

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

Anonymous ACL submission

Abstract

Legal research depends on headnotes—concise summaries that help lawyers quickly identify relevant cases. Yet, many court decisions lack them due to the high cost of manual annotation. To address this gap, we introduce the Swiss Landmark Decisions Summarization (SLDS) dataset containing 20K rulings from the Swiss Federal Supreme Court, each with headnotes in German, French, and Italian. SLDS has the potential to significantly improve access to legal information and transform legal research in Switzerland. We fine-tune open models (Qwen2.5, Llama 3.2, Phi-3.5) and compare them to larger general-purpose and reasoning-tuned LLMs, including GPT-4o, Claude 3.5 Sonnet, and the open-source DeepSeek R1. Using an LLM-as-a-Judge framework, we find that fine-tuned models perform well in terms of lexical similarity, while larger models generate more legally accurate and coherent summaries. Interestingly, reasoning-focused models show no consistent benefit, suggesting that factual precision is more important than deep reasoning in this task. We release SLDS under a CC BY 4.0 license to support future research in cross-lingual legal summarization.

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).

To alleviate this increasing need for efficient ways to navigate large amounts of legal texts, legal document summarization has become a critical area of interest in NLP (Jain et al., 2021). Over

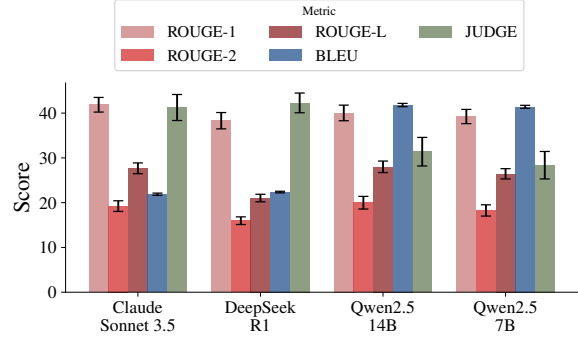


Figure 1: Two fine-tuned LLMs of the Qwen2.5 family and two frontier models evaluated on the SLDS test set. While fine-tuning dominates outcomes in terms of lexical metrics, the smaller fine-tuned models do not yet reach the same output quality as their larger pre-trained counterparts, as indicated by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

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 Giofr , 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 adequately 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 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. SLDS is openly available 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

Research on legal text summarization has increasingly shifted toward abstractive methods, supported by the emergence of dedicated datasets. Among monolingual English corpora, BillSum (Kornilova and Eidelman, 2019) offers 22K U.S. congressional and state bills with summaries, enabling cross-domain transfer for legal summarization. MultiLexSum (Shen et al., 2022) focuses on long civil rights lawsuits and supports multi-length evaluations. Bauer et al. (2023) extracted key passages from 430K U.S. court opinions, favoring reinforcement learning methods, although their dataset is not publicly available. RulingBR (de Vargas Feijó and Moreira, 2018) includes 10K Brazilian Supreme Court rulings with structured summaries. LAW-SUIT (Ragazzi et al., 2024) contains 14K Italian

verdicts with expert-authored maxims from the Constitutional Court. Multilingual datasets include EUR-Lex-Sum (Aumiller et al., 2022), which covers 24 EU languages and aligns 375 legal acts. Unlike court rulings, these acts follow a more structured format. In contrast, our dataset emphasizes case law within a single jurisdiction, offering over 13 times more French-to-German and more than twice as many Italian-to-German examples than EUR-Lex-Sum. MILDSum (Datta et al., 2023) addresses language barriers in India by translating 3K English judgments to Hindi. A key result was that Summarize-then-Translate outperformed direct cross-lingual summarization. Unlike MILDSum, our dataset excludes English and uses headnotes, which are harder to generate than summaries due to their legal specificity and structural requirements, making the task more challenging given the dominance of English in pretraining 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 accompanied 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

We scraped the decisions from the official Swiss Federal Supreme Court repository, covering 70 years and five legal volumes.² We extracted the full decision text, 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

¹Link available upon acceptance

²Available at <https://www.bger.ch/>

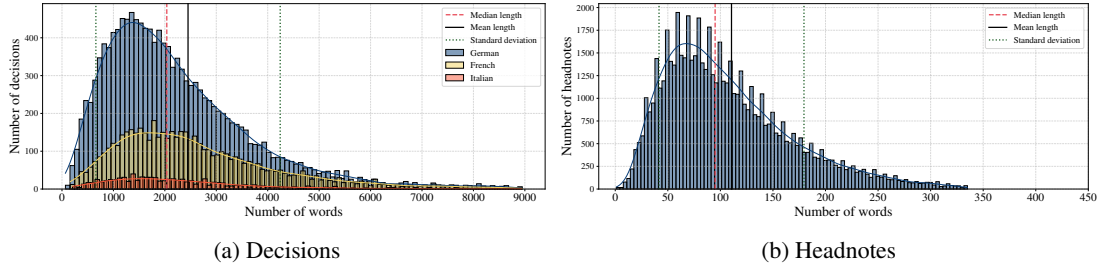


Figure 2: Distributions of token counts in (a) landmark decisions and (b) headnotes. To improve readability, only samples within the 99th percentile were included, as the long tail of the distribution would have otherwise skewed the visualization.

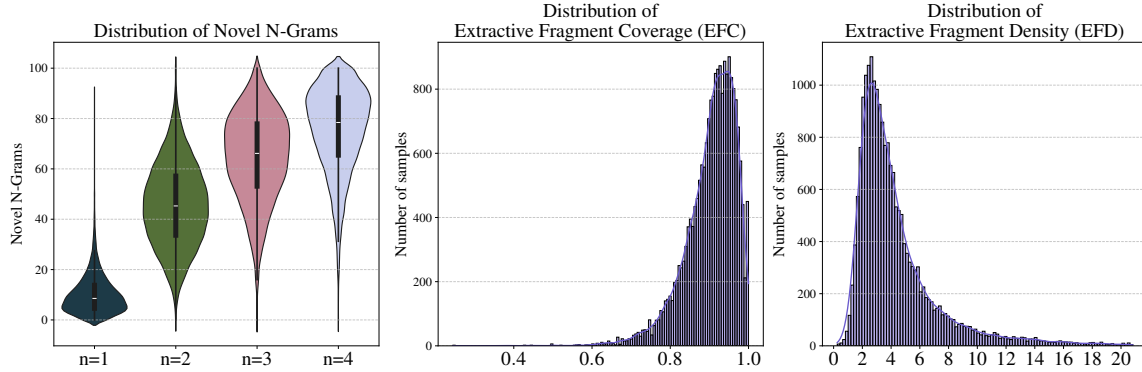


Figure 3: Distribution of Summarization Properties in SLDS. The figure illustrates N-Gram Novelty (left), Extractive Fragment Coverage (EFC) (center), and Extractive Fragment Density (EFD) (right), highlighting the dataset’s balance between abstraction and extractiveness. For the sake of readability, we only show EFD values within the 99th percentile.

over 60K samples. We show the exact fields of our dataset in [Appendix C.1](#).

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](#).

Split	Years	# Dec.	# Samp.	Languages (%)
Train	1954–2021	~20k	~60k	DE: 67.94, FR: 27.36, IT: 4.71
Val	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

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

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., 2018](#)), and N-Gram Novelty ([Narayan et al., 2018](#)). Given the dataset’s multilingual nature, we compare these properties to EUR-Lex-Sum ([Aumiller et al., 2022](#)) and MILD-Sum ([Datta et al., 2023](#)), but only for monolingual samples. We also report Coverage Increment (CI) and Formulaicness ([Ragazzi et al., 2024](#)). [Figure 3](#) visualizes key trends across the entire dataset.

Compression Ratio We compute Compression Ratio (CR) as the ratio of decision to headnote token counts, using language-specific spaCy tokenizers via the `spacy.blank` interface. The observed mean CR of 26.39 is notably higher than in EUR-Lex-Sum and MILDSum, reflecting the extreme conciseness of Swiss judicial headnotes. These headnotes highlight key legal principles that justify a decision’s landmark status. Higher CRs in the validation and test splits suggest a trend toward even shorter headnotes over time.

Extractive Fragments We compute EFC and EFD using spaCy ([Honnibal et al., 2020](#)) with `core_news_sm` models on monolingual samples. While EFC values match MILDSum, this may result from longer input texts and high CRs, which increase unigram overlap. The mean EFD of 4.63, however, is significantly lower than MILDSum’s

LLM-as-a-Judge Output	
Decision	(Sample ID: 61194)
150 III 223 Facts from page 225 A. A. was pursued for a claim of CHF 200 and a reminder fee of CHF 35 (debt collection no. w by the Zug Debt Collection Office) (...) The Federal Supreme Court has already dealt with the costs for a pickup invitation in the past. (...) Therefore, the insufficiently substantiated complaint is not to be admitted (cf. unpublished E. 1).	
Original Headnote	(Sample ID: 61194)
Art. 1, Art. 2, Art. 9 para. 1 let. a, Art. 10bis, (...); 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); (...)	
Generated Headnote	(Model: Claude 3.5 Sonnet)
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. (...)	
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, (...)	
ACCURACY_FAITHFULNESS_SCORE: 2 (Other judge categories skipped for brevity)	

Figure 4: Example of a DeepSeek V3 judgement for a headnote generated by Claude 3.5 Sonnet. The full sample can be seen in Appendix I.

24.42, indicating a more abstractive summarization style. Slightly higher EFC and EFD values in the validation/test sets align with their increased CRs.

N-Gram Novelty We measure abstractivity by the proportion of novel n-grams in headnotes versus source decisions (Narayan et al., 2018). Novelty ranges from 0 (fully extractive) to 100 (fully abstractive). On average, about 90% of headnote unigrams appear in the decision, and only 5% are novel in the test set. Novelty increases with longer n-grams, suggesting that headnotes often reuse the same words but in new combinations. Around 30% of quadgrams are copied verbatim, highlighting the dataset’s mix of extractive and abstractive styles. Compared to the unigram novelty of around 40% and bigram novelty of 64-67% reported for the German, French, and Italian subsets of EUR-Lex-Sum, the headnotes in SLDS adhere more closely to the original wording of the source decisions. More detailed statistics are in Appendix C.3.

Coverage Increment and Formulaicness Following Ragazzi et al. (2024), we compute CI and Formulaicness on monolingual samples. To obtain CI, we divide each decision into ten equal-length segments and compute the proportion of headnote unigrams that also appear in each segment. Figure 5 shows that SLDS exhibits CI values similar to BillSum, especially in German and Italian SLDS samples, while EUR-Lex-Sum displays slightly lower values, in line with its higher abstractivity.

Formulaicness is computed by averaging ROUGE-L F1 scores between headnotes across random subsets. Figure 6 shows that SLDS-DE and SLDS-IT have the lowest scores, indicating greater variability in phrasing. The French subset is similar to BillSum in this regard, while EUR-Lex-Sum exhibits the highest Formulaicness despite its higher N-Gram Novelty. This finding supports the hypothesis that SLDS headnotes, although largely composed of words found in the original decisions, frequently rearrange these words in novel ways. As a result, they strike a distinctive balance between extractiveness and abstractiveness.

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 before publishing them online.⁴

4 Experimental Setup

We evaluate four frontier Large Language Models (LLMs) (GPT-4o, Claude 3.5 Sonnet, DeepSeek R1, o3-mini) in a one-shot setting and fine-tune three Small Language Models (SLMs) (Llama 3.2 3B, Qwen2.5 3B, and Phi-3.5-mini) on the SLDS training split. To assess the effect of model size, we fine-tune additional Qwen2.5 variants (0.5B - 14B) and evaluate them in a zero-shot setting. Appendix F details the model versions, decoding parameters, and one-shot prompting. Fine-tuning hyperparameters are listed in Appendix G.

