+ words) and uniquely allows summary evaluation at different lengths (25, 130, and 650 words), leveraging BART and PEGASUS models. Additionally, Bauer et al. (2023) extracted key passages from 430K U.S. court opinions, with results favoring a reinforcement-learning-based model over transformers, though their dataset remains unavailable due to licensing issues. In Portuguese, RulingBR (de Vargas Feijó and Moreira, 2018) offers over 10K Brazilian Supreme Court rulings, with summaries divided into four distinct components, one of which serves as the reference for training.

Multilingual legal summarization datasets include EUR-Lex-Sum (Aumiller et al., 2022), covering 24 EU languages and aligning 375 legal acts across all languages. Unlike court decisions, which involve complex legal reasoning and factual narratives, legal acts follow a more structured format. Compared to this dataset, our work emphasizes case law within a single jurisdiction, providing over 13 times more cross-lingual samples for French-to-German summarization and more than double for Italian-to-German, enabling a more detailed analysis of these language pairs in the Swiss legal context. Another relevant dataset, MILDSum (Datta et al., 2023), addresses language barriers in the Indian legal system by translating 3,000 English legal case judgments into Hindi. A key finding was that a Summarize-then-Translate approach outperformed direct cross-lingual summarization, with summaries averaging 700 tokens. Unlike MILDSum, our dataset does not include English and is structured around headnotes, making the summarization task more challenging due to the dominance of English in pre-training corpora.

3 Data

We introduce SLDS, a novel dataset for cross-lingual summarization in the legal domain. It comprises over 20K landmark decisions published by the SFSC in German, French, or Italian, each ac-

[1] Link available upon acceptance



(a) Decisions (b) Headnotes

Figure 2: Distributions of token counts in (a) landmark decisions and (b) headnotes. To improve readability, only samples within the 99th percentile were included, as the long tail of the distribution would have otherwise skewed the visualization. Tokenization was performed using the `tiktoken` library with the `o200k_base` encoding.

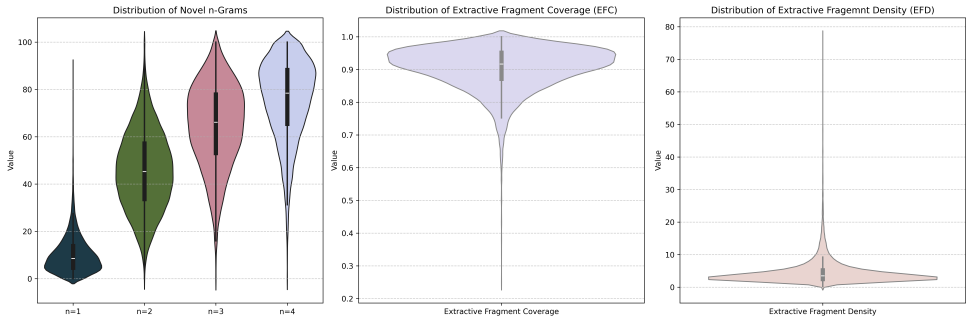


Figure 3: Distribution of Summarization Properties in SLDS. The figure illustrates n-gram novelty (left), Extractive Fragment Coverage (EFC) (center), and Extractive Fragment Density (EFD) (right), highlighting the dataset's balance between abstraction and extractiveness.

companied by paragraph-aligned summaries written by clerks and judges in all three languages. This dataset provides a valuable resource for studying cross-lingual summarization, a relatively underexplored area in legal NLP. Unlike datasets such as EUR-Lex-Sum, which focus on legislation, SLDS centers on judicial decisions, making it particularly relevant for developing tools to assist legal practitioners and researchers working with court rulings.

3.1 Data Collection

Decisions were scraped from the official Swiss Federal Supreme Court repository[2], covering 70 years and five legal volumes. We extracted the full decision text, which was either in German, French or Italian, along with the headnotes in all three languages. We also stored and inferred metadata including the year of the decision, the volume in which the decision was published, the law area of the decision which can be inferred from the volume and the year, and the url to the official published decision on the repository. To enable model training and cross-lingual evaluation, each row contains one decision-headnote pair, tripling the dataset to over 60K samples. The exact fields of our dataset can be seen in Appendix C.1.

[2]Available at https://www.bger.ch/

3.2 General information

Dataset Splits The dataset is partitioned by publication year to prevent data leakage and maintain consistency with current summarization styles. As shown in Table 1, the training set spans 1954–2021, the validation set covers 2022, and the test set includes 2023–2024, resulting in approximately 60k, 600, and 978 samples per split. For a detailed year-wise distribution, see Appendix C.2.

Table 1: Dataset splits by publication years and language distribution of decisions.

Split	Years	# Decisions	# Samples	Languages (%)
Train	1954–2021	∼20,000	∼60,000	DE: 67.94, FR: 27.36, IT: 4.71
Validation	2022	200	600	DE: 68.50, FR: 27.50, IT: 4.00
Test	2023–2024	326	978	DE: 63.50, FR: 32.82, IT: 3.68

Text Length Figure 2 shows the number of tokens for both decisions and the headnotes up to the 99th percentile. Decisions range from 102 to 44.3k tokens. The median decision length is 2971 tokens, and the mean decision length is 3585 tokens with a standard deviation of 2629 tokens.

3.3 Summarization-related properties

To analyze the summarization tendencies in SLDS, we examine Compression Ratio (CR), Extractive Fragment Coverage (EFC), Extractive Fragment Density (EFD) (Grusky et al., 2018a), and n-gram novelty (Narayan et al., 2018). Given the dataset's legal and multilingual nature, we compare these

properties to EUR-Lex-Sum (Aumiller et al., 2022) and MILDSum (Datta et al., 2023). Since EFC, EFD, and n-gram novelty rely on lexical overlap, we compute them only for headnotes in the same language as their reference decisions.

Figure 3 visualizes the distributions of these properties, capturing both abstractive (n-gram novelty) and extractive (EFC, EFD) characteristics across dataset splits. This approach provides a more comprehensive view than summary statistics alone.

Compression Ratio To calculate the Compression Ratio (CR), we use the `tiktoken` tokenizer with the `o200k_base` encoding that is currently used by GPT-4o and GPT-4o-mini. We calculate it by dividing the number of tokens in the decision by the number of tokens in the headnote. The observed mean CRs of 26.39 is much larger than the CRs reported in EUR-Lex-Sum and MILDSum, highlighting the conciseness of headnotes in the Swiss judicial system. In the case of landmark decisions, headnotes not only summarize the decision but most importantly highlight the key points that made the decision a *landmark* decision in the first place, serving as a reference for future jurisprudence. This elevated CR makes the generation of headnotes an even harder task than other summarization tasks in the legal domain. Interestingly, we observe even higher CRs in the validation and the test split, indicating that there is a trend towards shorter headnotes.

Extractive Fragments To measure how extractive the headnotes in our dataset are, i.e., how much of the headnote is directly copied from the decision, we compute both Extractive Fragment Coverage (EFC) and Extractive Fragment Density (EFD) values as introduced by Grusky et al. (2018b). Interestingly, we observe exactly the same mean EFC as the one reported in MILDSum. While this can be an indicator of high extractiveness, we argue that this value is naturally higher for longer reference texts with high compression ratios, since the few unigrams present in the summaries have a higher probability of appearing in the original reference text. On the other hand, we observe a mean EFD of 4.63, much lower than the 24.42 for MILDSum. This suggests a lower extractiveness and therefore higher degree of abstractivity of the headnotes in SLDS. We do however observe higher EFC and EFD values for the validation and the test set. This goes in line with the higher CRs observed in those splits.

n-Gram Novelty We further investigate the abstractivity of our dataset by computing the percentage of novel n-grams appearing in the headnotes when compared to the decisions, as proposed by Narayan et al. (2018) and also evaluated by Aumiller et al. (2022) in EUR-Lex-Sum. The minimum n-gram novelty of 0 for some samples suggests that there are headnotes that are entirely extractive. On the other hand, there also seem to be headnotes that are entirely abstractive, as indicated by the maximum n-gram novelty value of 100. The results in Table 3 show that on average about 90% of the words appearing in the headnote are present in the decisions as well. For the test set, only about 5% of words are novel. The picture changes when looking at the bigrams, trigrams and quadgrams. There we observe a much higher novelty, indicating that the headnotes may use the same words as in the decisions, but they are combined differently, making the texts more abstractive. Nonetheless, about 30% of all quadgrams in the headnotes appearing in the test set are copied from the corresponding decision. In conclusion, our dataset lies somewhere in between the spectrum of extractive and abstractive summaries, with some outlier headnotes being entirely extractive and others being fully abstractive.

3.4 Licensing

We release the dataset under the CC-BY-4.0 license, which complies with the SFSC licensing[3].

3.5 Ethical Considerations

Due to the sensitive nature of court cases and their corresponding rulings, the SFSC anonymizes personal or sensitive information according to their guidelines[4] before publishing them online.

4 Experimental Setup

To establish baselines, we evaluate four frontier Large Language Models (LLMs): GPT-4o, Claude 3.5 Sonnet, DeepSeek R1, and o3-mini[5]. We used a one-shot setting to make them familiar with the

[3]For more information, see https://www.bger.ch/files/live/sites/bger/files/pdf/de/urteilsveroeffentlichung_d.pdf

[4]Anonymization guidelines at https://www.bger.ch/home/juridiction/anonymisierungsregeln.html

[5]We used the following model names with timestamps: `gpt-4o-2024-08-06`, `o3-mini-2025-01-31`, `claude-3-5-sonnet-20241022`

Table 2: Results of the baseline experiments on the test set of SLDS. The reported metrics are macro-averages over the test subsets consisting of nine different language combinations of decision and headnote language. Standard errors are estimated using the bootstrapping mechanism implemented in lighteval (Fourrier et al., 2023). For BERTScore we report the F1 score. The ROUGE scores are multiplied with a factor of 100 for consistency. JUDGE = LLM-as-a-Judge. **Bold**: best overall; underlined: best within setup.