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

⁴Anonymization guidelines at <https://www.bger.ch/home/juridiction/anonymisierungsregeln.html>

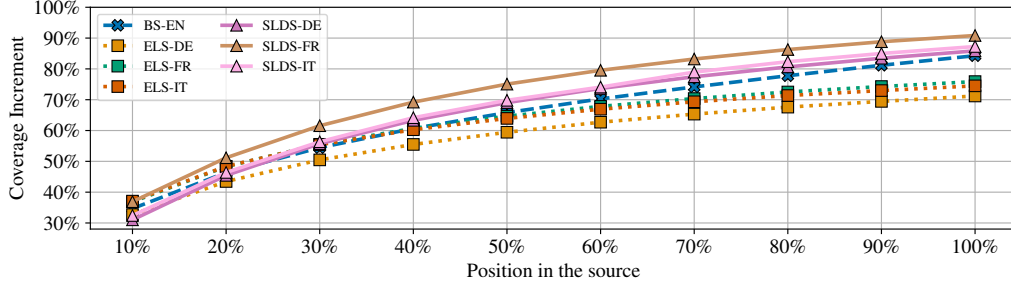


Figure 5: Percentage of unique unigrams in the headnote that also appear in the decision text, reported for the monolingual German, French, and Italian subsets in SLDS and EUR-Lex-Sum (ELS), and for the entire English dataset in BillSum (BS-EN).

Model	Setting	BERTScore $\uparrow$	BLEU $\uparrow$	ROUGE-1 $\uparrow$	ROUGE-2 $\uparrow$	ROUGE-L $\uparrow$	JUDGE $\uparrow$
Phi-3.5-mini	fine-tuned	11.24 $\pm$ 3.82	34.84 $\pm$ 0.41	31.20 $\pm$ 2.08	14.11 $\pm$ 1.27	20.96 $\pm$ 1.35	15.25 $\pm$ 2.32
Llama 3.2 3B	fine-tuned	15.20 $\pm$ 4.40	21.89 $\pm$ 0.42	31.89 $\pm$ 2.34	14.87 $\pm$ 1.61	22.49 $\pm$ 1.60	18.47 $\pm$ 2.99
Qwen2.5 0.5B	fine-tuned	-1.37 $\pm$ 3.85	32.20 $\pm$ 0.35	23.87 $\pm$ 1.68	9.46 $\pm$ 0.94	17.37 $\pm$ 1.09	5.80 $\pm$ 1.26
Qwen2.5 1.5B	fine-tuned	19.81 $\pm$ 2.72	36.79 $\pm$ 0.34	33.03 $\pm$ 1.73	14.14 $\pm$ 1.08	22.67 $\pm$ 1.13	15.92 $\pm$ 2.27
Qwen2.5 3B	fine-tuned	23.23 $\pm$ 2.80	38.42 $\pm$ 0.34	35.18 $\pm$ 1.79	15.66 $\pm$ 1.23	24.10 $\pm$ 1.17	20.31 $\pm$ 2.66
Qwen2.5 7B	fine-tuned	29.59 $\pm$ 1.97	41.40 $\pm$ 0.34	39.24 $\pm$ 1.59	18.26 $\pm$ 1.25	26.44 $\pm$ 1.15	28.37 $\pm$ 3.07
Qwen2.5 14B	fine-tuned	32.48 $\pm$ 1.98	41.80 $\pm$ 0.37	40.04 $\pm$ 1.74	19.99 $\pm$ 1.41	28.00 $\pm$ 1.28	31.38 $\pm$ 3.19
GPT-4o	one-shot	30.44 $\pm$ 1.74	31.89 $\pm$ 0.25	42.12 $\pm$ 1.79	18.92 $\pm$ 1.22	25.92 $\pm$ 1.05	39.70 $\pm$ 2.66
Claude 3.5 Sonnet	one-shot	-11.91 $\pm$ 18.88	21.88 $\pm$ 0.25	41.86 $\pm$ 1.64	19.23 $\pm$ 1.19	27.67 $\pm$ 1.20	41.25 $\pm$ 2.90
DeepSeek-R1	one-shot	20.28 $\pm$ 1.45	22.37 $\pm$ 0.18	38.30 $\pm$ 1.82	15.97 $\pm$ 0.85	21.03 $\pm$ 0.84	42.28 $\pm$ 2.21
o3-mini	one-shot	14.18 $\pm$ 1.31	20.55 $\pm$ 0.17	34.77 $\pm$ 1.43	11.92 $\pm$ 0.69	18.21 $\pm$ 0.67	34.82 $\pm$ 2.41

Table 2: Baselines on the SLDS test set, macro-averaged over the nine decision and headnote language combinations. Standard errors are estimated with bootstrapping as implemented in lighteval (Fourrier et al., 2023). For BERTScore we report the F1 score. The ROUGE scores are multiplied by 100 for consistency. **Bold**: best overall; underlined: best within setup.

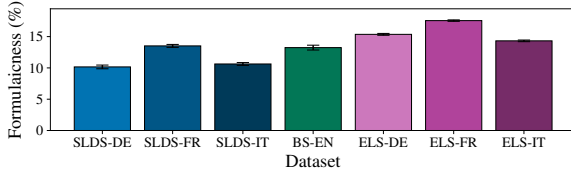


Figure 6: Average headnote formulaicness reported for the monolingual German, French, and Italian subsets in SLDS and EUR-Lex-Sum (ELS), and for the entire English dataset in the case of BillSum (BS-EN).

4.1 Traditional Metrics

We evaluate models on the SLDS test set using lighteval (Fourrier et al., 2023), reporting BERTScore (Zhang et al., 2020), BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004). Since each metric has known limitations (Zhang et al., 2020), we report all three to capture complementary aspects of summarization performance.⁵ For more details on the hyperparameters we used in certain metrics, refer to Appendix E.

4.2 LLM-as-a-Judge

We further adopt the LLM-as-a-Judge framework (Zheng et al., 2023), employing DeepSeek V3 (Liu et al., 2024) as the judge model due to its multilingual capabilities, low cost, and the fact that it was not among the evaluated models, avoiding bias toward self-generated outputs (Panickssery et al.,

2024). To enable a fair comparison with human evaluation, neither the LLM nor the human judges were shown the full decision text. Given the high quality of the gold headnotes, this setup provides a meaningful and token-efficient evaluation.

Evaluation Protocol The judge LLM evaluates generated headnotes against gold headnotes across five dimensions: (1) Accuracy & Faithfulness, (2) Completeness & Relevance, (3) Clarity & Coherence, (4) Articles (whether legal articles are correctly and completely referenced), and (5) Considerations (whether the correct legal considerations are identified and preserved). It provides a short analysis and assigns a score from 1 (major flaws) to 3 (close match) per category. Prompts and an example output are shown in Appendix H.3 and I. An example is shown in Fig. 4.⁶ For the full texts, refer to Appendix I.

Aggregation To compute the final score, each individual rating is first normalized from 1–3 to 0–2. The five normalized scores for a sample are then summed (for a maximum of 10) and multiplied by 10 to represent a percentage between 0 and 100. The final judge score is the average of these scaled values across all samples in the test set.

⁵We plan to contribute our evaluation task to lighteval to support reproducibility.

⁶The decision is also accessible on the SFSC repository under Decision 150 III 223

316 **4.3 Human Evaluations**

317 To obtain a trusted qualitative estimate of model
318 performance, we sampled 63 instances from the
319 test set, with seven per decision-headnote lan-
320 guage pairs across all nine subsets (such as de→de,
321 de→fr), resulting in a total of 189 generated head-
322 notes evaluated against 63 gold headnotes. Each
323 sample included the original headnote and outputs
324 from the top-performing models in three categories:
325 fine-tuned, frontier, and reasoning models. Two
326 co-authors, both professional lawyers fluent in the
327 relevant languages, assessed the samples using the
328 same protocol as the LLM judge. We prioritized
329 broader coverage across all language pairs in the
330 dataset over inter-annotator agreement, especially
331 since only one of the two experts is fluent in Italian.
332 Expanding the evaluation further was not feasible
333 due to the high cost of legal expertise, while using
334 less experienced annotators would have compro-
335 mised quality. Additionally, a third legal expert
336 and co-author of this paper conducted an in-depth
337 qualitative analysis of six selected samples, tak-
338 ing into account the full decision text. This setup
339 provided valuable expert insights while balancing
340 quality and feasibility.

341 **5 Results**

342 **5.1 Overall Results**

343 We present the results of our evaluations on the
344 SLDS test set in Table 2. We macro-averaged over
345 the scores in each of the nine language subsets of
346 decision and headnote language pairs to promote
347 fairness and robustness across languages. Below,
348 we highlight several interesting observations.

349 **Fine-tuned models perform well on automated**
350 **metrics, but lag in legal precision.** Although
351 smaller SLMs achieve lower JUDGE scores than
352 their larger counterparts, our results show that the
353 fine-tuned Qwen2.5 14B surpasses even signifi-
354 cantly larger proprietary models on standard met-
355 rics such as BERTScore, BLEU, ROUGE-2, and
356 ROUGE-L. ROUGE-1 scores for Qwen are also
357 notably high. This indicates that the fine-tuned
358 models excel in lexical similarity but still fall short
359 in legal correctness, completeness, and structural
360 fidelity when compared to large proprietary LLMs.
361 These findings underscore the limitations of tradi-
362 tional automated metrics and emphasize the need
363 for more nuanced evaluation methods based on
364 LLMs as judges. Nevertheless, fine-tuning on the

SLDS training split leads to a substantial improve-
ment in JUDGE scores on the test set, as illustrated
in Appendix J.

Large models are more accurate. Our 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 hypothesize that it
could be partially due to the one shot examples
provided in the prompt. Although we initially con-
sidered one-shot prompting for the fine-tuned mod-
els, it did not improve performance, likely because
these models had already learned the headnote for-
mat during training. Another interesting observa-
tion 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 cer-
tain metrics can be deceptive and that relying on a
single metric for evaluating summaries is usually
not sufficient.

Reasoning Models Offer Limited Gains Inter-
estingly, the reasoning models do not perform sig-
nificantly better. Even though DeepSeek R1 out-
performs all other models in terms of the JUDGE
score, the difference to Claude 3.5 Sonnet is only
one point. Moreover, o3-mini performs worse than
Claude 3.5 Sonnet and only slightly outperforms
our fine-tuned Qwen2.5 14B model by roughly 3.4
points. Our findings suggest that generating le-
gal headnotes primarily requires factual accuracy,
domain knowledge, and structured outputs, rather
than complex logical reasoning. The task primarily
demands models to faithfully extract and concisely
rephrase key legal principles, ensuring that refer-
ences 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 indi-
cates 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 deci-
sion and headnote language (*subsets*), e.g., de_fr
for decisions in German with French headnotes.
Key findings are summarized below with full de-
tails in Appendix Table 4. To facilitate the analysis
of JUDGE and BERTScores in the cross-lingual
settings, we provide heatmaps of selected models
in Figure 7 and Figure 8.

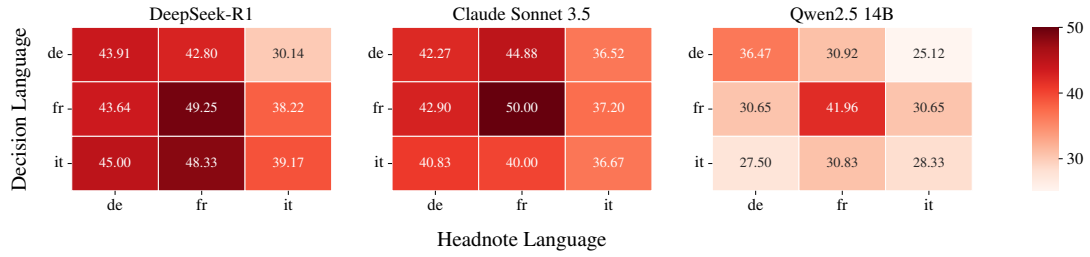


Figure 7: JUDGE scores for different cross-lingual language subsets and different models. Darker colors indicate better scores.

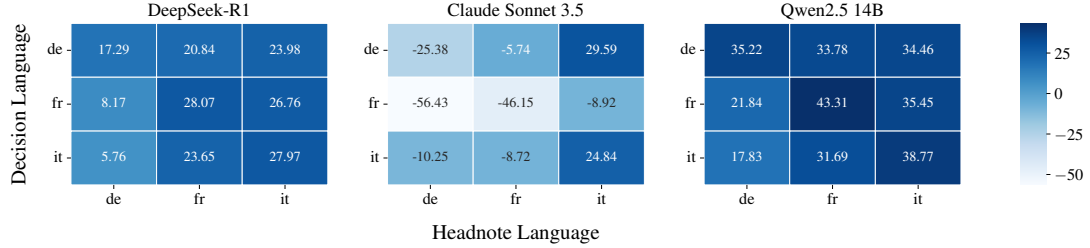


Figure 8: BERTScore for different cross-lingual language subsets and different models. Darker colors indicate better scores.

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 *weaker cross-lingual robustness* despite strong monolingual performance.

French headnotes often score highest. French headnotes tend to achieve higher JUDGE scores, particularly in the monolingual fr→fr setting. This trend also appears frequently, though not universally, in cross-lingual cases such as de→fr with Claude 3.5 Sonnet, it→fr with DeepSeek R1, and it→fr with Qwen2.5 14B. In cases where French is not the top-performing target language, the score differences are usually small. This may suggest either higher model proficiency in generating French legal text or that French headnotes are more systematically structured and easier to reproduce.

Limitations of general-purpose metrics. The heatmaps in Figures 7 and 8 reveal substantial inconsistencies between the two metrics. Some model outputs from Claude 3.5 Sonnet receive low BERTScores while achieving high JUDGE scores, indicating strong performance in legal correctness, completeness, and clarity. These observations underscore the limitations of general-purpose similarity metrics and emphasize the need for *domain-specific evaluation methods* in legal text generation.

6 Human Expert Evaluation

We perform two human expert evaluations. The first is based on the same evaluation process the LLM judge also follows. Two lawyers assess three

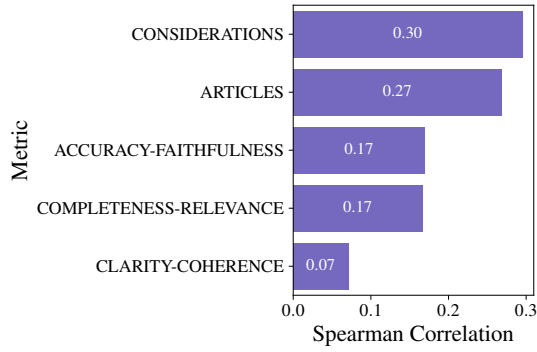
generated headnotes across 63 samples. This evaluation only considers the generated and the original headnote without taking into account the actual text of the landmark decision, assuming the gold headnote is the ideal headnote and that any deviation should be penalized. We refer to this evaluation as *Human-as-a-Judge*. In the second evaluation which we will refer to as *Contextualized Human Analysis*, another lawyer looked at six of those 63 samples and performed an in-depth analysis which involved studying the decision text as well.

6.1 Human-as-a-Judge

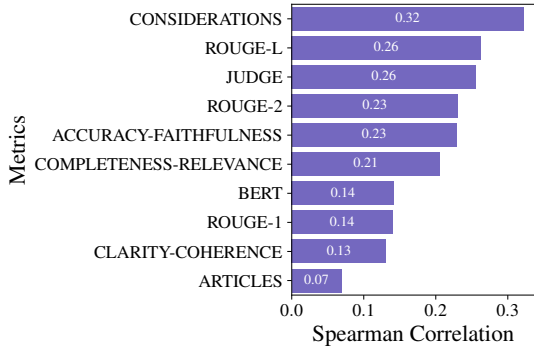
With 63 decisions and headnotes generated by three models, we obtained 189 annotated samples. Appendix Figure 13 illustrates score distributions assigned by both the LLM and the lawyers. The latter tend to give slightly higher scores than DeepSeek-V3, with a mean difference of 11.64, indicating that the LLM judge is stricter in its assessments.

Evaluation Metrics Figure 9 presents two correlation analyses assessing our legal headnote evaluation. Figure 9a shows Spearman correlations between DeepSeek-V3’s category-specific scores and human expert ratings across five dimensions. Figure 9b compares traditional metrics (ROUGE, BERTScore) and LLM-based judgments with aggregated human scores. We present our findings in the following paragraphs.

Correlation Analysis Figure 9 reveals important patterns in how automated evaluation approaches align with human judgment. Examining the category-wise correlations in Figure 9a, we find that objective elements of legal analysis show the



(a) Human Rubric to LLM Judge Rubric



(b) Overall Human Score to Rubrics

Figure 9: While LLM scores vary across categories, the overall JUDGE score remains highly correlated with human judgment. Notably, the considerations score, shows the strongest correlation with aggregated human scores.

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 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 9b 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.

6.2 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 deci-

sions, we observed significant variations in terms of reference accuracy, legal precision, and headnote appropriateness. The expert found that DeepSeek R1 produced closely aligned headnotes to 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. Finetuned Qwen2.5 14B 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, it 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. We show a additional analysis in Appendix L.

7 Conclusions and Future Work

We introduce SLDS, a large-scale cross-lingual resource for judicial summarization. We benchmark fine-tuned and proprietary models, revealing a trade-off between lexical similarity and legal accuracy. While fine-tuned models perform well on traditional summarization metrics, they struggle 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.

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, introducing potential subjectivity due to resource constraints, the high cost of legal annotations, and language barriers.

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 head-note 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

- 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.
- 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. 2018. [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.
- Ben Hachey and Claire Grover. 2006. [Extractive summarisation of legal texts](#). *Artificial Intelligence and Law*, 14(4):305–345.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- 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.
- Dominique Jakob. 2019. Merkblatt zum bundesgericht.
- 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. Billsun: 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.

659	Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang,	and their evaluation. In <i>Proceedings of the 2nd Con-</i>	714
660	Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi	<i>ference of the Asia-Pacific Chapter of the Association</i>	715
661	Deng, Chenyu Zhang, Chong Ruan, et al. 2024.	<i>for Computational Linguistics and the 12th Interna-</i>	716
662	Deepseek-v3 technical report. <i>arXiv preprint</i>	<i>tional Joint Conference on Natural Language Pro-</i>	717
663	<i>arXiv:2412.19437</i> .	<i>cessing</i> , pages 1048–1064.	718
664	I Loshchilov. 2017. Decoupled weight decay regulariza-	Tianyi Zhang, Varsha Kishore, Felix Wu, Kil-	719
665	<i>tion. arXiv preprint arXiv:1711.05101</i> .	ian Q. Weinberger, and Yoav Artzi. 2020.	720
666	Revital Ludewig and Juan Lallave. 2013. Professional	BERTScore: Evaluating Text Generation with	721
667	stress, discrimination and coping strategies: Similari-	BERT . <i>arXiv:1904.09675 [cs]</i> . ArXiv: 1904.09675.	722
668	ties and differences between female and male judges	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	723
669	in switzerland.	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	724
670	Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and	Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023.	725
671	Paolo Italiani. 2023. Multi-language transfer learn-	Judging llm-as-a-judge with mt-bench and chatbot	726
672	ing for low-resource legal case summarization. <i>Arti-</i>	<i>arena. Advances in Neural Information Processing</i>	727
673	<i>ficial Intelligence and Law</i> , pages 1–29.	<i>Systems</i> , 36:46595–46623.	728
674	Shashi Narayan, Shay B. Cohen, and Mirella Lapata.		
675	2018. Don’t give me the details, just the summary!		
676	topic-aware convolutional neural networks for ex-		
677	treme summarization . In <i>Proceedings of the 2018</i>		
678	<i>Conference on Empirical Methods in Natural Lan-</i>		
679	<i>guage Processing</i> , pages 1797–1807, Brussels, Bel-		
680	gium. Association for Computational Linguistics.		
681	Joel Niklaus and Daniele Giofr�. 2022. BudgetLong-		
682	former: Can we Cheaply Pretrain a SotA Legal		
683	Language Model From Scratch? <i>arXiv preprint</i> .		
684	ArXiv:2211.17135 [cs].		
685	Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christo-		
686	pher Hahn, Brian M. Rosen, Peter Henderson,		
687	Daniel E. Ho, Garrett Honke, Percy Liang, and		
688	Christopher Manning. 2024. FLawN-T5: An Em-		
689	pirical Examination of Effective Instruction-Tuning		
690	Data Mixtures for Legal Reasoning . <i>arXiv preprint</i> .		
691	ArXiv:2404.02127 [cs].		
692	Arjun Panickssery, Samuel R Bowman, and Shi Feng.		
693	2024. Llm evaluators recognize and favor their own		
694	generations. <i>arXiv preprint arXiv:2404.13076</i> .		
695	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-		
696	Jing Zhu. 2002. Bleu: A method for automatic evalu-		
697	ation of machine translation . In <i>Proceedings of the</i>		
698	<i>40th Annual Meeting on Association for Computa-</i>		
699	<i>tional Linguistics</i> , ACL ’02, page 311–318, USA.		
700	Association for Computational Linguistics.		
701	Luca Ragazzi, Gianluca Moro, Stefano Guidi, and Gia-		
702	como Frisoni. 2024. Lawsuit: a large expert-written		
703	summarization dataset of italian constitutional court		
704	verdicts. <i>Artificial Intelligence and Law</i> , pages 1–37.		
705	Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg,		
706	Margo Schlanger, and Doug Downey. 2022. Multi-		
707	LexSum: Real-World Summaries of Civil Rights		
708	Lawsuits at Multiple Granularities . <i>arXiv preprint</i> .		
709	ArXiv:2206.10883 [cs].		
710	Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Ra-		
711	jdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal,		
712	and Saptarshi Ghosh. 2022. Legal case document		
713	summarization: Extractive and abstractive methods		

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

Dataset Creation Pipeline

We developed a two-stage pipeline to collect and preprocess the dataset of decisions from the SFSC. The source documents are available via the official online archive, which publishes court decisions along with headnotes in the three official languages of Switzerland (German, French, and Italian).

Scraping We implemented an asynchronous scraping script to systematically retrieve all decisions published between 1954 and 2024 across five official volumes (I–V). For each entry, the script first accesses an index page for a given year and volume, then follows hyperlinks to individual decision pages. From each decision page, we extract the decision ID, metadata, the full German headnote, and the corresponding headnotes in French and Italian via linked language-specific pages. The full decision text is extracted from the website source while removing page breaks other irrelevant elements. To ensure robustness, the script uses exponential backoff to retry failed HTTP or timeout requests. Existing data are cached to allow resumable scraping.

Postprocessing. Once all raw data is collected, a postprocessing script performs several operations: (1) assignment of each decision to a *law area* (e.g., civil law, criminal law) based on the year and volume, following the official classification rules (Jakob, 2019) and historical documentation,⁷ (2) automatic detection of the *language of the decision* using the `langdetect` library, and (3) transformation of the dataset into a long format by *melting* the multilingual headnote columns into a single column with an associated language label.

The mapping from volumes to law areas is historically defined and has changed over time. Up to 1994 (volume 120), the structure included separate sub-volumes **Ia** and **Ib**, with the following assignments:

- **Ia** – Constitutional law
- **Ib** – Administrative law and international public law
- **II** – Civil law
- **III** – Debt enforcement and bankruptcy law
- **IV** – Criminal law and criminal procedure
- **V** – Social security law (successor of the EVGE series, 1926–1969)

Since 1995 (volume 121), the structure has been simplified to five volumes:

- **I** – Constitutional law
- **II** – Administrative law and international public law
- **III** – Civil law and debt enforcement/bankruptcy
- **IV** – Criminal law and criminal procedure
- **V** – Social security law

These assignments were implemented programmatically using a mapping table informed by both the court’s own documentation and secondary academic references.

The dataset is split into *training* (decisions from 1954 to 2021), *validation* (2022), and *test* (2023–2024) sets. Each entry receives a unique

⁷Summarized at https://de.wikipedia.org/w/index.php?title=Entscheidungen_des_Schweizerischen_Bundesgerichts&oldid=253293997#Gliederung.

sample_id. A predefined set of one-shot examples is additionally marked for each language pair based on the smallest sequence length in the validation set. The final dataset is then pushed to the Hugging Face Hub, including separate configurations for each decision-headnote language pair (e.g., de_fr, it_it).