Model	Setting	BERTScore ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	JUDGE ↑
Phi-3.5-mini	fine-tuned	11.24 ± 3.82	34.84 ± 0.41	0.31 ± 0.02	0.14 ± 0.01	0.21 ± 0.01	15.25 ± 2.32
Llama 3.2 3B	fine-tuned	15.20 ± 4.40	21.89 ± 0.42	0.32 ± 0.02	0.15 ± 0.02	0.22 ± 0.02	18.47 ± 2.99
Qwen2.5 0.5B	fine-tuned	-1.37 ± 3.85	32.20 ± 0.35	0.24 ± 0.02	0.09 ± 0.01	0.17 ± 0.01	5.80 ± 1.26
Qwen2.5 1.5B	fine-tuned	19.81 ± 2.72	36.79 ± 0.34	0.33 ± 0.02	0.14 ± 0.01	0.23 ± 0.01	15.92 ± 2.27
Qwen2.5 3B	fine-tuned	23.23 ± 2.80	38.42 ± 0.34	0.35 ± 0.02	0.16 ± 0.01	0.24 ± 0.01	20.31 ± 2.66
Qwen2.5 7B	fine-tuned	29.59 ± 1.97	41.40 ± 0.34	0.39 ± 0.02	0.18 ± 0.01	0.26 ± 0.01	28.37 ± 3.07
Qwen2.5 14B	fine-tuned	**32.48** ± 1.98	**41.8** ± 0.37	0.40 ± 0.02	**0.20** ± 0.01	**0.28** ± 0.01	31.38 ± 3.19
GPT-4o	one-shot	30.44 ± 1.74	31.89 ± 0.25	**0.42** ± 0.02	0.19 ± 0.01	0.26 ± 0.01	39.7 ± 2.66
Claude 3.5 Sonnet	one-shot	-11.91 ± 18.88	21.88 ± 0.25	**0.42** ± 0.02	0.19 ± 0.01	**0.28** ± 0.01	41.25 ± 2.90
DeepSeek-R1	one-shot	20.28 ± 1.45	22.37 ± 0.18	0.38 ± 0.02	0.16 ± 0.01	0.21 ± 0.01	**42.28** ± 2.21
o3-mini	one-shot	14.18 ± 1.31	20.55 ± 0.17	0.35 ± 0.01	0.12 ± 0.01	0.18 ± 0.01	34.82 ± 2.41

expected headnote format. Additionally, we fine-tuned three Small Language Models (SLMs) in the 3-4B parameter range: Llama 3.2 3B (Dubey et al., 2024), Qwen2.5 3B (Yang et al., 2024), and Phi-3.5-mini (Abdin et al., 2024) on the training split of our dataset. To measure the effect of model size on summarization performance, we also fine-tuned several variants of the Qwen2.5 model family with 0.5B, 1.5B, 3B, 7B, and 14B parameters. Since these models have learned how to generate headnotes, we then evaluate them in a zero-shot manner. Appendix Appendix E shows more infromation about fine-tuning hyperparameters.

4.1 Traditional Metrics

We evaluate the models on the test set of our dataset using the lighteval library (Fourrier et al., 2023)[6]. The tasks are evaluated using BERTScore (Zhang et al., 2020), BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004). Since each individual metric has inherent weaknesses (Zhang et al., 2020), it is necessary to employ multiple metrics for a more comprehensive assessment.

4.2 LLM-as-a-Judge

We also employ an evaluation of the generated headnote using a LLM-as-a-Judge (Zheng et al., 2023) approach. Because judge models favor their own output (Panickssery et al., 2024) and because DeepSeek V3 (Liu et al., 2024) is both cheap and powerful, we used this model as our judge LLM. For cost reasons we did not show the humans the decision texts. To make the llm judge comparabe to the humans we also did not show the decision to the LLMs judge. Because the original headnotes are manually written and of good quality, we believe that this approach still works well while being much more token-efficient.

Implementation Details To further evaluate the quality of generated headnotes, we employ an LLM-as-a-**Judge** approach (Zheng et al., 2023), where a LLM assesses the generated headnote against the official (gold) headnote based on five key evaluation criteria. We show the system and user prompt used for the evaluations in Appendix F.3, and an example of the generated reasoning and the scores in Appendix G.1.

The system prompt instructs the LLM to assess the generated headnote in comparison to the official headnote across five categories: (1) Accuracy & Faithfulness, (2) Completeness & Relevance, (3) Clarity & Coherence, (4) Articles, and (5) Considerations. For each category, the LLM provides a brief analysis followed by a score on a scale from 1 to 3, where 1 indicates major flaws, 2 denotes minor omissions or inaccuracies, and 3 signifies a close match to the official headnote.

The user prompt presents both headnotes and explicitly guides the LLM through the evaluation process. The output format is predefined to ensure structured and parsable results, with category-specific analyses followed by categorical scores. An example evaluation is provided within the prompt to reinforce consistency.

The structured format allows for both automated parsing and human review, contributing to a scalable and cost-effective evaluation process. The exact instructions and descriptions of the five evaluation categories can be seen in the prompts provided in Appendix F.3. To see an example of the generated reasoning and the individual scores, please

[6]We will add several pull requests to include the community task to the official ligtheval repository upon acceptance to ensure reproducibility and make future evaluations easier to perform.

refer to Appendix G.1.

Aggregation To compute the finial score, we first transform the generated scores to a range from 0 to 2. Then we sum up the five scores belonging to a single sample and divide it by 10 - the maximum number of points that a sample can receive in total. These values ranging from 0 to 1 are then multiplied by 100 and aggregated over the entire test set into a mean judge score.

4.3 Human Evaluations

To get a trusted estimate of the quality of generated headnotes, we randomly sampled 7 samples per decision-headnote language pair from the test set, resulting in a set of 63 samples, which were evaluated in the same way as by the LLM judge but this time by two co-authors of this paper who are professional and experienced lawyers. Both of them are fluent in the corresponding languages (German, French or Italian). The evaluation was performed using the original headnotes along with three generated headnotes from the best performing models out of the following three categories: fine-tuned models, frontier models, reasoning models.

The two lawyers were instructed to perform the same evaluation as the LLM judge. Additionally, we selected a subset of nine of these samples and instructed another professional and experienced lawyer and co-author of this paper to perform an in-depth qualitative analysis of the generated headnotes, taking into account also the landmark decisions themselves and not only the original headnote.

5 Results

5.1 Overall Results

The results of our evaluations on the test split of SLDS are in Table 2. We macro-averaged over the scores in each of the nine language subsets of decision and headnote language pairs to promote model's fairness and robustness across languages. We highlight several interesting observations.

Most automated metrics favor fine-tuned models While the SLMs perform worse in terms of the JUDGE metric as their larger counterparts, the data shows that the fine-tuned Qwen2.5 14B outperforms even much larger proprietary models according to BERTScore, BLEU, ROUGE-2, and ROUGE-L. For and ROUGE-2, the scores for Qwen are also comparably high. This suggests that our fine-tuned models prioritize lexical similarity but struggle with legal accuracy, completeness, and structure when compared to the larger proprietary LLMs. Moreover, it highlights the limitations of traditional evaluation metrics and underlines the importance of more sophisticated evaluations using LLMs as judges.

Large models are more accurate The results indicate that larger models are better at generating headnotes that are legally accurate, complete and faithful, as indicated by the higher judge scores. While this was expected, we think that it could be partially due to the one shot examples provided in the prompt. We initially planned to do a one-shot evaluation for the fine-tuned models as well, but we found that it did not improve the performance, possibly because the model already learned what a headnote is supposed to look like during fine-tuning. Another interesting observation is that Claude 3.5 Sonnet performs second best in the judge score but has a negative BERTScore, worse than any other model. This shows that certain metrics can be deceptive and that relying on a single metric for evaluating summaries is usually not sufficient.

Not that much reasoning required An interesting finding is that the reasoning models do not perform significantly better. Even though DeepSeek R1 outperforms all other models in terms of the judge score, the margin to Claude 3.5 Sonnet is only small. Moreover, o3-mini only beats our fine-tuned Qwen2.5 14B model by roughly 3.4 points. These results suggest that summarization in the context of headnotes may not require additional deep logical reasoning but rather strong factual accuracy, domain-specific knowledge, and structured content generation. The task primarily demands models to faithfully extract and concisely rephrase key legal principles, ensuring that references to legal articles and considerations remain intact. Given that general-purpose models such as GPT-4o and Claude 3.5 Sonnet achieve similar or better judge scores than reasoning models, this indicates that current LLMs already possess sufficient reasoning capabilities for this summarization task.

5.2 Cross-lingual Subsets

We report cross-lingual results based on the decision and headnote language (*subsets*), e.g., `de_fr` for decisions in German with French headnotes.

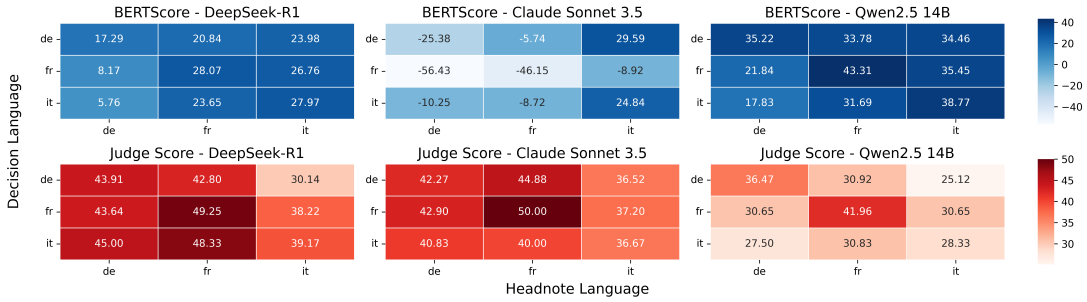


Figure 4: Visualization of the BERT and JUDGE scores for different cross-lingual language subsets and different models. Darker colors indicate better scores.

Key findings are summarized below (full details in Appendix Table 4).

Qwen2.5 14B struggles with cross-lingual consistency While Qwen2.5 14B performs well in monolingual French (fr→fr), its scores drop significantly when the headnote language differs from the decision language, particularly for German and Italian sources. This suggests that its *cross-lingual robustness is weaker* despite strong monolingual performance.

French subsets perform best Both monolingual (fr→fr) and cross-lingual (de→fr, it→fr) French subsets tend to achieve higher scores than their German or Italian counterparts. This may indicate *stronger model proficiency in French legal text generation* or that *French headnotes are more systematic and easier to reproduce*.

Challenges with Italian Italian monolingual generation (it→it) consistently yields lower scores across models. Possible reasons include *less training exposure* to Italian legal texts or the complexity of faithfully replicating Italian legal writing.

Limitations of general-purpose metrics BERTScore heatmaps reveal discrepancies with judge scores, highlighting the need for *domain-specific evaluation*. Some model outputs with low BERTScore still score highly in *legal correctness and completeness*, emphasizing that BERTScore alone is insufficient for assessing legal precision.

5.3 Human Evaluation

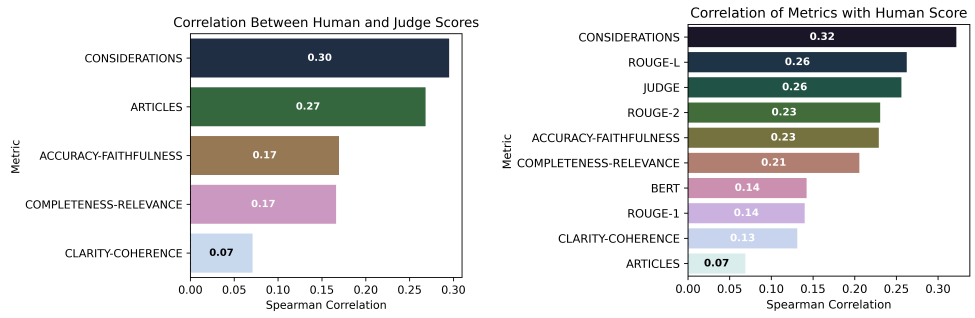
We perform two human evaluations. The first is based on the same evaluation process that the LLM judge also follows. Two lawyers assess three generated headnotes across 63 samples. This evaluation only considers the generated and the original headnote without taking into account the actual text of the landmark decision, assuming that the gold headnote is the ideal headnote and that any deviation should be penalized. We refer to this evaluation as *Human-as-a-Judge*. In the second evaluation which we will refer to as *Contextualized Human Analysis*, another lawyer looked at 6 of those 63 samples and performed an in-depth analysis which involved looking at the decision text as well.

5.3.1 Human-as-a-Judge

With 63 decisions and headnotes generated by 3 models, we obtained 189 annotated samples. Figure 8 illustrates the score distributions assigned by both the LLM and the lawyers. The latter tend to give slightly higher scores than the DeepSeek-V3 judge, with a mean difference of 11.64, indicating that the LLM is stricter in its assessments.

Evaluation Metrics Figure 5 presents two correlation analyses assessing our legal headnote evaluation. Figure 5a shows Spearman correlations between DeepSeek-V3's category-specific scores and human expert ratings across five dimensions: Considerations, Articles, Accuracy-Faithfulness, Completeness-Relevance, and Clarity-Coherence. Figure 5b compares traditional metrics (ROUGE, BERTScore) and LLM-based judgments with aggregated human scores. These analyses reveal how well automated evaluation aligns with expert assessments.

Correlation Analysis Figure 5 reveals important patterns in how automated evaluation approaches align with human judgment. Examining the category-wise correlations in Figure 5a, we find that objective elements of legal analysis show the strongest agreement between human and LLM evaluators. The *Considerations* and *Articles* categories demonstrate the highest correlations (0.30 and 0.27 respectively), suggesting that LLMs are most reliable when evaluating concrete, verifiable aspects of legal headnotes. However, the markedly lower correlation in *Clarity & Coherence* (0.07) highlights a crucial limitation: automated systems



(a) Correlations across categories

(b) Correlations across metrics

Figure 5: Spearman correlations between (a) DeepSeek-V3 and human-assigned category scores and (b) various metrics and aggregated human scores. While LLM scores vary across categories, the overall JUDGE score remains highly correlated with human judgment. Notably, the considerations score, comprising 1/5 of JUDGE, shows the strongest correlation with aggregated human scores.

struggle to assess the more nuanced, subjective aspects of legal writing that human experts evaluate with ease.

Metric Comparison The analysis of different evaluation metrics in Figure 5b reveals the complementary strengths of traditional and LLM-based evaluation approaches. While ROUGE-L and the overall JUDGE score show moderate correlation with human assessment (both at 0.26), the distribution of correlations across metrics suggests that no single automated measure fully captures the complexity of human evaluation. Traditional metrics like BERTScore and ROUGE variants (ranging from 0.14 to 0.26) perform comparably to LLM-based assessments, indicating that the challenges in automated evaluation persist even with advanced language models. This finding underscores the importance of combining multiple evaluation approaches when assessing legal document generation, as different metrics capture distinct aspects of document quality that align with human judgment.

5.3.2 Contextualized Human Analysis

In addition to quantitative evaluation metrics, we conducted a qualitative assessment of model-generated headnotes with a lawyer. The expert reviewed six Swiss landmark decisions along with their original headnotes and the outputs generated by Claude 3.5 Sonnet, DeepSeek R1, and our fine-tuned Qwen2.5 14B model. While all models successfully captured the general themes of the decisions, significant variations were observed in terms of reference accuracy, legal precision, and headnote appropriateness.

The expert found that DeepSeek R1 produced headnotes that aligned closely with the original ones in terms of coverage and completeness, but often included excessive detail, making them more akin to case summaries than concise headnotes. Claude 3.5 Sonnet demonstrated strengths in readability and in capturing the core judgment but introduced occasional legal misinterpretations, including statements that contradicted or over-simplified aspects of the decision. Qwen2.5 14B, fine-tuned on our dataset, showed notable improvements in referencing relevant legal provisions, including the European Convention on Human Rights (ECHR), which was not cited in the original headnote but was deemed relevant. However, the model also introduced incorrect legal references in some cases and sometimes inferred conclusions absent from the decision text. Additionally, all models exhibited inconsistencies in how they structured information, affecting their suitability for legal practitioners.[7]

6 Conclusions and Future Work

We introduced SLDS, a large-scale cross-lingual resource for judicial summarization. Our benchmarking study compared fine-tuned and proprietary models, revealing a trade-off between lexical similarity and legal accuracy. While fine-tuned models performed well on traditional summarization metrics, they struggled with legal correctness, as shown by our LLM-as-a-Judge evaluation. Proprietary models demonstrated higher legal faithfulness and structured output. Notably, reasoning models did not significantly outperform general-purpose LLMs, suggesting that headnote generation requires domain-specific precision rather than complex reasoning.

[7] We will provide a detailed breakdown of the expert analysis in the appendix in the camera-ready version of the paper.

Limitations

Our LLM-as-a-Judge evaluation showed only a moderate correlation with human judgments, suggesting that more sophisticated prompting strategies could improve alignment in future work. Additionally, we lack Inter-Annotator Agreement, as each lawyer annotated a different subset of samples, introducing potential subjectivity due to resource constraints and the high cost of legal annotations.

While we experimented with fine-tuned small and mid-sized models, we did not explore fine-tuning larger-scale models that benefit from scaling laws. It remains an open question whether such models could close the gap with proprietary systems while maintaining efficiency. Future research should investigate the impact of scaling laws on legal coherence and factual accuracy, as well as refine prompting techniques to enhance both headnote generation and LLM-as-a-Judge evaluation. We hope that SLDS will foster progress in multilingual legal NLP and the development of more reliable judicial summarization systems.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. *arXiv preprint*. ArXiv:2210.13448 [cs].

Emmanuel Bauer, Dominik Stammbach, Nianlong Gu, and Elliott Ash. 2023. Legal extractive summarization of u.s. court opinions. *Preprint*, arXiv:2305.08428.

Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 22–31.

Peter Bieri. 2015. Law clerks in switzerland-a solution to cope with the caseload? In *IJCA*, volume 7, page 29. HeinOnline.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.

Diego de Vargas Feijó and Viviane Pereira Moreira. 2018. Rulingbr: A summarization dataset for legal texts. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 255–264. Springer.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Clémentine Fourrier, Nathan Habib, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Claire Grover, Ben Hachey, and Ian Hughson. 2004. The HOLJ Corpus. Supporting Summarisation of Legal Texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 47–54, Geneva, Switzerland. COLING.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018a. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018b. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. *arXiv:1804.11283 [cs]*. ArXiv: 1804.11283.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

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*.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2024. Summarization of lengthy legal documents via abstractive dataset building: An extract-then-assign approach. *Expert Systems with Applications*, 237:121571.

Mi-Young Kim, Ying Xu, and Randy Goebel. 2013. Summarization of legal texts with high cohesion and automatic compression rate. In *New Frontiers in Artificial Intelligence*, pages 190–204, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Revital Ludewig and Juan Lallave. 2013. Professional stress, discrimination and coping strategies: Similarities and differences between female and male judges in switzerland.

Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and Paolo Italiani. 2023. Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law*, pages 1–29.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? *arXiv preprint*. ArXiv:2211.17135 [cs].

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 [cs].

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. MultiLexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. *arXiv preprint*. ArXiv:2206.10883 [cs].

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 1048–1064.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*. ArXiv: 1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Potential Risks

We believe the release of SLDS poses minimal risk. On the contrary, we expect our dataset to foster further research and encourage the development of assistive technologies that can make the work of lawyers, judges, and clerks more efficient. However, it is crucial not to rely on these summaries blindly. We recommend using such systems as tools to enhance efficiency, rather than as substitutes for human oversight. Users must ensure that the generated summaries accurately reflect the decisions and do not introduce any misleading content, since lawyers will rely on these summaries to find relevant cases faster.

B Use of AI Assistants

We used ChatGPT to improve the content of this article. It was used to rephrase certain passages, as well as condense them to make the text less redundant and easier to understand. We carefully checked that the generated paraphrases corresponded to our own ideas and that no errors were introduced during this process.

C Additional Details on Dataset

C.1 Fields

The dataset includes the following fields:

- `sample_id`: Unique identifier for a sample.
- `decision_id`: Identifier for a specific decision. Since each decision has headnotes in three languages, this ID appears three times in the dataset.
- `decision`: Full text of the landmark decision in either German, French or Italian.
- `decision_language`: ISO language code of the decision (one of `de`, `fr`, `it`).
- `headnote`: Text of the headnote/summary, comprising: i) Key legal citations, including laws and prior cases, ii) Thematic keywords from a legal thesaurus, and iii) A free-form summary of key considerations.
- `headnote_language`: ISO language code of the headnote (one of `de`, `fr`, `it`).
- `law_area`: Legal domain of the decision.
- `year`: Year the decision was issued.
- `volume`: Publication volume of the decision.
- `url`: Link to the official decision on the SFSC website.

C.2 Number of landmark decisions by Year

In Figure 6, we provide a distribution of Landmark Decisions (LDs) over the years.

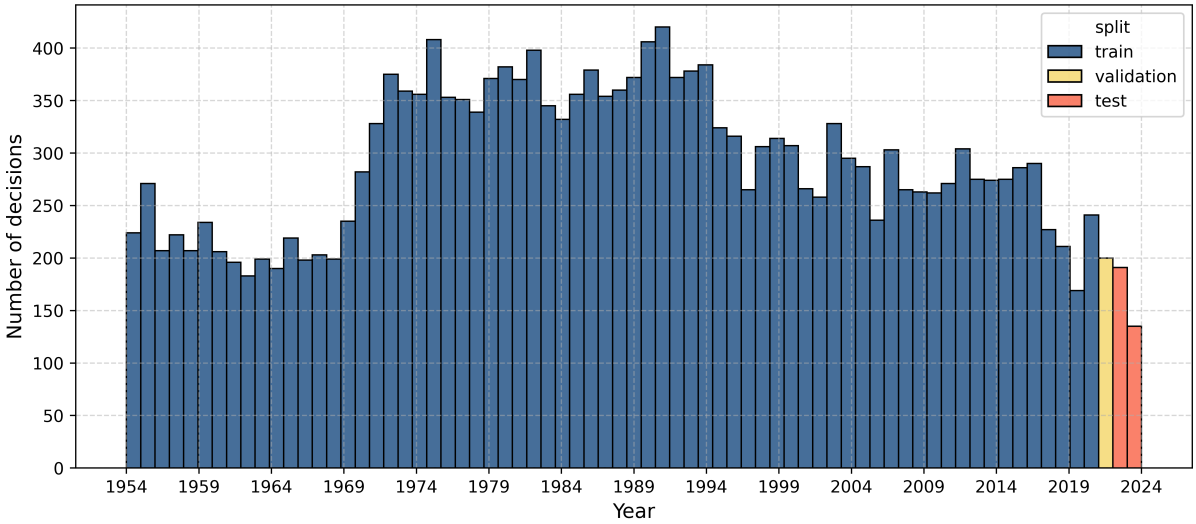


Figure 6: Number of landmark decisions published per year.