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 10, we provide a distribution of Landmark Decisions (LDs) over the years.

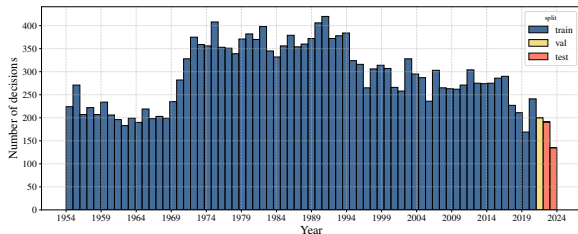


Figure 10: Number of landmark decisions published per year.

C.3 Properties related to Summarization

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

D Resources Used

For fine-tuning and the learning rate sweeps, we mostly used a single NVIDIA H100 GPU with 96 GB of VRAM. Some runs were performed on another node with two NVIDIA A100 GPUs with 80 GB of VRAM each. The total runtime of these experiments was 15.363 days.

E Hyperparameters Used in Metrics

For ROUGE, we employed the lighteval wrapper, which internally uses the rouge_score library with default settings—specifically, whitespace-based tokenization without stemming or additional preprocessing.

For BERTScore, we used xlm-roberta-large through the BERTScorer implementation in lighteval, setting rescale_with_baseline=True and num_layers=24. Language-specific baselines were obtained from the official BERTScore GitHub repository.

F Experiment Details

F.1 Exact Model Versions

For the proprietary models, we used the following model versions in our experiments: gpt-4o-2024-08-06, o3-mini-2025-01-31, claude-3-5-sonnet-20241022.

F.2 One-Shot Example Selection Strategy

To reduce the input sequence length and the associated costs, we selected the sample with the shortest sequence length in the validation split for each decision-headnote language pair and use them as the one-shot examples in our experiment. The one-shot example was provided in terms of a user- and assistant message pair in a multi-turn chat completions format.

F.3 Decoding Parameters

We used the default vLLM settings, with some modifications for Llama3.2 3B and the Qwen model family:

- repetition_penalty: 1.05
- temperature: 0.7
- top_k: 20
- top_p: 0.8

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

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 languages.

We used seed 2025 for reproducibility. Proprietary model APIs accepted only some of these parameters. More specifically, for OpenAI and DeepSeek models, we had to drop the repetition_penalty and the top_k parameter. The Anthropic API did not accept a repetition_penalty parameter either.

G 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 per-

formed 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 configuration along with the training and evaluation scripts can be found on our GitHub repository.

H Prompts

All 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.

H.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*.

H.2 Headnote generation

During the evaluation, we used the default system prompt of the model and [Prompt 2](#) as the user message 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*.