C.3 Properties related to Summarization

We provide detailed statistics about summarization-related properties across different dataset splits in Table 3 below.

Table 3: Summarization-related properties of our dataset for each split. CR = Compression Ratio, EFC/EFD = Extractive Fragment Coverage/Density, 1GN-4GN = n-Gram Novelty percentages. CRs are calculated across all samples, the other metrics only across samples where the decision language matches the headnote language to prevent distorted results due to non-matching n-gram pairs in different lanaguages.

Metric	Subset	Mean	Std	Min	Median	Max
CR	Overall	26.39	30.09	1.89	21.42	3710.5
	Train	26.21	30.01	1.89	21.29	3710.5
	Validation	29.86	19.74	4.84	25.29	150.96
	Test	35.47	37.68	3.22	28.02	634.61
EFC	Overall	0.90	0.07	0.24	0.92	1.00
	Train	0.90	0.07	0.24	0.92	1.00
	Validation	0.95	0.04	0.78	0.96	1.00
	Test	0.95	0.04	0.78	0.96	1.00
EFD	Overall	4.63	4.05	0.25	3.51	77.65
	Train	4.59	3.98	0.25	3.48	77.65
	Validation	6.90	6.31	1.76	4.80	45.56
	Test	6.02	5.49	1.58	4.54	66.40
1GN	Overall	10.15	7.85	0.00	8.55	90.38
	Train	10.26	7.89	0.00	8.70	90.38
	Validation	5.52	4.30	0.00	4.40	24.29
	Test	5.73	4.80	0.00	4.58	26.79
2GN	Overall	45.63	16.39	0.00	45.28	100.0
	Train	45.86	16.39	0.00	45.53	100.0
	Validation	36.25	13.70	7.31	37.50	76.92
	Test	37.15	13.82	9.57	36.55	76.36
3GN	Overall	64.62	17.50	0.00	66.15	100.0
	Train	64.84	17.47	0.00	66.67	100.0
	Validation	55.38	16.87	15.06	58.49	100.0
	Test	56.95	16.25	17.65	58.14	96.30
4GN	Overall	75.46	16.86	0.00	78.43	100.0
	Train	75.65	16.82	0.00	78.65	100.0
	Validation	66.70	17.31	20.16	70.67	100.0
	Test	68.87	16.30	22.32	70.36	100.0

D Resources Used

E Fine-Tuning Hyperparameters

We fine-tuned our models using the Unsloth library (Daniel Han and team, 2023). We followed a Parameter Efficient Fine-Tuning (PEFT) training scheme by only fine-tuning a small set of additional weights using LoRA (Hu et al., 2021). We used 16 for both the LoRA rank and the alpha. LoRA was applied to the following target modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. Whenever possible, we used a batch size of 32. Where this was not possible, we used gradient accumulation steps to still train with an effective batch size of 32. For each model, we performed a learning rate sweep across three different learning rates (1e-5, 5e-5, 1e-4) for 500 steps. The 1e-4 learning rate performed best across all models, so we used it for fine-tuning all of our models with 200 warmup steps and a linear learning rate scheduler. We used an 8-bit version of AdamW (Loshchilov, 2017) as the optimizer and trained the models for 3 epochs. Due to memory limitations, the maximum sequence length of the models was set to 8192, which is long enough to cover roughly 95% of all decisions in the training set when estimated using the `tiktoken` tokenizer. The rest of the decisions was truncated during training. The exact training configuratio along with the training and evaluation scripts can be found on our GitHub repository.

E.1 Evaluation

We used a single H100 GPU with 96 GB VRAM both during fine-tuning and evaluation of local LLMs. We estimate a total of roughly 250 GPU hours for all of our learning rate sweeps and fine-tuning runs.

F Prompts

All of the models that we used during our experiments use chat templates. Below, we report the different system and user messages that were used in our experiments.

F.1 Fine-Tuning

During fine-tuning, we did not specify the system message, which means that the individual default system message for each model was used. The user message that we used to teach the model to map decisions to headnotes was a simple prefix that can be seen below in Prompt 1.

```
Generate a headnote in {language} for the following leading decision: {decision}
```

Prompt 1: The user prompt that was used during fine-tuning. The blue text wrapped with curly brackets represent variables. The decision text was inserted directly from dataset column. For the language, we converted the language ISO code into the corresponding written out language first, i.e. either *German*, *French*, or *Italian.*

F.2 Headnote generation

During the evaluation, we used the default system prompt of the model and Prompt as the user message 2 to generate the headnotes. Unlike during fine-tuning, we decided to use a suffix rather than a prefix for the instruction to benefit from prompt caching. In the case of the pre-trained models (OpenAI and Anthropic models as well as DeepSeek R1), we used one-shot prompting as implemented in lighteval: an additional initial turn of conversation is added where the assistant response is already provided with the gold headnote as content.

````
Leading decision:
```{decision}```

Generate a headnote in {language} for the leading decision above.
````

Prompt 2: The user prompt that was used during the generation of the headnotes. The blue text wrapped with curly brackets represent variables. The decision text was inserted directly from dataset column. For the language, we converted the language ISO code into the corresponding written out language first, i.e. either *German*, *French*, or *Italian.*

F.3 Evaluation

For the LLM-as-a-Judge evaluation, we used Prompt 3 as the system message and Prompt 4 as the user message. In the user prompt, we provided a one-shot example in German, French or Italian, depending on the language of the generated headnote that was evaluated. For these examples, we use the gold headnotes from the validation set that had the least number of tokens in the respective language. The model generated output in these examples stems from DeepSeek V3 and the scores in these demonstrations were assigned manually. The content of these one-shot examples is presented in Examples 1 to 3.

G Example Outputs

G.1 Judge
````

```
You are a senior legal expert and quality assurance specialist with over 20 years of experience in Swiss law. You possess
↪  native-level proficiency in German, French, and Italian, enabling you to evaluate Swiss Federal Supreme Court headnotes
↪  with precision. Your task is to compare the **Official (Gold) Headnote** with a **Model-Generated Headnote** and provide
↪  a structured evaluation in five categories. You will carefully analyze each category and provide a short analysis before
↪  committing to a score. The categories are:

1. Accuracy & Faithfulness: How well does the Model-Generated Headnote match the essential legal meaning and intent of the
↪  Official Headnote?
2. Completeness & Relevance: Does the Model-Generated Headnote include all important points that the Official Headnote
↪  emphasizes, without adding irrelevant details?
3. Clarity & Coherence: Is the text well-organized, easy to understand, and coherent in style and structure?
4. Articles: Do the same legal articles (prefixed "Art.") appear correctly and completely in the Model-Generated Headnote as
↪  in the Official Headnote?
5. Considerations: Do the same considerations (prefixed "E." in German or "consid." in French/Italian) appear correctly and
↪  completely in the Model-Generated Headnote as in the Official Headnote?

For each category, provide a short and concise explanation followed by a score on a scale from 1 to 3:

1: Fails or is substantially flawed.
Major omissions or inaccuracies that fundamentally alter the legal meaning.

2: Largely correct but missing key element(s).
Generally captures the substance, yet lacks one or more important details or references.

3: Closely matches the Official Headnote.
Covers all critical aspects and references with only minor wording variations that do not affect the legal content.

Your output must follow the exact structure provided below to ensure consistency and ease of parsing.
```

Prompt 3: The system prompt that was used for the DeepSeek V3 judge in the LLM-as-a-Judge evaluation. It describes the five categories that the judge should use to compare the generated headnotes with the original (gold) headnotes as well as the grading system.

```
Below are two headnotes for the same leading decision from the Swiss Federal Supreme Court. Please compare the
↪  Model-Generated Headnote to the Official (Gold) Headnote according to the following five categories: Accuracy &
↪  Faithfulness, Completeness & Relevance, Clarity & Coherence, Articles, and Considerations.

1. Analyze the Model-Generated Headnote in comparison to the Official Headnote for each category.
2. Provide a short explanation for your evaluation in each category.
3. Conclude each category with a score in the exact format: CATEGORYNAME_SCORE: [X], where X is an integer from 1 to 3.

Required Output Format:

ACCURACY_FAITHFULNESS:
Analysis: [Your concise analysis here]
ACCURACY_FAITHFULNESS_SCORE: [X]

COMPLETENESS_RELEVANCE:
Analysis: [Your concise analysis here]
COMPLETENESS_RELEVANCE_SCORE: [X]

CLARITY_COHERENCE:
Analysis: [Your concise analysis here]
CLARITY_COHERENCE_SCORE: [X]

ARTICLES:
Analysis: [Your concise analysis here]
ARTICLES_SCORE: [X]

CONSIDERATIONS:
Analysis: [Your concise analysis here]
CONSIDERATIONS_SCORE: [X]

Please adhere strictly to this structure.

Here is an example how your response could look like:

<EXAMPLE>
{one_shot_example}
</EXAMPLE>

Now evaluate the Model-Generated Headnote against the Official (Gold) Headnote on the headnotes below.

Official (Gold) Headnote:
{original_headnote}

Model-Generated Headnote:
{generated_headnote}
```

Prompt 4: The user prompt that was used for the DeepSeek V3 judge in the LLM-as-a-Judge evaluation. The blue text wrapped with curly brackets represent variables. The `one_shot_example` used either the German, French or Italian demonstration illustrated in Examples 1 to 3.

```
Official (Gold) Headnote:
Art. 391 Abs. 2 StPO; die erstmalige Anordnung einer ambulanten Massnahme durch das Berufungsgericht verstösst gegen das
↪  Verschlechterungsverbot (reformatio in peius). Verzichtet das erstinstanzliche Gericht auf die Anordnung einer beantragten
↪  ambulanten Massnahme und hat die Staatsanwaltschaft in ihrer Anschlussberufung deren Anordnung nicht erneut beantragt,
↪  verletzt das Berufungsgericht das Verschlechterungsverbot, wenn es eine ambulante Massnahme anordnet (E. 4.1-4.4).

Model-Generated Headnote:
Art. 391 Abs. 2 StPO; Verschlechterungsverbot; ambulante Massnahme nach Art. 63 StGB. Die erstmalige Anordnung einer
↪  ambulanten Massnahme im Berufungsverfahren verletzt das Verschlechterungsverbot nach Art. 391 Abs. 2 StPO, wenn
↪  erstinstanzlich auf eine solche Massnahme verzichtet wurde und die Staatsanwaltschaft in ihrer Anschlussberufung keinen
↪  entsprechenden Antrag gestellt hat. Eine solche Anordnung stellt eine unzulässige reformatio in peius dar, da der
↪  Beschuldigte das Risiko einer nachträglichen Anpassung oder Umwandlung der Massnahme nur dann trägt, wenn bereits
↪  erstinstanzlich eine therapeutische Massnahme angeordnet wurde (E. 4.4). Die freiwillige Teilnahme an einer Therapie
↪  durch den Beschuldigten begründet keine ausreichende Grundlage für die erstmalige Anordnung einer ambulanten Massnahme im
↪  Berufungsverfahren (E. 4.3).

ACCURACY_FAITHFULNESS:
Analysis: The model-generated headnotes captures the legal essence and intent of the official headnote accurately. It
↪  correctly explains the legal principle of reformatio in peius and the violation of the prohibition of worsening in the
↪  context of ordering outpatient measures.
ACCURACY_FAITHFULNESS_SCORE: 3

COMPLETENESS_RELEVANCE:
Analysis: The model-generated headnote inclundes all relevant aspects of the official headnote, such as the prohibition of
↪  worsening, the legal context of ordering outpatient measures, and the implications of the prosecution's appeal. However,
↪  it also adds additional details regarding the voluntary participation in therapy, which are not explicitly mentioned in
↪  the official headnote. These are very specific details that may not be necessary for a concise headnote.
COMPLETENESS_RELEVANCE_SCORE: 2

CLARITY_COHERENCE:
Analysis: The model-generated headnote is easy to understand and coherent. The clarity could be improved by not repeating the
↪  fact that the outpatient measure violates the prohibition of worsening twice. Other than that, the structure and style
↪  are clear and well-organized.
CLARITY_COHERENCE_SCORE: 2

ARTICLES:
Analysis: While both the official and model-generated headnotes correctly reference Art. 391 Abs. 2 StPO, the model-generated
↪  headnote also includes Art. 63 StGB, which is not explicitly mentioned in the official headnote. This additional
↪  reference is relevant to the context of outpatient measures but may not be necessary for a concise headnote.
ARTICLES_SCORE: 2

CONSIDERATIONS:
Analysis: The model-generated headnote only cites considerations E. 4.3 and E. 4.4, whereas the official headnote includes
↪  considerations E. 4.1 to E. 4.4. This omission reduces the completeness of the model-generated headnote.
CONSIDERATIONS_SCORE: 2
```

Example 1: The German one-shot example that was used in Prompt 4 in the LLM-as-a-Judge evaluation in case the generated and gold headnotes to evaluate were in German. For token efficiency, we used the sample with `sample_id` 60312 as it was the shortest possible example from the validation set in German.

```
Official (Gold) Headnote:
Travail externe (art. 77a CP). Le condamné à une peine privative de liberté qui a passé une longue période en détention
↪  provisoire peut exécuter sa peine (restante) directement sous la forme de travail externe s'il réalise les conditions
↪  posées à l'art. 77a al. 1 CP. Il n'a pas nécessairement besoin d'être privé de sa liberté au moment où est prise la
↪  décision ordonnant le travail externe (consid. 2.5.2).

Model-Generated Headnote:
Art. 77a CP; art. 164 ss RSPC; conditions pour l'octroi du travail externe. Le travail externe, en tant que phase de
↪  l'élargissement progressif de l'exécution de la peine, ne nécessite pas impérativement un séjour préalable en milieu
↪  ouvert ou dans une section ouverte d'un établissement fermé. L'utilisation du terme "en principe" à l'art. 77a al. 2 CP
↪  indique que cette condition n'est pas absolue et peut être adaptée aux circonstances individuelles, notamment lorsque le
↪  condamné a déjà purgé une partie significative de sa peine en détention provisoire ou pour des motifs de sûreté. La cour
↪  cantonale a violé le droit fédéral en exigeant de manière rigide un séjour en milieu ouvert comme condition préalable à
↪  l'octroi du travail externe. Le recours est admis, l'arrêt cantonal est annulé, et la cause est renvoyée pour une
↪  nouvelle décision sur la base des conditions prévues à l'art. 77a CP (consid. 2.5.1 à 2.5.3).

ACCURACY_FAITHFULNESS:
Analysis: The model-generated headnote accurately reflects the legal principle and conditions for granting external work
↪  under Art. 77a CP.
ACCURACY_FAITHFULNESS_SCORE: 3

COMPLETENESS_RELEVANCE:
Analysis: The model-generated headnote includes all relevant aspects of the official headnote. However, it adds additional
↪  details regarding the use of the term "en principe" and the violation of federal law by the cantonal court. While these
↪  details provide context, they are not essential for a concise headnote that shapes future legislation.
COMPLETENESS_RELEVANCE_SCORE: 2

CLARITY_COHERENCE:
Analysis: The model-generated headnote is clear and well-organized, but the inclusion of specific details may obscure the
↪  broader legal principle.
CLARITY_COHERENCE_SCORE: 2

ARTICLES:
Analysis: The model-generated headnote includes extra legal articles (Art. 164 ff. RSPC) not cited in the official headnote,
↪  deviating from its intended focus. Besides this, the reference to Art. 77a CP aligns with the official headnote.
ARTICLES_SCORE: 2

CONSIDERATIONS:
Analysis: The official headnote only cites consideration E 2.5.2, while the model-generated headnote references
↪  considerations E 2.5.1 to E 2.5.3. This makes the model-generated headnote less precise.
CONSIDERATIONS_SCORE: 2
```

Example 2: The French one-shot example that was used in Prompt 4 in the LLM-as-a-Judge evaluation in case the generated and gold headnotes to evaluate were in French. For token efficiency, we used the sample with `sample_id` 60247 as it was the shortest possible example from the validation set in German.

```
Official (Gold) Headnote:
Art. 24 Cost.; libertà di domicilio; residenza secondaria per motivi professionali; autorizzazione di soggiorno.  Contenuto e
↪  portata della libertà di domicilio garantita dall'art. 24 Cost. (consid. 3.2.1). In tale ambito le competenze dei Cantoni
↪  e dei Comuni sono circoscritte al controllo e alla registrazione degli abitanti (consid. 3.3.1). L'obbligo di annunciarsi,
↪  in caso di soggiorno in un Comune differente da quello del domicilio per motivi professionali, non limita l'esercizio
↪  della libertà di domicilio. Detto annuncio non può però essere sottoposto ad un regime autorizzativo (consid. 3.3.3).

Model-Generated Headnote:
Art. 24 Cost.; libertà di domicilio; regime autorizzativo per il soggiorno; competenze comunali. La libertà di domicilio
↪  garantita dall'art. 24 Cost. consente a ogni cittadino svizzero di eleggere liberamente il proprio luogo di residenza o
↪  soggiorno sul territorio nazionale, senza che ciò possa essere sottoposto a un regime autorizzativo di polizia. Le
↪  autorità comunali possono richiedere unicamente una notifica al momento dell'arrivo nel Comune, finalizzata alla
↪  registrazione e al controllo degli abitanti, ma non possono imporre un'autorizzazione di soggiorno con scadenza annuale,
↪  poiché ciò costituirebbe una violazione della libertà di domicilio (consid. 3.3.3). Il sistema instaurato dalle Direttive
↪  MOVPOP, che prevede il rilascio di un'"autorizzazione di soggiorno" con validità limitata, deve essere interpretato nel
↪  senso che l'autorità comunale può solo certificare formalmente la notifica del soggiorno, senza sottoporre quest'ultimo a
↪  un regime autorizzativo (consid. 3.3.2 e 3.3.3). La conferma di un tale regime da parte del Tribunale cantonale
↪  amministrativo viola pertanto l'art. 24 Cost. e deve essere annullata (consid. 3.4).

ACCURACY_FAITHFULNESS:
Analysis: The model-generated headnote aligns with the core legal meaning but includes additional details (e.g., MOVPOP
↪  directives) not in the official headnote. These do not conflict but shift the focus slightly.
ACCURACY_FAITHFULNESS_SCORE: 2

COMPLETENESS_RELEVANCE:
Analysis: The model-generated headnote captures key points but omits emphasis on secondary residence for professional reasons
↪  and cantonal/communal roles. Irrelevant details (e.g., MOVPOP) add complexity.
COMPLETENESS_RELEVANCE_SCORE: 2

CLARITY_COHERENCE:
Analysis: The model-generated headnote is clear and organized, but additional elements like MOVPOP reduce coherence by
↪  shifting focus away from the main points and making the text longer and more complex.
CLARITY_COHERENCE_SCORE: 2

ARTICLES:
Analysis: References to Art. 24 Cost. are correct and complete.
ARTICLES_SCORE: 3

CONSIDERATIONS:
Analysis: The model-generated headnote correctly references consid. 3.3.3 but adds consid. 3.3.2 and 3.4, which are beyond
↪  the official headnote's scope. Moreover, it leaves out consid 3.2.1 and 3.3.1, reducing precision. Instead, it mentiones
↪  consid. 3.3.3 twice, which is redundant.
CONSIDERATIONS_SCORE: 1
```

Example 3: The Italian one-shot example that was used in Prompt 4 in the LLM-as-a-Judge evaluation in case the generated and gold headnotes to evaluate were in Italian. For token efficiency, we used the sample with `sample_id` 59894 as it was the shortest possible example from the validation set in German.