## H.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](#).

```
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 includes 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	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 mentions consid.  ↳ 3.3.3 twice, which is redundant.   CONSIDERATIONS_SCORE: 1
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

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 mentions 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.
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## **I Judge Example Output**

We show an example output of the DeepSeek V3 judge below.

## 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 erfolgreicher 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 lit. 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 11: 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.

## **I.1 Results on Language Subsets**

We provide the detailed results for the cross-lingual evaluations in our experiment in [Table 4](#).

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	31.30 ± 1.19	15.65 ± 0.91	22.76 ± 0.94	20.97 ± 1.55
Llama 3.2 3B	de_de	27.27 ± 1.43	47.59 ± 0.45	39.58 ± 1.02	20.72 ± 0.99	28.84 ± 0.94	28.50 ± 1.56
Qwen2.5 0.5B	de_de	16.37 ± 1.65	72.53 ± 0.41	31.51 ± 0.87	14.45 ± 0.64	23.08 ± 0.67	13.00 ± 1.15
Qwen2.5 1.5B	de_de	23.19 ± 1.49	<b>74.22 ± 0.44</b>	36.05 ± 0.92	17.72 ± 0.84	26.37 ± 0.86	21.88 ± 1.38
Qwen2.5 3B	de_de	28.22 ± 1.40	67.40 ± 0.41	39.31 ± 0.93	20.20 ± 0.88	29.10 ± 0.87	29.42 ± 1.62
Qwen2.5 7B	de_de	32.21 ± 1.24	72.18 ± 0.43	42.26 ± 0.98	22.78 ± 1.06	31.64 ± 1.01	33.09 ± 1.50
Qwen2.5 14B	de_de	<b>35.22 ± 1.22</b>	66.74 ± 0.43	<b>43.82 ± 0.94</b>	<b>24.54 ± 1.08</b>	<b>33.48 ± 1.02</b>	36.47 ± 1.60
GPT-4o	de_de	27.96 ± 0.90	39.94 ± 0.26	40.78 ± 0.69	18.72 ± 0.62	26.97 ± 0.56	40.58 ± 1.33
DeepSeek-R1	de_de	17.29 ± 0.74	29.69 ± 0.19	36.04 ± 0.64	15.01 ± 0.44	21.02 ± 0.38	<b>43.91 ± 1.16</b>
o3-mini	de_de	13.78 ± 0.73	31.34 ± 0.20	33.04 ± 0.54	11.85 ± 0.40	18.18 ± 0.34	36.52 ± 1.09
Claude 3.5 Sonnet	de_de	-25.38 ± 25.95	27.00 ± 0.28	40.50 ± 0.77	18.66 ± 0.72	29.24 ± 0.76	42.27 ± 1.41
Phi-3.5-mini	de_fr	4.71 ± 2.47	50.73 ± 0.52	27.36 ± 1.18	11.48 ± 0.63	18.50 ± 0.76	13.57 ± 1.33
Llama 3.2 3B	de_fr	24.84 ± 1.62	18.07 ± 0.41	35.29 ± 0.92	15.16 ± 0.53	24.03 ± 0.62	19.08 ± 1.40
Qwen2.5 0.5B	de_fr	-3.81 ± 2.18	20.30 ± 0.50	22.33 ± 0.92	7.23 ± 0.42	15.77 ± 0.55	3.29 ± 0.48
Qwen2.5 1.5B	de_fr	21.71 ± 1.61	25.19 ± 0.38	33.69 ± 0.87	13.08 ± 0.53	22.28 ± 0.55	11.79 ± 1.09
Qwen2.5 3B	de_fr	26.37 ± 1.32	40.22 ± 0.32	35.87 ± 0.76	14.39 ± 0.48	24.06 ± 0.50	18.55 ± 1.29
Qwen2.5 7B	de_fr	32.61 ± 1.06	<b>52.55 ± 0.32</b>	40.56 ± 0.74	17.94 ± 0.59	26.69 ± 0.56	26.47 ± 1.52
Qwen2.5 14B	de_fr	33.78 ± 1.15	40.47 ± 0.41	40.67 ± 0.80	19.44 ± 0.63	28.30 ± 0.63	30.92 ± 1.55
GPT-4o	de_fr	<b>33.97 ± 0.76</b>	30.45 ± 0.21	<b>45.47 ± 0.61</b>	<b>20.65 ± 0.50</b>	27.59 ± 0.42	40.14 ± 1.42
DeepSeek-R1	de_fr	20.84 ± 0.61	24.25 ± 0.15	39.69 ± 0.66	16.36 ± 0.37	21.49 ± 0.29	42.80 ± 1.24
o3-mini	de_fr	15.68 ± 0.62	20.86 ± 0.15	36.99 ± 0.56	13.11 ± 0.31	18.78 ± 0.25	35.70 ± 1.33
Claude 3.5 Sonnet	de_fr	-5.74 ± 0.94	27.23 ± 0.21	43.15 ± 0.64	19.34 ± 0.55	<b>28.88 ± 0.56</b>	<b>44.88 ± 1.48</b>
Phi-3.5-mini	de_it	8.06 ± 2.28	30.39 ± 0.47	25.85 ± 1.03	9.89 ± 0.52	18.08 ± 0.72	9.61 ± 1.09
Llama 3.2 3B	de_it	22.81 ± 1.60	14.32 ± 0.41	31.47 ± 0.78	12.65 ± 0.50	22.46 ± 0.56	13.72 ± 1.28
Qwen2.5 0.5B	de_it	4.48 ± 1.89	<b>48.16 ± 0.38</b>	22.19 ± 0.76	7.62 ± 0.34	16.35 ± 0.50	2.17 ± 0.40
Qwen2.5 1.5B	de_it	22.99 ± 1.30	41.46 ± 0.33	30.71 ± 0.69	10.86 ± 0.42	21.31 ± 0.50	8.16 ± 0.88
Qwen2.5 3B	de_it	23.86 ± 1.50	31.39 ± 0.33	32.30 ± 0.77	12.41 ± 0.47	22.53 ± 0.56	12.46 ± 1.24
Qwen2.5 7B	de_it	30.75 ± 1.00	31.86 ± 0.34	35.74 ± 0.71	14.77 ± 0.53	24.99 ± 0.56	20.39 ± 1.44
Qwen2.5 14B	de_it	<b>34.46 ± 0.95</b>	45.34 ± 0.35	37.86 ± 0.68	16.38 ± 0.54	26.72 ± 0.54	25.12 ± 1.44
GPT-4o	de_it	32.12 ± 0.69	30.40 ± 0.25	39.05 ± 0.58	15.51 ± 0.49	24.78 ± 0.44	29.66 ± 1.29
DeepSeek-R1	de_it	23.98 ± 0.55	12.77 ± 0.17	36.07 ± 0.53	12.72 ± 0.34	20.15 ± 0.29	30.14 ± 1.26
o3-mini	de_it	15.90 ± 0.52	15.63 ± 0.14	30.70 ± 0.45	7.79 ± 0.24	15.58 ± 0.23	27.83 ± 1.23
Claude 3.5 Sonnet	de_it	29.59 ± 0.88	29.52 ± 0.26	<b>43.26 ± 0.67</b>	<b>20.46 ± 0.63</b>	<b>29.56 ± 0.62</b>	<b>36.52 ± 1.46</b>
Phi-3.5-mini	fr_de	-6.11 ± 3.27	38.47 ± 0.41	24.14 ± 1.27	8.92 ± 0.61	16.55 ± 0.82	8.69 ± 1.56
Llama 3.2 3B	fr_de	1.58 ± 2.44	49.67 ± 0.37	25.75 ± 1.11	10.72 ± 0.67	19.26 ± 0.83	10.65 ± 1.56
Qwen2.5 0.5B	fr_de	-10.66 ± 2.47	33.38 ± 0.39	21.14 ± 0.95	6.93 ± 0.51	15.61 ± 0.64	2.71 ± 0.60
Qwen2.5 1.5B	fr_de	0.62 ± 2.21	27.16 ± 0.35	26.46 ± 0.97	9.37 ± 0.57	18.88 ± 0.65	7.10 ± 1.18
Qwen2.5 3B	fr_de	7.68 ± 2.03	28.04 ± 0.32	28.78 ± 0.96	10.87 ± 0.58	20.36 ± 0.65	13.36 ± 1.48
Qwen2.5 7B	fr_de	15.63 ± 1.80	<b>50.67 ± 0.31</b>	33.45 ± 0.91	12.38 ± 0.63	22.58 ± 0.62	22.90 ± 2.01
Qwen2.5 14B	fr_de	<b>21.84 ± 1.51</b>	41.26 ± 0.34	35.59 ± 0.90	14.74 ± 0.71	24.70 ± 0.66	30.65 ± 1.97
GPT-4o	fr_de	21.02 ± 1.03	31.29 ± 0.21	<b>38.97 ± 0.72</b>	<b>15.74 ± 0.54</b>	24.48 ± 0.50	41.12 ± 1.64
DeepSeek-R1	fr_de	8.17 ± 1.01	20.77 ± 0.17	33.01 ± 0.70	12.30 ± 0.42	19.03 ± 0.41	<b>43.64 ± 1.40</b>
o3-mini	fr_de	0.81 ± 0.88	19.15 ± 0.18	28.94 ± 0.56	7.89 ± 0.34	15.78 ± 0.33	28.69 ± 1.72
Claude 3.5 Sonnet	fr_de	-56.43 ± 50.10	0.00 ± 0.26	36.83 ± 0.74	14.80 ± 0.60	<b>24.78 ± 0.60</b>	42.90 ± 1.93
Phi-3.5-mini	fr_fr	18.62 ± 3.27	49.91 ± 0.54	36.72 ± 1.64	18.45 ± 1.22	24.61 ± 1.15	24.58 ± 2.09
Llama 3.2 3B	fr_fr	24.86 ± 3.03	4.32 ± 0.61	39.08 ± 1.83	21.49 ± 1.42	26.75 ± 1.30	33.36 ± 2.22
Qwen2.5 0.5B	fr_fr	14.65 ± 3.22	<b>51.91 ± 0.50</b>	32.02 ± 1.59	15.80 ± 1.08	22.12 ± 1.03	14.30 ± 1.81
Qwen2.5 1.5B	fr_fr	33.37 ± 2.17	41.51 ± 0.47	42.66 ± 1.35	23.66 ± 1.09	29.17 ± 1.04	31.50 ± 1.92
Qwen2.5 3B	fr_fr	34.57 ± 2.18	47.78 ± 0.41	44.14 ± 1.37	24.20 ± 1.18	30.24 ± 1.13	35.42 ± 1.93
Qwen2.5 7B	fr_fr	39.91 ± 1.48	51.20 ± 0.42	47.91 ± 1.08	26.80 ± 1.04	32.55 ± 0.93	38.97 ± 1.90
Qwen2.5 14B	fr_fr	<b>43.31 ± 1.26</b>	42.67 ± 0.44	50.06 ± 1.10	<b>29.13 ± 1.17</b>	<b>34.69 ± 1.03</b>	41.96 ± 1.99
GPT-4o	fr_fr	40.20 ± 0.96	44.32 ± 0.28	<b>50.66 ± 0.81</b>	26.53 ± 0.83	31.05 ± 0.69	48.04 ± 1.48
DeepSeek-R1	fr_fr	28.07 ± 0.85	31.18 ± 0.20	43.28 ± 0.93	21.53 ± 0.61	23.95 ± 0.50	49.25 ± 1.38
o3-mini	fr_fr	25.92 ± 0.86	34.85 ± 0.21	44.01 ± 0.82	20.09 ± 0.60	23.58 ± 0.45	43.93 ± 1.47
Claude 3.5 Sonnet	fr_fr	-46.15 ± 42.17	17.32 ± 0.24	46.57 ± 0.85	22.12 ± 0.75	30.57 ± 0.76	<b>50.00 ± 1.99</b>
Phi-3.5-mini	fr_it	17.03 ± 2.96	25.76 ± 0.47	31.07 ± 1.43	12.63 ± 0.77	20.79 ± 0.94	13.18 ± 1.62
Llama 3.2 3B	fr_it	22.19 ± 2.42	4.98 ± 0.47	32.31 ± 1.32	14.29 ± 0.87	22.77 ± 0.95	17.57 ± 1.82
Qwen2.5 0.5B	fr_it	5.93 ± 2.73	21.94 ± 0.37	24.88 ± 1.15	9.53 ± 0.64	17.93 ± 0.76	3.36 ± 0.70
Qwen2.5 1.5B	fr_it	26.50 ± 1.77	38.52 ± 0.34	34.46 ± 0.92	13.10 ± 0.66	22.93 ± 0.68	12.80 ± 1.34
Qwen2.5 3B	fr_it	28.52 ± 1.93	39.51 ± 0.34	35.37 ± 1.08	15.02 ± 0.76	24.62 ± 0.84	17.76 ± 1.82
Qwen2.5 7B	fr_it	31.50 ± 1.79	<b>45.05 ± 0.31</b>	37.51 ± 1.17	16.43 ± 0.79	25.69 ± 0.80	24.30 ± 2.04
Qwen2.5 14B	fr_it	35.45 ± 1.53	44.31 ± 0.33	40.03 ± 1.17	19.37 ± 0.92	28.54 ± 0.95	30.65 ± 1.98
GPT-4o	fr_it	<b>36.37 ± 1.01</b>	31.56 ± 0.25	42.97 ± 0.79	18.84 ± 0.66	26.81 ± 0.65	32.71 ± 1.66
DeepSeek-R1	fr_it	26.76 ± 0.91	21.21 ± 0.17	38.08 ± 0.86	15.46 ± 0.54	21.31 ± 0.48	<b>38.22 ± 1.66</b>
o3-mini	fr_it	22.98 ± 0.88	15.31 ± 0.19	36.12 ± 0.65	11.22 ± 0.41	19.34 ± 0.43	29.91 ± 1.60
Claude 3.5 Sonnet	fr_it	-8.92 ± 38.98	24.62 ± 0.29	<b>45.12 ± 0.96</b>	<b>22.30 ± 0.84</b>	<b>30.11 ± 0.79</b>	37.20 ± 1.80
Phi-3.5-mini	it_de	0.53 ± 6.69	20.35 ± 0.23	27.05 ± 3.61	10.75 ± 1.89	17.19 ± 1.69	5.83 ± 2.60
Llama 3.2 3B	it_de	-3.89 ± 5.97	15.89 ± 0.21	24.22 ± 3.08	10.13 ± 1.91	17.67 ± 2.38	7.50 ± 3.92
Qwen2.5 0.5B	it_de	-23.28 ± 5.94	9.64 ± 0.18	16.15 ± 2.65	5.97 ± 1.11	12.09 ± 1.66	0.00 ± 0.00
Qwen2.5 1.5B	it_de	4.91 ± 2.90	15.66 ± 0.23	27.51 ± 2.49	9.62 ± 1.36	18.77 ± 1.52	4.17 ± 2.29
Qwen2.5 3B	it_de	4.32 ± 5.98	10.03 ± 0.26	28.31 ± 3.07	9.06 ± 1.41	18.70 ± 1.71	10.83 ± 3.36
Qwen2.5 7B	it_de	14.69 ± 3.46	21.69 ± 0.27	33.39 ± 2.81	12.95 ± 2.20	21.07 ± 1.92	23.33 ± 6.20
Qwen2.5 14B	it_de	<b>17.83 ± 3.40</b>	<b>28.24 ± 0.36</b>	31.46 ± 2.54	14.68 ± 2.07	22.35 ± 2.15	27.50 ± 6.17
GPT-4o	it_de	14.71 ± 2.94	21.30 ± 0.20	34.98 ± 3.34	14.19 ± 1.76	21.21 ± 1.82	41.67 ± 5.34
DeepSeek-R1	it_de	5.76 ± 2.42	22.03 ± 0.18	35.15 ± 3.76	13.41 ± 1.41	17.94 ± 1.55	<b>45.00 ± 3.99</b>
o3-mini	it_de	-6.59 ± 1.74	5.54 ± 0.13	25.97 ± 2.53	6.71 ± 0.68	13.16 ± 0.93	34.17 ± 3.79
Claude 3.5 Sonnet	it_de	-10.25 ± 3.24	22.41 ± 0.20	<b>37.18 ± 2.77</b>	<b>14.86 ± 1.53</b>	<b>23.24 ± 2.04</b>	40.83 ± 5.29
Phi-3.5-mini	it_fr	15.30 ± 8.17	30.01 ± 0.32	33.66 ± 4.87	15.59 ± 2.84	21.46 ± 3.02	13.33 ± 3.76

Model	Subset	BERTScore $\uparrow$	BLEU $\uparrow$	ROUGE-1 $\uparrow$	ROUGE-2 $\uparrow$	ROUGE-L $\uparrow$	JUDGE $\uparrow$
Llama 3.2 3B	it_fr	11.77 $\pm$ 9.72	9.48 $\pm$ 0.36	31.36 $\pm$ 5.09	14.07 $\pm$ 3.08	20.35 $\pm$ 3.01	17.50 $\pm$ 6.64
Qwen2.5 0.5B	it_fr	-23.29 $\pm$ 6.14	8.88 $\pm$ 0.18	17.07 $\pm$ 3.10	5.40 $\pm$ 1.45	12.95 $\pm$ 1.66	9.17 $\pm$ 3.36
Qwen2.5 1.5B	it_fr	20.02 $\pm$ 5.31	24.91 $\pm$ 0.22	32.04 $\pm$ 3.87	13.53 $\pm$ 2.01	20.87 $\pm$ 1.87	17.50 $\pm$ 4.63
Qwen2.5 3B	it_fr	27.60 $\pm$ 3.78	<b>39.09 <math>\pm</math> 0.32</b>	36.43 $\pm$ 3.68	15.66 $\pm$ 2.70	22.57 $\pm$ 1.85	25.00 $\pm$ 5.71
Qwen2.5 7B	it_fr	31.67 $\pm$ 2.34	23.05 $\pm$ 0.24	39.93 $\pm$ 2.92	19.09 $\pm$ 1.84	25.36 $\pm$ 1.55	34.17 $\pm$ 4.99
Qwen2.5 14B	it_fr	31.69 $\pm$ 3.27	35.41 $\pm$ 0.28	37.40 $\pm$ 3.25	16.76 $\pm$ 2.31	22.95 $\pm$ 1.45	30.83 $\pm$ 7.12
GPT-4o	it_fr	<b>33.10 <math>\pm</math> 3.64</b>	31.58 $\pm$ 0.23	<b>45.76 <math>\pm</math> 4.22</b>	<b>20.92 <math>\pm</math> 2.48</b>	<b>26.60 <math>\pm</math> 1.98</b>	43.33 $\pm$ 4.66
DeepSeek-R1	it_fr	23.65 $\pm$ 3.24	19.29 $\pm$ 0.19	43.50 $\pm$ 4.22	19.51 $\pm$ 1.87	22.92 $\pm$ 1.75	<b>48.33 <math>\pm</math> 4.41</b>
o3-mini	it_fr	17.25 $\pm$ 3.07	16.06 $\pm$ 0.14	39.77 $\pm$ 3.90	13.73 $\pm$ 1.77	20.26 $\pm$ 1.57	38.33 $\pm$ 4.41
Claude 3.5 Sonnet	it_fr	-8.72 $\pm$ 3.58	19.08 $\pm$ 0.23	42.18 $\pm$ 3.96	18.76 $\pm$ 2.71	25.64 $\pm$ 2.67	40.00 $\pm$ 5.50
Phi-3.5-mini	it_it	36.33 $\pm$ 2.62	27.64 $\pm$ 0.21	<b>43.65 <math>\pm</math> 2.48</b>	23.63 $\pm$ 2.08	28.72 $\pm$ 2.12	27.50 $\pm$ 5.24
Llama 3.2 3B	it_it	5.40 $\pm$ 11.34	32.69 $\pm$ 0.52	27.97 $\pm$ 5.91	14.61 $\pm$ 4.54	20.23 $\pm$ 3.83	18.33 $\pm$ 6.49
Qwen2.5 0.5B	it_it	7.31 $\pm$ 8.42	23.08 $\pm$ 0.28	27.58 $\pm$ 3.09	12.25 $\pm$ 2.32	20.48 $\pm$ 2.37	4.17 $\pm$ 2.88
Qwen2.5 1.5B	it_it	24.95 $\pm$ 5.68	<b>42.49 <math>\pm</math> 0.35</b>	33.68 $\pm$ 3.47	16.30 $\pm$ 2.21	23.44 $\pm$ 2.50	28.33 $\pm$ 5.75
Qwen2.5 3B	it_it	27.92 $\pm$ 5.05	42.30 $\pm$ 0.34	36.14 $\pm$ 3.46	19.11 $\pm$ 2.62	24.70 $\pm$ 2.43	20.00 $\pm$ 5.50
Qwen2.5 7B	it_it	37.34 $\pm$ 3.52	24.37 $\pm$ 0.41	42.38 $\pm$ 2.96	21.22 $\pm$ 2.61	27.41 $\pm$ 2.40	31.67 $\pm$ 6.01
Qwen2.5 14B	it_it	<b>38.77 <math>\pm</math> 3.58</b>	31.79 $\pm$ 0.36	43.45 $\pm$ 4.30	<b>24.88 <math>\pm</math> 3.30</b>	<b>30.33 <math>\pm</math> 3.13</b>	28.33 $\pm$ 4.90
GPT-4o	it_it	34.48 $\pm$ 3.73	26.14 $\pm$ 0.34	40.44 $\pm$ 4.33	19.15 $\pm$ 3.11	23.81 $\pm$ 2.41	<b>40.00 <math>\pm</math> 5.08</b>
DeepSeek-R1	it_it	27.97 $\pm$ 2.70	20.12 $\pm$ 0.19	39.91 $\pm$ 4.11	17.47 $\pm$ 1.71	21.47 $\pm$ 1.91	39.17 $\pm$ 3.36
o3-mini	it_it	21.87 $\pm$ 2.50	26.18 $\pm$ 0.17	37.37 $\pm$ 2.83	14.92 $\pm$ 1.48	19.27 $\pm$ 1.49	38.33 $\pm$ 5.05
Claude 3.5 Sonnet	it_it	24.84 $\pm$ 4.07	29.71 $\pm$ 0.29	41.98 $\pm$ 3.36	21.75 $\pm$ 2.39	27.05 $\pm$ 2.05	36.67 $\pm$ 5.27

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 and 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. ROUGE scores are multiplied by 100 for readability. JUDGE = LLM as Judge. **Bold**: best within subset.

## J Off-the-Shelf Performance

To investigate how well smaller pre-trained models perform in a zero-shot setting, we compare them with their finetuned counterparts in Figure 12. We observe a large performance gap in terms of the JUDGE score between the two settings, highlighting the benefits of fine-tuning on SLDS.

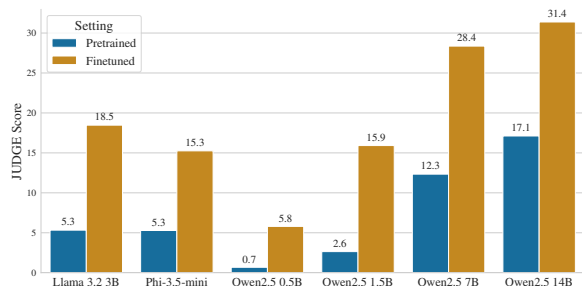


Figure 12: Comparison of the JUDGE scores between pre-trained and fine-tuned models on the test split of SLDS. Fine-tuned models outperform the pre-trained models by a large margin.

## K Distribution of Judgment Scores

Figure 13 provides an overview of the scores that were assigned by the LLM judge (left) and the human judges (right).

## L Contextualized Expert Evaluation of LLM-Generated Headnotes

Below, we present summaries of the expert commentary provided by our legal expert as part of the contextualized evaluation. The expert reviewed each case with access to the full decision, the official (gold) headnote, and the headnotes generated by different models, without being informed of the model identities. To save space, we do not reproduce the full decisions here; however, they remain accessible via their sample IDs or through the SFSC repository. We present English translations of the official headnote as well as the model-generated versions from Claude 3.5 Sonnet, DeepSeek R1, and our fine-tuned Qwen2.5 14B.

### Decision 1 (Sample ID: 60465)

#### Original Headnote

Art. 10 and 13 ECHR; Art. 16, 29a, 35 and 93 para. 3 Cst.; Art. 2 let. cbis, Art. 5a, 25 para. 3 let. b, Art. 83 para. 1 let. a and Art. 93 para. 1 and 95 para. 1 RTVA; Art. 28 ff. CC; Art. 1, 3, 5 para. 4 and Art. 18 of the SRG Concession;

Deletion of a user comment on Instagram by the SRG in other journalistic offerings (OJO); Legal recourse. The SRG is bound by fundamental rights in its other journalistic offerings (OJO); this applies - because of the close content-related connection between its editorial contribution and the user comments thereto - also insofar as it deletes such comments based on its "Netiquette" (E. 2). Since, due to lack of effectiveness, neither civil, criminal nor supervisory legal recourse in this context meets the requirements of Art. 29a Cst., the administrative legal recourse via the SRG Ombuds Office, which has a mediating function, is to be opened to the Independent Complaints Authority for Radio and Television (ICA) (E. 3). The specifications in the "Netiquette" are to be handled in individual cases analogous to the jurisprudence on the fundamental rights commitment of the SRG in the advertising sector (BGE 139 I 306) (E. 4).

### Model 1 (Claude 3.5 Sonnet)

Art. 29a BV, Art. 35 para. 2 BV, Art. 83 para. 1 lit. a RTVG; Fundamental rights obligation of SRG in the deletion of social media comments and legal process. The SRG is bound by fundamental rights when deleting user comments on editorial contributions in the remaining journalistic offerings (RJO) on social media platforms. The deletion of a comment constitutes an evaluative editorial act that interferes with the freedom of expression of the commenting person. The guarantee of legal recourse (Art. 29a BV) requires effective judicial protection. Neither the civil and criminal legal route nor the supervisory procedure at BAKOM meets these constitutional requirements. The UBI is responsible for assessing the legality of comment deletions, and complaints must be submitted there after an unsuccessful ombudsman procedure (E. 2-4).

**Expert Commentary:** The system generates only a limited number of references compared to the original headnote and excerpt. It identifies the core

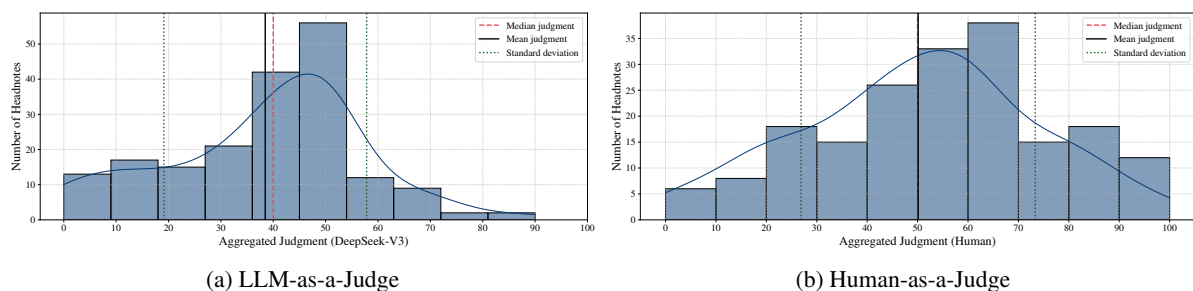


Figure 13: Distributions of (a) the scores generated by DeepSeek-V3 and (b) the scores assigned by two lawyers. The scores are aggregates of the individual scores per evaluation category, ranging from 0 to 100. The scores issued by the lawyers are slightly higher than the ones assigned by DeepSeek-V3.

of the decision quite accurately and uses terminology similar to both the excerpt and the original headnote. While the part on the “deletion of comments” is correct, the addition of “interference with the personal freedom” is problematic, as it is not mentioned in the decision or the original headnote. Regarding “legal recourse,” the output partially contradicts the main judgment: whereas the decision affirms that a civil law remedy can meet the requirements, the headnote suggests otherwise. However, the part on judicial protection is accurate. The role of the UBI is summarized precisely.

## Model 2 (DeepSeek R1)

Art. 35 Cst, Art. 29a Cst, Art. 83 LRTV; Obligation of the SRG to adhere to fundamental rights when deleting user-generated comments in the rest of the journalistic offer. The Swiss Broadcasting Corporation (SRG) is bound to fundamental rights when moderating comments in its social media channels within the framework of the rest of the journalistic offer (üjO), as it acts in the exercise of a public task (Art. 35 para. 2 Cst). The close factual connection between editorial contributions and user comments establishes a functional unit, which obliges the SRG to balance conflicting interests in conformity with fundamental rights (para. 2.2.3, 2.3.1). The Independent Complaints Authority for Radio and Television (ICA) is competent under Art. 83 para. 1 letter a RTVG to decide on complaints against the deletion of comments, as these are to be qualified as editorial factual acts in the sense of Art. 2 letter cbis RTVG (para. 3.3.1, 3.3.4). The civil or supervisory legal process does not meet the requirements of the

guarantee of legal recourse (Art. 29a Cst) and effective legal protection (Art. 13 ECHR) because it does not allow for individual fundamental rights control (para. 3.2.2, 3.2.3). The SRG must justify the deletion of comments based on objective criteria, whereby the “netiquette” is to be interpreted in the light of freedom of expression (Art. 16 Cst) and may only intervene if there are significant reasons (e.g., discrimination, glorification of violence) (para. 4.1, 4.2). The ICA must examine in individual cases whether the SRG has respected its obligation to fundamental rights.

**Expert Commentary:** The generated headnote includes fewer references than the original headnote and excerpt. It states the context of the case precisely. However, it focuses more on the role of the SRG and includes the appellant’s claim under Art. 35 para. 2. Since such claims are part of a party’s pleading, they should not be included in a headnote. This is why the original headnote does not mention them. Overall, the output reads more like a summary than a proper headnote.

## Model 3 (Qwen2.5 14B)

Art. 35 para. 2, Art. 16 para. 2, Art. 29a Cst.; Art. 10 ECHR; Art. 2 let. cbis, Art. 83 para. 1 let. a, Art. 86 para. 1, Art. 94 et seq. RTVA; Comment deletion by SRG in the remaining journalistic offering (RJO). The comment function on social media contributions within the remaining journalistic offering (RJO) is closely linked to the concessionary tasks assigned to SRG in this regard and financed by radio and television fees in the program area (E. 2.2.2). Therefore, SRG is also bound by fundamental

1137	rights in this context (E. 2.2.3). The ap-	Art. 60a LEaux; art. 127 para. 1 Cst.;	1185
1138	peals body (UBI) is competent to assess	cost coverage principle. The cost cover-	1186
1139	whether SRG impermissibly infringed	age principle does not have the status of	1187
1140	on the freedom of expression of the com-	an independent constitutional right. Its	1188
1141	ment author by deleting a comment in	constitutional function is solely to relax	1189
1142	the remaining journalistic offering (RJO)	the requirements of the formal legal ba-	1190
1143	(E. 3 and 4).	sis when the nature of the contribution	1191
1144	<b>Expert Commentary:</b> This is the only output that	allows it. Art. 60a LEaux implements	1192
1145	includes a reference to the ECHR along with other	the principle of causality and obliges the	1193
1146	legal provisions. Although not mentioned in the	cantons to pass on to the responsible par-	1194
1147	original headnote, this reference is legally impor-	ties the costs of construction, operation,	1195
1148	tant. The roles of the named bodies are stated	maintenance, remediation, and replace-	1196
1149	precisely without unnecessary detail, which is ap-	ment of wastewater disposal facilities	1197
1150	propriate for a headnote.	used for public purposes, but does not	1198
1151	<b>Decision 2 (Sample ID: 60487)</b>	confer upon the responsible parties an in-	1199
1152	<b>Original Headnote</b>	dividual right under federal law to com-	1200
1153	Art. 127 para. 1 Cst.; art. 60a LEaux;	pliance with the cost coverage principle	1201
1154	principle of legality; principle of cost	(considerations 3.5 and 3.8).	1202
1155	coverage; principle of causality. Accord-	<b>Expert Commentary:</b> The references are men-	1203
1156	ing to jurisprudence, the principle of cost	tioned correctly, matching the original headnote.	1204
1157	coverage, together with the principle of	The output identifies the principles of the case and	1205
1158	equivalence, serves as a substitute for	states them with precision. While the original head-	1206
1159	a formal legal basis in determining the	note leaves open the constitutional status of the	1207
1160	amount of a causal contribution (consid-	cost coverage principle, this version clearly states	1208
1161	erations 3.2 and 3.3). The question of	that it is not constitutionally protected, which re-	1209
1162	whether the principle of cost coverage it-	fects a direct reading of the judgment. However,	1210
1163	self has constitutional status can remain	in referring to Art. 60a, it incorrectly names the	1211
1164	open. In any case, it does not consti-	principle of causality instead of the “polluter pays”	1212
1165	tute a distinct and autonomous constitu-	principle.	1213
1166	tional right from the principle of legality,	<b>Model 2 (DeepSeek R1)</b>	1214
1167	which should be respected even when the	Art. 127 para. 1 Cst., 60a EPA; cost cov-	1215
1168	formal law provides sufficiently precise	erage principle. The cost coverage prin-	1216
1169	criteria for calculating the contribution	ciple does not constitute an autonomous	1217
1170	(considerations 3.4 and 3.5). If the law	constitutional right but acts as a substi-	1218
1171	on contributions stipulates, in addition	tute for the principle of legality when the	1219
1172	to a sufficiently precise calculation ba-	formal legal basis of a fee does not suf-	1220
1173	sis, compliance with the principle of cost	ficiently determine its extent. If the law	1221
1174	coverage, the latter applies by virtue of	precisely defines the calculation methods	1222
1175	the law and not the Constitution (clari-	of a fee (in this case, water and wastew-	1223
1176	fication of the case law; consideration	ater connection charges), the cost cover-	1224
1177	3.6). Art. 60a LEaux does not result	age principle does not apply to constitu-	1225
1178	in an individual right under federal leg-	tional review (recital 3.5). Art. 60a EPA,	1226
1179	islation for cantonal or communal con-	which enshrines the polluter-pays princi-	1227
1180	tributions levied in accordance with this	ple in the financing of purification infras-	1228
1181	provision to comply with the principle	tructure, does not entail a principle of full	1229
1182	of cost coverage (clarification of the case	cost coverage by the fees collected. Its	1230
1183	law; considerations 3.7-3.10).	objective is to transfer the burdens from	1231
1184	<b>Model 1 (Claude 3.5 Sonnet)</b>	public communities to users, without pro-	1232
		hibiting a financial surplus as long as an	1233
		adequate legal basis exists (recital 3.7.4).	1234