LLM-as-a-Judge Output

Generated Headnote (Model: Claude 3.5 Sonnet)

Original: **Art. 9, 13, 15, 20 und 10bis GebV SchKG; Gebühren und Entschädigungen im Betreibungsverfahren.** 1. Für die Zustellung eines Zahlungsbefehls können neben der Gebühr nach Art. 16 Abs. 1 GebV SchKG Auslagen für Posttaxen (Art. 13 Abs. 1 GebV SchKG) verrechnet werden. Ein erfolgloser Zustellversuch löst jedoch erst ab dem zweiten Versuch eine zusätzliche Gebühr nach Art. 16 Abs. 3 GebV SchKG aus (E. 3.2.1–3.2.2). 2. **Abholungseinladungen** für Zahlungsbefehle stellen keine gesetzlich vorgeschriebenen Amtshandlungen dar. Für sie dürfen weder Gebühren nach Art. 9 GebV SchKG noch Auslagen nach Art. 13 GebV SchKG erhoben werden, selbst wenn Art. 10bis GebV SchKG dies seit 2022 vorsieht, da die Verordnung hierfür keine hinreichende gesetzliche Grundlage bietet (E. 3.2.3). 3. **Pfändungsankündigungen** sind vom Pfändungsvollzug zu trennen und können separat mit einer Gebühr nach Art. 9 Abs. 1 lit. a GebV SchKG sowie Auslagen für eingeschriebene Zustellung belastet werden. Ein zusätzlicher Versand per A-Post ist jedoch mangels gesetzlicher Grundlage nicht kostenpflichtig (E. 3.3.1–3.3.2). 4. **Verlustscheine** (Art. 115 SchKG) dürfen nur für die Abschrift der Pfändungsurkunde (Art. 24 GebV SchKG) und deren eingeschriebene Zustellung belastet werden. Die Zustellung per A-Post an den Schuldner ist unzulässig und damit nicht erstattungsfähig (E. 3.4). 5. **Wegentschädigungen** nach Art. 14 GebV SchKG setzen voraus, dass das Betreibungsamt mehrere Verrichtungen gemäss Art. 15 GebV SchKG berücksichtigt. Die Nichtbeachtung entsprechender Parteivorbringen verletzt das rechtliche Gehör (Art. 29 Abs. 2 BV) und führt zur Zurückweisung des Entscheids (E. 3.5.1). Die Gebührenverordnung ist restriktiv auszulegen: Kosten dürfen nur für gesetzlich vorgeschriebene Amtshandlungen erhoben werden, wobei die Praxis der Betreibungsämter keine eigenständige Rechtsgrundlage schafft (E. 3.2.3, 3.3.2).

Translated: **Art. 9, 13, 15, 20, and 10bis GebV SchKG; Fees and compensations in debt enforcement proceedings.** In addition to the fee under Art. 16 para. 1 GebV SchKG, expenses for postal charges (Art. 13 para. 1 GebV SchKG) may be charged for the delivery of a payment order. However, an unsuccessful delivery attempt only triggers an additional fee under Art. 16 para. 3 GebV SchKG starting from the second attempt (consid. 3.2.1–3.2.2). Collection notices for payment orders do not constitute legally prescribed official acts. Therefore, no fees under Art. 9 GebV SchKG or expenses under Art. 13 GebV SchKG may be charged for them, even though Art. 10bis GebV SchKG has provided for this since 2022, as the ordinance lacks a sufficient legal basis for such charges (consid. 3.2.3). Seizure announcements must be distinguished from the execution of the seizure itself and may be charged separately with a fee under Art. 9 para. 1 lit. a GebV SchKG, along with expenses for registered delivery. However, an additional dispatch by A-Post is not chargeable due to the lack of a legal basis (consid. 3.3.1–3.3.2). Loss certificates (Art. 115 SchKG) may only be charged for the copy of the seizure record (Art. 24 GebV SchKG) and its registered delivery. Delivery by A-Post to the debtor is not permissible and therefore not reimbursable (consid. 3.4). Travel compensations under Art. 14 GebV SchKG require that the debt enforcement office considers multiple tasks in accordance with Art. 15 GebV SchKG. Failure to consider relevant submissions by the parties violates the right to be heard (Art. 29 para. 2 BV) and results in the annulment of the decision (consid. 3.5.1). The fee ordinance must be interpreted restrictively: Costs may only be charged for legally prescribed official acts, and the practices of the debt enforcement offices do not constitute an independent legal basis (consid. 3.2.3, 3.3.2).

Original Headnote (Sample ID: 61194)

Original: Art. 1, Art. 2, Art. 9 Abs. 1 lit. a, Art. 10bis, Art. 13 Abs. 1, Art. 14, Art. 15 Abs. 1, Art. 16 Abs. 1 und Abs. 3, Art. 20, Art. 24 GebV SchKG; Art. 16, Art. 34, Art. 72 Abs. 1, Art. 90, Art. 112, Art. 114, Art. 115 Abs. 1 SchKG; Kosten von Zahlungsbefehlen, Pfändungsankündigungen und Verlustscheinen. Allgemeines zu Gebühren und Entschädigungen gemäss GebV SchKG (E. 3.1). Kosten für die Zustellung von Zahlungsbefehlen (E. 3.2.1); Gebühr bei einem erfolglosen Zustellversuch (E. 3.2.2) und für eine Abholungseinladung. Art. 10bis GebV SchKG stellt keine genügende gesetzliche Grundlage dar, um für die Einladung zur Abholung eines Zahlungsbefehls Kosten in Rechnung zu stellen (E. 3.2.3). Die Kosten für eine Pfändungsankündigung sind nicht in Art. 20 GebV SchKG geregelt (E. 3.3.1). Die Pfändungsankündigung ist nach Art. 34 SchKG zuzustellen. Die Zustellung mit A-Post ist nicht vorgesehen und kann nicht in Rechnung gestellt werden (E. 3.3.2). Pfändungsurkunde als Verlustschein (Art. 115 Abs. 1 SchKG). Art. 20 Abs. 1 GebV SchKG bezieht sich nur auf die Abfassung der Pfändungsurkunde für das Amt (Art. 112 SchKG) und nicht auf die Abschriften für den Schuldner und die Gläubiger (Art. 114 SchKG). Gebühren für diese Abschriften (Art. 24 GebV SchKG). Die Abschriften sind nach Art. 34 SchKG zuzustellen. Die Zustellung mit A-Post ist nicht vorgesehen und kann nicht in Rechnung gestellt werden (E. 3.4). Wegentschädigungen (Art. 14 und 15 GebV SchKG). Verletzung des rechtlichen Gehörs; Sachverhaltsfeststellung von Amtes wegen (Art. 20a Abs. 2 Ziff. 2 SchKG) und Pflicht der Aufsichtsbehörden, die Anwendung der GebV SchKG zu überwachen (Art. 2 GebV SchKG) (E. 3.5).

Translated: Art. 1, Art. 2, Art. 9 para. 1 let. a, Art. 10bis, Art. 13 para. 1, Art. 14, Art. 15 para. 1, Art. 16 para. 1 and para. 3, Art. 20, Art. 24 GebV SchKG; Art. 16, Art. 34, Art. 72 para. 1, Art. 90, Art. 112, Art. 114, Art. 115 para. 1 SchKG; Costs of payment orders, seizure notices, and loss certificates. General principles on fees and compensations according to the GebV SchKG (consid. 3.1). Costs for the delivery of payment orders (consid. 3.2.1); fee for an unsuccessful delivery attempt (consid. 3.2.2) and for a pickup invitation. Art. 10bis GebV SchKG does not constitute a sufficient legal basis to charge costs for the invitation to pick up a payment order (consid. 3.2.3). The costs for a seizure notice are not regulated in Art. 20 GebV SchKG (consid. 3.3.1). The seizure notice must be delivered according to Art. 34 SchKG. Delivery by A-Post is not provided for and cannot be charged (consid. 3.3.2). Seizure certificate as a loss certificate (Art. 115 para. 1 SchKG). Art. 20 para. 1 GebV SchKG only refers to the drafting of the seizure certificate for the authority (Art. 112 SchKG) and not to the copies for the debtor and creditors (Art. 114 SchKG). Fees for these copies (Art. 24 GebV SchKG). The copies must be delivered according to Art. 34 SchKG. Delivery by A-Post is not provided for and cannot be charged (consid. 3.4). Travel compensations (Art. 14 and 15 GebV SchKG). Violation of the right to be heard; determination of facts ex officio (Art. 20a para. 2 no. 2 SchKG) and the duty of supervisory authorities to monitor the application of the GebV SchKG (Art. 2 GebV SchKG) (consid. 3.5).

Judge Output (Aggregated Score: 40)

ACCURACY_FAITHFULNESS:
Analysis: The Model-Generated Headnote captures the general legal principles and main points of the Official Headnote but lacks some specific details and precise references. For instance, it omits references to Articles 1, 2, 10bis, 14, 16, 72, 90, 112, and 115 SchKG, which are crucial in the Official Headnote. Additionally, the Model-Generated Headnote introduces new elements like the reference to BGE 138 III 25, which is not present in the Official Headnote.
ACCURACY_FAITHFULNESS_SCORE: 2

COMPLETENESS_RELEVANCE:
Analysis: The Model-Generated Headnote includes several key points from the Official Headnote, such as the treatment of Abholungseinladung, Pfändungsankündigung, Verlustschein, and Wegentschädigung. However, it omits significant details, such as the broader context of general fees and compensations (E. 3.1), the specific mention of Articles 10bis, 14, 15, 16, 72, 90, 112, and 115 SchKG, and the role of supervisory authorities (Art. 2 GebV SchKG). These omissions reduce the completeness and relevance of the Model-Generated Headnote.
COMPLETENESS_RELEVANCE_SCORE: 2

CLARITY_COHERENCE:
Analysis: The Model-Generated Headnote is well-organized and coherent, with clear sections addressing different aspects of the case. However, the inclusion of an unnecessary reference to BGE 138 III 25 and the omission of certain key articles and considerations slightly detract from its clarity. The structure is logical, but the lack of completeness affects the overall coherence.
CLARITY_COHERENCE_SCORE: 2

ARTICLES:
Analysis: The Model-Generated Headnote correctly references Articles 9, 13, 15, 20, 24 GebV SchKG and Articles 34, 114 SchKG. However, it omits several critical articles mentioned in the Official Headnote, such as Articles 1, 2, 10bis, 14, 16, 72, 90, 112, and 115 SchKG. This omission significantly impacts the accuracy and completeness of the legal references. ARTICLES_SCORE: 1

CONSIDERATIONS:
Analysis: The Model-Generated Headnote cites considerations E. 3.2.3, E. 3.3.1, E. 3.3.2, E. 3.4, and E. 3.5.1, which are relevant and align with the Official Headnote. However, it misses other critical considerations, such as E. 3.1, E. 3.2.1, E. 3.2.2, and broader references to the role of supervisory authorities. This omission reduces the completeness of the considerations.
CONSIDERATIONS_SCORE: 2

Figure 7: Example of a judgment made by the DeepSeek V3 judge for a headnote generated for sample with id 61194 using Claude 3.5 Sonnet. Given the five scores that the judge produces across the five evaluation category, an aggregated score of 40 is computed.

H Cross-Lingual Evaluations

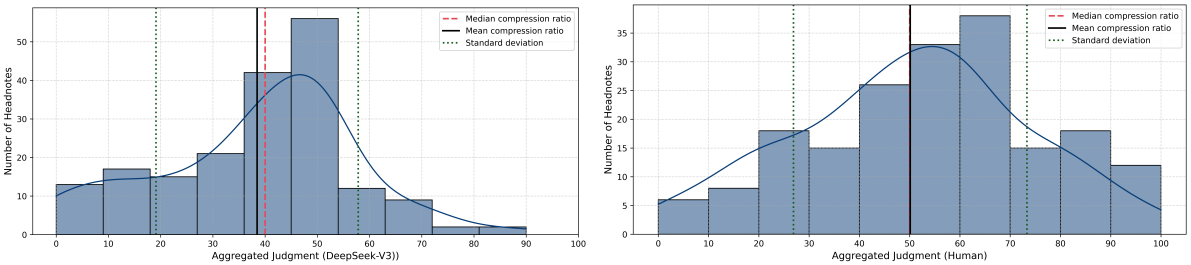
We provide the detailed results for the cross-lingual evaluations in our experiment in Table 4.