1235	The review of cantonal or communal fees	obligations of the State. An interven-	1283
1236	concerning the cost coverage principle	tion by the Federal Council with France	1284
1237	falls, in the absence of alleged violation	on the grounds that the authorities of	1285
1238	of federal law or cantonal constitutional	that State allegedly violated the principle	1286
1239	rights, within the sovereign discretion of	of specialty pertains to foreign relations	1287
1240	local authorities. The Federal Court con-	within the meaning of Art. 32 para. 1	1288
1241	ducts only a limited arbitrary review (art.	let. a LTAF (paras. 4 and 5). However,	1289
1242	9 Cst.) when formal legality is respected	access to the courts may be opened under	1290
1243	(recitals 3.6, 3.9).(Judgment 149 I 305 of	the counter-exception of Art. 32 para. 1	1291
1244	the Federal Court)	let. a in fine LTAF (paras. 6.1 and 6.2).	1292
1245	<b>Expert Commentary:</b> The references are correctly	In this case, it is justified, even if the	1293
1246	identified, along with the theme of the judgment.	challenged decision originates from the	1294
1247	This output defines the cost coverage principle	Federal Council in a matter not listed in	1295
1248	more clearly than the first one and aligns more	Art. 33 let. a and b LTAF, to appeal to	1296
1249	closely with the original headnote. It also supports	the Federal Administrative Court before	1297
1250	its content with references to the recitals in the	the Federal Supreme Court (para. 6.3).	1298
1251	judgment, which adds clarity to the headnote.	Art. 8 ECHR did not, in this instance, im-	1299
1252	<b>Model 3 (Qwen2.5 14B)</b>	pose a positive obligation on the Federal	1300
1253	Art. 127 paras. 1 and 2, Art. 8 para. 1	Council to issue an injunction to France	1301
1254	and Art. 9 FC; cost coverage principle.	and therefore cannot open access to the	1302
1255	The cost coverage is a principle of con-	courts under Art. 13 ECHR in connec-	1303
1256	stitutional law, but it cannot be invoked	tion with Art. 8 ECHR (paras. 6.4 and	1304
1257	independently of the principle of legal-	6.5).	1305
1258	ity. When the collection of a contribution	<b>Model 1 (Claude 3.5 Sonnet)</b>	1306
1259	is regulated with sufficient precision in	Art. 32 para. 1 let. a LTAF; art. 8 and	1307
1260	formal law, the cost coverage principle	13 ECHR; international administrative	1308
1261	does not need to be examined from the	assistance in stock market matters; viola-	1309
1262	perspective of the Federal Constitution	tion of the principle of specialty; foreign	1310
1263	(para. 3.5). Art. 60a EPA does not con-	relations. The Federal Council's refusal	1311
1264	tain a cost coverage principle (para. 3.7).	to intervene with a foreign state due to	1312
1265	<b>Expert Commentary:</b> The output correctly identi-	an alleged violation of the principle of	1313
1266	fies both the reference and the theme of the judg-	specialty constitutes an act related to for-	1314
1267	ment. Its language differs from the other two and	foreign relations within the meaning of Art.	1315
1268	presents the cost coverage principle as constitu-	32 para. 1 let. a LTAF (consideration	1316
1269	tionally grounded, based on an explanation from	5). Art. 8 ECHR does not require the	1317
1270	the judgment itself. While the original headnote	requested state to intervene with the re-	1318
1271	left this question open, this version takes a clear	questing state when the latter violates,	1319
1272	position by engaging more deeply with the text.	after obtaining the information, the con-	1320
1273	The use of recitals further supports and clarifies the	ditions under which their transmission	1321
1274	headnote content.	had been subordinated. An individual	1322
1275	<b>Decision 3 (Sample ID: 60490)</b>	cannot, therefore, rely on Arts. 8 and	1323
1276	<b>Original Headnote</b>	13 ECHR to obtain judicial review of	1324
1277	Art. 83 let. a LTF; art. 32 para. 1 let. a	the Federal Council's refusal to intervene	1325
1278	LTAF; art. 8 and 13 ECHR; violation of	(consideration 6).	1326
1279	the principle of specialty in international	<b>Expert Commentary:</b> The output does not include	1327
1280	administrative assistance; standing to ap-	all the references found in the original headnote	1328
1281	peal against acts of the Federal Coun-	but identifies the theme of the judgment precisely.	1329
1282	cil concerning foreign relations; positive	It clearly states the source of the decision and sup-	1330
		ports its content with references to the relevant	1331
		recitals.	1332

## Model 2 (DeepSeek R1)

Art. 32 para. 1 let. a LTAF; art. 83 let. a LTF; art. 8 and 13 ECHR; principle of specialty in matters of international administrative assistance. The decision of the Federal Council not to intervene with a foreign state in the case of an alleged violation of the principle of specialty falls under foreign relations and constitutes a political government act, not subject to appeal unless international law guarantees a right to judicial review (paras. 5.2-5.3).

The transmission of bank data within the framework of international administrative assistance, in accordance with legal requirements (art. 38 para. 2 LBVM), does not generate a positive obligation for the requested state (Switzerland) to monitor a posteriori the respect of the principle of specialty by the requesting state (France), insofar as the alleged violation falls under the foreign territorial jurisdiction and escapes Swiss control in the sense of art. 1 ECHR (paras. 6.5.3-6.5.4).

Art. 8 ECHR does not require Switzerland to exert diplomatic pressure to correct a non-compliant use of the transmitted data, with the protection of financial data enjoying a wide margin of appreciation and positive obligations being limited to the preliminary phase of transmission (paras. 6.5.1-6.5.2).

Art. 13 ECHR does not confer an autonomous right to an effective remedy against a refusal of political intervention by the Federal Council, as Switzerland has ensured a preliminary review of the legal conditions for administrative assistance and the individual has legal remedies in the requesting state to contest the violation of the principle of specialty (para. 6.6). Conclusion: The appeal is inadmissible, as decisions related to the management of diplomatic relations fall under the political discretion of the Federal Council and do not require judicial review in the absence of binding international obligation.

**Expert Commentary:** This output identifies all the references and the theme of the judgment, closely matching the original headnote. It supports the main findings with all relevant recitals and summarizes the case contentions precisely. The inclusion of every article and the reference to recitals is accurate, and the language used is clear and appropriate.

## Model 3 (Qwen2.5 14B)

Art. 32 para. 1 let. a APAF; art. 13 ECHR; art. 8 ECHR; art. 184 para. 2 and 182 para. 2 Const.; art. 38 SESTA; art. 6 ECHR; art. 184 para. 3 Const.; art. 86 SCA; art. 32 para. 1 let. a APAF; art. 83 let. a SCA; art. 86 SCA; art. 33 let. a and b APAF; competence of the Federal Administrative Court to rule on complaints regarding international administrative assistance in stock exchange matters. The complaint against the French Republic, for violation of the principle of specialty by the French authorities in relation to international administrative assistance in stock exchange matters, pertains to foreign relations within the meaning of art. 32 para. 1 let. a APAF (consid. 5). The counter-exception of art. 32 para. 1 let. a in fine APAF, namely the right of the interested party to have his complaint judged by a court, is not applicable in this case. Indeed, art. 8 ECHR does not oblige the requested state to intervene with the requesting state when the latter, after having obtained the required information, violates the conditions under which their transmission was subject, especially when such a violation was not foreseeable (consid. 6).

**Expert Commentary:** This output includes some irrelevant references but correctly identifies the theme and states it upfront, unlike the other two versions which do not mention the complaint against the French Republic and instead focus on the role of the Federal Council. The remaining references to the recitals are supported with clear explanations.

## Decision 4 (Sample ID: 60495)

### Original Headnote

Art. 51 para. 1 Cst.; cantonal popular initiative "for a more democratic urban

1432	development in Geneva"; compatibility	5 LGZD) - does not require a constitu-	1480
1433	with superior law; introduction of a mu-	tional basis in the cantonal constitution,	1481
1434	nicipal fundamental vote in the field of	provided it concerns a clearly defined	1482
1435	neighborhood planning (plan localisé de	matter and no systematic provisions of	1483
1436	quartier [PLQ]). The creation of a mu-	the cantonal constitution oppose it (con-	1484
1437	nicipal fundamental vote in the presence	sid. 5.1-5.3).	1485
1438	of several PLQ drafts through a cantonal	According to Art. 51 para. 1 Cst., the	1486
1439	law in the formal sense does not violate	cantons are only required to have a demo-	1487
1440	Art. 51 para. 1 Cst. (E. 5.1). The intro-	cratic constitution with an elected parlia-	1488
1441	duction of this new political right without	ment and a guarantee of popular rights.	1489
1442	a basis in the cantonal constitution itself	The introduction of additional participa-	1490
1443	is compatible with Geneva constitutional	tory procedures at the legislative level	1491
1444	law (E. 5.2).	- particularly in the field of municipal	1492
1445	<b>Model 1 (Claude 3.5 Sonnet)</b>	spatial planning - does not violate the	1493
1446	Art. 51 para. 1 Cst.; Art. 34 para. 2 Cst.;	federal distribution of competences nor	1494
1447	Introduction of new political rights at the	the cantonal constitutional law, provided	1495
1448	municipal level through a formal law.	the cantonal constitution does not foresee	1496
1449	The introduction of a new political right	a *numerus clausus* of popular rights	1497
1450	at the municipal level (here: municipal	and the regulation is limited and system-	1498
1451	vote on several neighborhood develop-	consistent (consid. 5.2.1-5.2.2).	1499
1452	ment plans) can be done through a for-	The qualification of a norm as consti-	1500
1453	mal law and does not necessarily have to	tutionally relevant in the material sense	1501
1454	be enshrined in the cantonal constitution.	does not create an obligation to enshrine	1502
1455	Art. 51 para. 1 Cst. does not require that	it in the formal cantonal constitution. Par-	1503
1456	all political rights at the cantonal and mu-	ticularly for municipal political rights,	1504
1457	nicipal level be exhaustively regulated in	statutory regulation is sufficient, espe-	1505
1458	the formal cantonal constitution (E. 5.1).	cially since public participation in spa-	1506
1459	An extension of political rights by law	tial planning is provided for by federal	1507
1460	without a specific constitutional basis is	and cantonal law (Art. 4 RPG, Art. 134	1508
1461	exceptionally permissible when it con-	Cst./GE; consid. 5.2.2).	1509
1462	cerns a clearly defined area and does not	The partial invalidation of the initiative	1510
1463	contradict the will of the cantonal consti-	by the cantonal authority thus violates	1511
1464	tutional legislator (E. 5.2).	the constitutional right of initiative (Art.	1512
1465	<b>Expert Commentary:</b> The output correctly iden-	34 para. 2 Cst.), as the contested provi-	1513
1466	tifies the references and the theme. However, its	sion is compatible with superior law.	1514
1467	explanation, especially regarding the recital and	*(BGE 149 I 33; references to ATF 104	1515
1468	Art. 51, does not align with the original judgment.	Ia 343, 142 I 216; HANGARTNER/K-	1516