Table 4: Results of the baseline experiments on different subsets of the test set of SLDS. Each subset is a combination of the decision language and the headnote language. Standard errors are estimated using the bootstrapping mechanism implemented in lighteval (Fourrier et al., 2023). The Phi-3.5-mini, Llama 3.2 und Qwen 2.5 models were fine-tuned and evaluated in a zero-shot manner, the other models were not fine-tuned and evaluated in a one-shot setting. BERT = BERTScore, MET = METEOR, R1/R2/RL = ROUGE 1/2/L, JUDGE = LLM as Judge. **Bold**: best within subset.

Model	Subset	BERTScore ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	JUDGE ↑
Phi-3.5-mini	de_de	6.74 ± 2.63	40.34 ± 0.54	0.31 ± 0.01	0.16 ± 0.01	0.23 ± 0.01	20.97 ± 1.55
Llama 3.2 3B	de_de	27.27 ± 1.43	47.59 ± 0.45	0.4 ± 0.01	0.21 ± 0.01	0.29 ± 0.01	28.5 ± 1.56
Qwen2.5 0.5B	de_de	16.37 ± 1.65	72.53 ± 0.41	0.32 ± 0.01	0.14 ± 0.01	0.23 ± 0.01	13.0 ± 1.15
Qwen2.5 1.5B	de_de	23.19 ± 1.49	**74.22** ± 0.44	0.36 ± 0.01	0.18 ± 0.01	0.26 ± 0.01	21.88 ± 1.38
Qwen2.5 3B	de_de	28.22 ± 1.4	67.4 ± 0.41	0.39 ± 0.01	0.2 ± 0.01	0.29 ± 0.01	29.42 ± 1.62
Qwen2.5 7B	de_de	32.21 ± 1.24	72.18 ± 0.43	0.42 ± 0.01	0.23 ± 0.01	0.32 ± 0.01	33.09 ± 1.5
Qwen2.5 14B	de_de	**35.22** ± 1.22	66.74 ± 0.43	**0.44** ± 0.01	**0.25** ± 0.01	**0.33** ± 0.01	36.47 ± 1.6
GPT-4o	de_de	27.96 ± 0.9	39.94 ± 0.26	0.41 ± 0.01	0.19 ± 0.01	0.27 ± 0.01	40.58 ± 1.33
Claude Sonnet 3.5	de_de	-25.38 ± 25.95	27.0 ± 0.28	0.4 ± 0.01	0.19 ± 0.01	0.29 ± 0.01	42.27 ± 1.41
DeepSeek-R1	de_de	17.29 ± 0.74	29.69 ± 0.19	0.36 ± 0.01	0.15 ± 0.0	0.21 ± 0.0	**43.91** ± 1.16
o3-mini	de_de	13.78 ± 0.73	31.34 ± 0.2	0.33 ± 0.01	0.12 ± 0.0	0.18 ± 0.0	36.52 ± 1.09
Phi-3.5-mini	de_fr	4.71 ± 2.47	50.73 ± 0.52	0.27 ± 0.01	0.11 ± 0.01	0.19 ± 0.01	13.57 ± 1.33
Llama 3.2 3B	de_fr	24.84 ± 1.62	18.07 ± 0.41	0.35 ± 0.01	0.15 ± 0.01	0.24 ± 0.01	19.08 ± 1.4
Qwen2.5 0.5B	de_fr	-3.81 ± 2.18	20.3 ± 0.5	0.22 ± 0.01	0.07 ± 0.0	0.16 ± 0.01	3.29 ± 0.48
Qwen2.5 1.5B	de_fr	21.71 ± 1.61	25.19 ± 0.38	0.34 ± 0.01	0.13 ± 0.01	0.22 ± 0.01	11.79 ± 1.09
Qwen2.5 3B	de_fr	26.37 ± 1.32	40.22 ± 0.32	0.36 ± 0.01	0.14 ± 0.0	0.24 ± 0.01	18.55 ± 1.29
Qwen2.5 7B	de_fr	32.61 ± 1.06	**52.55** ± 0.32	0.41 ± 0.01	0.18 ± 0.01	0.27 ± 0.01	26.47 ± 1.52
Qwen2.5 14B	de_fr	33.78 ± 1.15	40.47 ± 0.41	0.41 ± 0.01	0.19 ± 0.01	0.28 ± 0.01	30.92 ± 1.55
GPT-4o	de_fr	**33.97** ± 0.76	30.45 ± 0.21	**0.45** ± 0.01	**0.21** ± 0.01	0.28 ± 0.0	40.14 ± 1.42
Claude Sonnet 3.5	de_fr	-5.74 ± 0.94	27.23 ± 0.21	0.43 ± 0.01	0.19 ± 0.01	**0.29** ± 0.01	**44.88** ± 1.48
DeepSeek-R1	de_fr	20.84 ± 0.61	24.25 ± 0.15	0.4 ± 0.01	0.16 ± 0.0	0.21 ± 0.0	42.8 ± 1.24
o3-mini	de_fr	15.68 ± 0.62	20.86 ± 0.15	0.37 ± 0.01	0.13 ± 0.0	0.19 ± 0.0	35.7 ± 1.33
Phi-3.5-mini	de_it	8.06 ± 2.28	30.39 ± 0.47	0.26 ± 0.01	0.1 ± 0.01	0.18 ± 0.01	9.61 ± 1.09
Llama 3.2 3B	de_it	22.81 ± 1.6	14.32 ± 0.41	0.31 ± 0.01	0.13 ± 0.0	0.22 ± 0.01	13.72 ± 1.28
Qwen2.5 0.5B	de_it	4.48 ± 1.89	**48.16** ± 0.38	0.22 ± 0.01	0.08 ± 0.0	0.16 ± 0.01	2.17 ± 0.4
Qwen2.5 1.5B	de_it	22.99 ± 1.3	41.46 ± 0.33	0.31 ± 0.01	0.11 ± 0.0	0.21 ± 0.0	8.16 ± 0.88
Qwen2.5 3B	de_it	23.86 ± 1.5	31.39 ± 0.33	0.32 ± 0.01	0.12 ± 0.0	0.23 ± 0.01	12.46 ± 1.24
Qwen2.5 7B	de_it	30.75 ± 1.0	31.86 ± 0.34	0.36 ± 0.01	0.15 ± 0.01	0.25 ± 0.01	20.39 ± 1.44
Qwen2.5 14B	de_it	**34.46** ± 0.95	45.34 ± 0.35	0.38 ± 0.01	0.16 ± 0.01	0.27 ± 0.01	25.12 ± 1.44
GPT-4o	de_it	32.12 ± 0.69	30.4 ± 0.25	0.39 ± 0.01	0.16 ± 0.0	0.25 ± 0.0	29.66 ± 1.29
Claude Sonnet 3.5	de_it	29.59 ± 0.88	29.52 ± 0.26	**0.43** ± 0.01	**0.2** ± 0.01	**0.3** ± 0.01	**36.52** ± 1.46
DeepSeek-R1	de_it	23.98 ± 0.55	12.77 ± 0.17	0.36 ± 0.01	0.13 ± 0.0	0.2 ± 0.0	30.14 ± 1.26
o3-mini	de_it	15.9 ± 0.52	15.63 ± 0.14	0.31 ± 0.0	0.08 ± 0.0	0.16 ± 0.0	27.83 ± 1.23
Phi-3.5-mini	fr_de	-6.11 ± 3.27	38.47 ± 0.41	0.24 ± 0.01	0.09 ± 0.01	0.17 ± 0.01	8.69 ± 1.56
Llama 3.2 3B	fr_de	1.58 ± 2.44	49.67 ± 0.37	0.26 ± 0.01	0.11 ± 0.01	0.19 ± 0.01	10.65 ± 1.56
Qwen2.5 0.5B	fr_de	-10.66 ± 2.47	33.38 ± 0.39	0.21 ± 0.01	0.07 ± 0.01	0.16 ± 0.01	2.71 ± 0.6
Qwen2.5 1.5B	fr_de	0.62 ± 2.21	27.16 ± 0.35	0.26 ± 0.01	0.09 ± 0.01	0.19 ± 0.01	7.1 ± 1.18
Qwen2.5 3B	fr_de	7.68 ± 2.03	28.04 ± 0.32	0.29 ± 0.01	0.11 ± 0.01	0.2 ± 0.01	13.36 ± 1.48
Qwen2.5 7B	fr_de	15.63 ± 1.8	**50.67** ± 0.31	0.33 ± 0.01	0.12 ± 0.01	0.23 ± 0.01	22.9 ± 2.01
Qwen2.5 14B	fr_de	**21.84** ± 1.51	41.26 ± 0.34	0.36 ± 0.01	0.15 ± 0.01	**0.25** ± 0.01	30.65 ± 1.97
GPT-4o	fr_de	21.02 ± 1.03	31.29 ± 0.21	**0.39** ± 0.01	**0.16** ± 0.01	0.24 ± 0.0	41.12 ± 1.64
Claude Sonnet 3.5	fr_de	-56.43 ± 50.1	0.0 ± 0.26	0.37 ± 0.01	0.15 ± 0.01	**0.25** ± 0.01	42.9 ± 1.93
DeepSeek-R1	fr_de	8.17 ± 1.01	20.77 ± 0.17	0.33 ± 0.01	0.12 ± 0.0	0.19 ± 0.0	**43.64** ± 1.4
o3-mini	fr_de	0.81 ± 0.88	19.15 ± 0.18	0.29 ± 0.01	0.08 ± 0.0	0.16 ± 0.0	28.69 ± 1.72
Phi-3.5-mini	fr_fr	18.62 ± 3.27	49.91 ± 0.54	0.37 ± 0.02	0.18 ± 0.01	0.25 ± 0.01	24.58 ± 2.09
Llama 3.2 3B	fr_fr	24.86 ± 3.03	4.32 ± 0.61	0.39 ± 0.02	0.21 ± 0.01	0.27 ± 0.01	33.36 ± 2.22
Qwen2.5 0.5B	fr_fr	14.65 ± 3.22	**51.91** ± 0.5	0.32 ± 0.02	0.16 ± 0.01	0.22 ± 0.01	14.3 ± 1.81
Qwen2.5 1.5B	fr_fr	33.37 ± 2.17	41.51 ± 0.47	0.43 ± 0.01	0.24 ± 0.01	0.29 ± 0.01	31.5 ± 1.92
Qwen2.5 3B	fr_fr	34.57 ± 2.18	47.78 ± 0.41	0.44 ± 0.01	0.24 ± 0.01	0.3 ± 0.01	35.42 ± 1.93
Qwen2.5 7B	fr_fr	39.91 ± 1.48	51.2 ± 0.42	0.48 ± 0.01	0.27 ± 0.01	0.33 ± 0.01	38.97 ± 1.9
Qwen2.5 14B	fr_fr	**43.31** ± 1.26	42.67 ± 0.44	0.5 ± 0.01	**0.29** ± 0.01	**0.35** ± 0.01	41.96 ± 1.99
GPT-4o	fr_fr	40.2 ± 0.96	44.32 ± 0.28	**0.51** ± 0.01	0.27 ± 0.01	0.31 ± 0.01	48.04 ± 1.48
Claude Sonnet 3.5	fr_fr	-46.15 ± 42.17	17.32 ± 0.24	0.47 ± 0.01	0.22 ± 0.01	0.31 ± 0.01	**50.0** ± 1.99
DeepSeek-R1	fr_fr	28.07 ± 0.85	31.18 ± 0.2	0.43 ± 0.01	0.22 ± 0.01	0.24 ± 0.0	49.25 ± 1.38
o3-mini	fr_fr	25.92 ± 0.86	34.85 ± 0.21	0.44 ± 0.01	0.2 ± 0.01	0.24 ± 0.0	43.93 ± 1.47
Phi-3.5-mini	fr_it	17.03 ± 2.96	25.76 ± 0.47	0.31 ± 0.01	0.13 ± 0.01	0.21 ± 0.01	13.18 ± 1.62
Llama 3.2 3B	fr_it	22.19 ± 2.42	4.98 ± 0.47	0.32 ± 0.01	0.14 ± 0.01	0.23 ± 0.01	17.57 ± 1.82
Qwen2.5 0.5B	fr_it	5.93 ± 2.73	21.94 ± 0.37	0.25 ± 0.01	0.1 ± 0.01	0.18 ± 0.01	3.36 ± 0.7
Qwen2.5 1.5B	fr_it	26.5 ± 1.77	38.52 ± 0.34	0.34 ± 0.01	0.13 ± 0.01	0.23 ± 0.01	12.8 ± 1.34
Qwen2.5 3B	fr_it	28.52 ± 1.93	39.51 ± 0.34	0.35 ± 0.01	0.15 ± 0.01	0.25 ± 0.01	17.76 ± 1.82
Qwen2.5 7B	fr_it	31.5 ± 1.79	**45.05** ± 0.31	0.38 ± 0.01	0.16 ± 0.01	0.26 ± 0.01	24.3 ± 2.04
Qwen2.5 14B	fr_it	35.45 ± 1.53	44.31 ± 0.33	0.4 ± 0.01	0.19 ± 0.01	0.29 ± 0.01	30.65 ± 1.98
GPT-4o	fr_it	**36.37** ± 1.01	31.56 ± 0.25	0.43 ± 0.01	0.19 ± 0.01	0.27 ± 0.01	32.71 ± 1.66
Claude Sonnet 3.5	fr_it	-8.92 ± 38.98	24.62 ± 0.29	**0.45** ± 0.01	**0.22** ± 0.01	**0.3** ± 0.01	37.2 ± 1.8
DeepSeek-R1	fr_it	26.76 ± 0.91	21.21 ± 0.17	0.38 ± 0.01	0.15 ± 0.01	0.21 ± 0.0	**38.22** ± 1.66
o3-mini	fr_it	22.98 ± 0.88	15.31 ± 0.19	0.36 ± 0.01	0.11 ± 0.0	0.19 ± 0.0	29.91 ± 1.6
Phi-3.5-mini	it_de	0.53 ± 6.69	20.35 ± 0.23	0.27 ± 0.04	0.11 ± 0.02	0.17 ± 0.02	5.83 ± 2.6

Model	Subset	BERTScore ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	JUDGE ↑
Llama 3.2 3B	it_de	-3.89 ± 5.97	15.89 ± 0.21	0.24 ± 0.03	0.1 ± 0.02	0.18 ± 0.02	7.5 ± 3.92
Qwen2.5 0.5B	it_de	-23.28 ± 5.94	9.64 ± 0.18	0.16 ± 0.03	0.06 ± 0.01	0.12 ± 0.02	0.0 ± 0.0
Qwen2.5 1.5B	it_de	4.91 ± 2.9	15.66 ± 0.23	0.28 ± 0.02	0.1 ± 0.01	0.19 ± 0.02	4.17 ± 2.29
Qwen2.5 3B	it_de	4.32 ± 5.98	10.03 ± 0.26	0.28 ± 0.03	0.09 ± 0.01	0.19 ± 0.02	10.83 ± 3.36
Qwen2.5 7B	it_de	14.69 ± 3.46	21.69 ± 0.27	0.33 ± 0.03	0.13 ± 0.02	0.21 ± 0.02	23.33 ± 6.2
Qwen2.5 14B	it_de	**17.83** ± 3.4	**28.24** ± 0.36	0.31 ± 0.03	**0.15** ± 0.02	0.22 ± 0.02	27.5 ± 6.17
GPT-4o	it_de	14.71 ± 2.94	21.3 ± 0.2	0.35 ± 0.03	0.14 ± 0.02	0.21 ± 0.02	41.67 ± 5.34
Claude Sonnet 3.5	it_de	-10.25 ± 3.24	22.41 ± 0.2	**0.37** ± 0.03	**0.15** ± 0.02	**0.23** ± 0.02	40.83 ± 5.29
DeepSeek-R1	it_de	5.76 ± 2.42	22.03 ± 0.18	0.35 ± 0.04	0.13 ± 0.01	0.18 ± 0.02	**45.0** ± 3.99
o3-mini	it_de	-6.59 ± 1.74	5.54 ± 0.13	0.26 ± 0.03	0.07 ± 0.01	0.13 ± 0.01	34.17 ± 3.79
Phi-3.5-mini	it_fr	15.3 ± 8.17	30.01 ± 0.32	0.34 ± 0.05	0.16 ± 0.03	0.21 ± 0.03	13.33 ± 3.76
Llama 3.2 3B	it_fr	11.77 ± 9.72	9.48 ± 0.36	0.31 ± 0.05	0.14 ± 0.03	0.2 ± 0.03	17.5 ± 6.64
Qwen2.5 0.5B	it_fr	-23.29 ± 6.14	8.88 ± 0.18	0.17 ± 0.03	0.05 ± 0.01	0.13 ± 0.02	9.17 ± 3.36
Qwen2.5 1.5B	it_fr	20.02 ± 5.31	24.91 ± 0.22	0.32 ± 0.04	0.14 ± 0.02	0.21 ± 0.02	17.5 ± 4.63
Qwen2.5 3B	it_fr	27.6 ± 3.78	**39.09** ± 0.32	0.36 ± 0.04	0.16 ± 0.03	0.23 ± 0.02	25.0 ± 5.71
Qwen2.5 7B	it_fr	31.67 ± 2.34	23.05 ± 0.24	0.4 ± 0.03	0.19 ± 0.02	0.25 ± 0.02	34.17 ± 4.99
Qwen2.5 14B	it_fr	31.69 ± 3.27	35.41 ± 0.28	0.37 ± 0.03	0.17 ± 0.02	0.23 ± 0.01	30.83 ± 7.12
GPT-4o	it_fr	**33.1** ± 3.64	31.58 ± 0.23	**0.46** ± 0.04	**0.21** ± 0.02	**0.27** ± 0.02	43.33 ± 4.66
Claude Sonnet 3.5	it_fr	-8.72 ± 3.58	19.08 ± 0.23	0.42 ± 0.04	0.19 ± 0.03	0.26 ± 0.03	40.0 ± 5.5
DeepSeek-R1	it_fr	23.65 ± 3.24	19.29 ± 0.19	0.43 ± 0.04	0.2 ± 0.02	0.23 ± 0.02	**48.33** ± 4.41
o3-mini	it_fr	17.25 ± 3.07	16.06 ± 0.14	0.4 ± 0.04	0.14 ± 0.02	0.2 ± 0.02	38.33 ± 4.41
Phi-3.5-mini	it_it	36.33 ± 2.62	27.64 ± 0.21	**0.44** ± 0.02	0.24 ± 0.02	0.29 ± 0.02	27.5 ± 5.24
Llama 3.2 3B	it_it	5.4 ± 11.34	32.69 ± 0.52	0.28 ± 0.06	0.15 ± 0.05	0.2 ± 0.04	18.33 ± 6.49
Qwen2.5 0.5B	it_it	7.31 ± 8.42	23.08 ± 0.28	0.28 ± 0.03	0.12 ± 0.02	0.2 ± 0.02	4.17 ± 2.88
Qwen2.5 1.5B	it_it	24.95 ± 5.68	**42.49** ± 0.35	0.34 ± 0.03	0.16 ± 0.02	0.23 ± 0.03	28.33 ± 5.75
Qwen2.5 3B	it_it	27.92 ± 5.05	42.3 ± 0.34	0.36 ± 0.03	0.19 ± 0.03	0.25 ± 0.02	20.0 ± 5.5
Qwen2.5 7B	it_it	37.34 ± 3.52	24.37 ± 0.41	0.42 ± 0.03	0.21 ± 0.03	0.27 ± 0.02	31.67 ± 6.01
Qwen2.5 14B	it_it	**38.77** ± 3.58	31.79 ± 0.36	0.43 ± 0.04	**0.25** ± 0.03	**0.3** ± 0.03	28.33 ± 4.9
GPT-4o	it_it	34.48 ± 3.73	26.14 ± 0.34	0.4 ± 0.04	0.19 ± 0.03	0.24 ± 0.02	**40.0** ± 5.08
Claude Sonnet 3.5	it_it	24.84 ± 4.07	29.71 ± 0.29	0.42 ± 0.03	0.22 ± 0.02	0.27 ± 0.02	36.67 ± 5.27
DeepSeek-R1	it_it	27.97 ± 2.7	20.12 ± 0.19	0.4 ± 0.04	0.17 ± 0.02	0.21 ± 0.02	39.17 ± 3.36
o3-mini	it_it	21.87 ± 2.5	26.18 ± 0.17	0.37 ± 0.03	0.15 ± 0.01	0.19 ± 0.01	38.33 ± 5.05

I Human Evaluation



(a) LLM-as-a-Judge

(b) Human-as-a-Judge

Figure 8: Distributions of (a) the scores generated by DeepSeek-V3 and (b) the scores assigned by two lawyers. The scores are aggregates of the individual scores per evaluation category, ranging from 0 to 100. The scores issued by the lawyers are slightly higher than the ones assigned by DeepSeek-V3.