1469	The language is also more complex than necessary,	LEY, GRISEL)*	1517
1470	which may reduce clarity.	<b>Expert Commentary:</b> This output includes addi-	1518
1471	<b>Model 2 (DeepSeek R1)</b>	tional references and demonstrates an understand-	1519
1472	**Art. 34 para. 2 Cst.; Art. 51 para. 1	ing of the judgment's context. However, it elabo-	1520
1473	Cst.; Art. 5A para. 5 LGZD (GE); In-	rates on aspects that are not relevant to a headnote.	1521
1474	troduction of new political rights at the	It does not rely on exact references and adds exces-	1522
1475	municipal level through ordinary law.**	sive detail that goes beyond what is appropriate for	1523
1476	The creation of a new political right at	this format.	1524
1477	the municipal level - here a binding vote	<b>Model 3 (Qwen2.5 14B)</b>	1525
1478	to choose between competing neighbor-	Art. 51 para. 1 Cst; Art. 34 para. 2	1526
1479	hood development plans (Art. 5A para.	Cst; Art. 44 para. 1 and Art. 60 para.	1527

1528	4 Cst/GE; Inclusion of a municipal vot-	4.4.1). In journalistic research and inter-	1576
1529	ing right in a formal cantonal legislative	views, inadequate information provided	1577
1530	proposal. The introduction of a municip-	to those involved about the context and	1578
1531	al voting right on several neighborhood	the dissemination channel is only rele-	1579
1532	plan projects can be incorporated into	vant if it leads to a violation of Art. 4	1580
1533	a formal cantonal legislative proposal	para. 2 RTVG, i.e., if it impairs the audi-	1581
1534	in the area of spatial planning without	ence's ability to form their own opinion	1582
1535	violating Art. 51 para. 1 Cst (E. 5.1).	(E. 4.4.2). Investigative journalism with	1583
1536	The constitutional law of the canton of	a critical attitude is also permissible as	1584
1537	Geneva does not contain a general norm	long as the program as a whole does not	1585
1538	stipulating that material constitutional	appear manipulative and allows the audi-	1586
1539	norms must necessarily be included in	ence to form their own opinion (E. 3.4).	1587
1540	the constitution in the formal sense (E.		
1541	5.2).		
1542	<b>Expert Commentary:</b> The output identifies the	<b>Expert Commentary:</b> The output provides insuf-	1588
1543	references but includes some inaccuracies, such as	ficient references and fails to identify the primary	1589
1544	citing Art. 44 instead of p. 40 n. 44. Despite this, it	theme of the case. Instead, it refers to several other	1590
1545	is more concise than the other outputs and correctly	aspects that are part of the judgment but are less	1591
1546	captures the theme of the case.	central to the decision.	1592
1547	<b>Decision 5 (Sample ID: 60555)</b>	<b>Model 2 (DeepSeek R1)</b>	1593
1548	<b>Original Headnote</b>		
1549	Art. 17, 29 para. 1 and 93 para. 2	Art. 17 and 93 Swiss Constitution; Art.	1594
1550	Cst; Art. 4 RTVG; television and radio	4 of the Swiss Federal Act on Radio and	1595
1551	broadcasts about deaths in Ticino nursing	Television (RTVG); journalistic duty of	1596
1552	homes during the first wave of the pan-	care in reporting on COVID-19 deaths	1597
1553	demic; investigative journalism. Scope	in nursing homes. The Federal Supreme	1598
1554	and content of media freedom protected	Court confirms that critical reporting on	1599
1555	by Art. 17 and 93 para. 2 Cst (E. 3.1).	pandemic management in nursing homes	1600
1556	Minimum requirements according to Art.	does not violate program law require-	1601
1557	4 para. 1 and 2 RTVG for an editorial	ments, as long as the principles of fac-	1602
1558	program with informational content (E.	tual accuracy and transparency are main-	1603
1559	3.2 and 3.3). Scope and limits of inves-	tained (E. 3.3, 5.2.5). The freedom of	1604
1560	tigative journalism (E. 3.4). Review au-	the media according to Art. 17 of the	1605
1561	thority of the UBI (E. 4.3 and 4.4). The	Swiss Constitution also includes inves-	1606
1562	contested television and radio broadcasts	tigative journalism, provided that the pre-	1607
1563	in the present case did not violate Art. 4	sentation is not manipulative and allows	1608
1564	RTVG (E. 5).	the audience to form independent judg-	1609
1565	<b>Model 1 (Claude 3.5 Sonnet)</b>	ments (E. 3.4, 5.2.4). There is no vio-	1610
1566	Art. 4 para. 2 RTVG; Art. 97 para. 2 lit.	lation of Art. 4 para. 2 of the RTVG	1611
1567	a RTVG; Requirement of impartiality in	if emotive elements (e.g., music, inter-	1612
1568	radio and television broadcasts; Limits	view conduct) can be considered appro-	1613
1569	of the UBI's examination authority.	appropriate in the context of a complex and	1614
1570	The UBI exclusively examines the con-	already publicly discussed topic, and the	1615
1571	tent of already broadcast editorial pro-	responsible parties involved were able	1616
1572	grams and not their creation process. The	to comprehensively express their views	1617
1573	examination is conducted in the public	(E. 5.1.3, 5.2.6). The allegation of a	1618
1574	interest and not for the protection of the	formal denial of justice (Art. 29 of the	1619
1575	personality rights of those involved (E.	Swiss Constitution) fails because the su-	1620
		pervisory complaint with the Indepen-	1621
		dent Complaints Authority for Radio and	1622
		Television (AIRR) according to Art. 97	1623
		RTVG only examines the final program	1624
		content and not internal editorial agree-	1625

1626	ments (E. 4.4.1). The dignity of those	The principle of correctness of fact pre-	1676
1627	affected (Art. 7 of the Swiss Constitu-	sensation according to Art. 4 para. 2	1677
1628	tion) is preserved, as the reporting did	RTVG has not been violated (E. 5).	1678
1629	not equate to a "trial by media" and did		
1630	not distort the core statements of the in-	<b>Expert Commentary:</b> The output correctly identi-	1679
1631	terviewees (E. 5.2.5). The jurisprudence	fies the references and the theme at the beginning.	1680
1632	on balance (cf. BGE 137 I 340) remains	However, it introduces the principle of correctness,	1681
1633	unaffected: A quantitative equal distri-	which does not appear in the original excerpt of the	1682
1634	bution of perspectives is not required as	judgment.	1683
1635	long as divergent viewpoints remain rec-		
1636	ognizable (E. 3.3, 5.2.4).	<b>Decision 6 (Sample ID: 60782)</b>	1684
		<b>Original Headnote</b>	1685
1637	<b>Expert Commentary:</b> The references are correctly		
1638	identified, and the primary theme of the case is	Art. 276 and 285 CC; child support; dis-	1686
1639	captured more accurately than in the first output.	tribution of excess in the case of children	1687
1640	However, the headnote includes too much detail,	of unmarried parents. Summary of the	1688
1641	which makes it overly bulky for its purpose.	jurisprudence on the concrete two-stage	1689
		method with distribution of the excess	1690
1642	<b>Model 3 (Qwen2.5 14B)</b>	(para. 2.4-2.6). If only one parent is re-	1691
1643	Art. 4 para. 2 and Art. 97 para. 2 lit. a	quired to pay the support contribution,	1692
1644	RTVG; Art. 29 para. 1 BV; Transmis-	the calculation is made between this par-	1693
1645	sion of information about the conditions	ent and the beneficiary children of the	1694
1646	in a nursing home during the first pan-	contribution. An excess that remains af-	1695
1647	demic wave; Examination of the correct-	ter covering the family law subsistence	1696
1648	ness of the presentation of facts; Proce-	minimum is distributed (in the ratio of	1697
1649	dure. The appellate instance can only	two to one) between this parent and the	1698
1650	resort to the subsequent review of the	children (para. 2.7).	1699
1651	content of broadcasts to examine com-		
1652	pliance with the provisions on the con-	<b>Model 1 (Claude 3.5 Sonnet)</b>	1700
1653	tent of editorial broadcasts. The prepara-		
1654	tions and discussions between journalists	Art. 285 para. 1 CC; child maintenance;	1701
1655	and those affected, which preceded, fall	distribution of surplus between unmar-	1702
1656	outside its scope of examination (E. 4).	ried parents. In the case of unmarried par-	1703
1657	The review of compliance with the pro-	ents, the calculation of the maintenance	1704
1658	visions on the content of editorial broad-	contribution according to the two-phase	1705
1659	casts is not aimed at protecting the per-	concrete method with surplus distribu-	1706
1660	sonal rights of those who participated in	tion occurs only between the obligated	1707
1661	the broadcast. The question of whether	parent and the child. A virtual share of	1708
1662	the broadcast, aired as part of an "investi-	the surplus cannot be attributed to the	1709
1663	gation," presents the position established	unmarried custodial parent who is not	1710
1664	by the journalists without depicting it as	entitled to maintenance and is not autho-	1711
1665	an absolute truth must be assessed based	rized to actually participate in the surplus	1712
1666	on the specific circumstances. In this	of the other parent. The distribution of	1713
1667	context, the severity of the issue and the	the surplus must be limited to the persons	1714
1668	significance of the criticism must also	actually involved in the maintenance re-	1715
1669	be considered. In the present matter, the	lationship (consideration 2.7).	1716
1670	broadcast did not present the journalists'		
1671	established position as the only truth,	<b>Expert Commentary:</b> The output does not iden-	1717
1672	and those affected, who participated in	tify all relevant references but correctly states the	1718
1673	the broadcast, were able to engage with	theme of the case. It presents the content concisely	1719
1674	the accusations made against them. The	and with appropriate reference, demonstrating a	1720
1675	broadcast was therefore not manipulative.	good understanding of the judgment.	1721

1722	<b>Model 2 (DeepSeek R1)</b>		
1723	Art. 285 CC; concrete two-phase method	any other possible justified distribution	1771
1724	with excess distribution. In calculating	of the profit for the specific case (consid.	1772
1725	child support for children of unmarried	2.1-2.5). If the parents are not married	1773
1726	parents, for the purpose of distributing	and the child is under the care of one	1774
1727	the excess according to the principle of	of them, the profit must be shared be-	1775
1728	"big and small heads," an additional "big	tween the parent obliged to contribute	1776
1729	head" for the parent not obliged to make	and the child, without considering the	1777
1730	monetary contributions is not to be con-	non-contributing parent as a "big head"	1778
1731	sidered virtually (cons. 2.7). The ex-	(consid. 2.6 and 2.7).	1779
1732	cess is to be distributed exclusively be-	<b>Expert Commentary:</b> The output only partially	1780
1733	tween the child ("small head") and the	identifies the relevant reference. It focuses on child	1781
1734	obligated parent ("big head"), in the pro-	support in relation to education, while the judgment	1782
1735	portion of 1/3 and 2/3, without fictitious	addresses child support in the context of caregiving	1783
1736	deductions in favor of the unmarried cus-	for children of unmarried parents. The language	1784
1737	todial parent, who is not entitled to an	used is more complex than necessary and does not	1785
1738	excess. This ensures a fair assessment	reflect the clarity of the original judgment.	1786
1739	of the obligated parent's ability to con-		
1740	tribute, in accordance with the dynamic		
1741	nature of maintenance under Art. 285		
1742	CC, avoiding artificial reductions to the		
1743	detriment of the child's rights (reference		
1744	to BGE 147 III 265). The claimed auto-		
1745	matic equalization with the children of		
1746	married parents, through the insertion of		
1747	an additional "big head," is contrary to		
1748	the principles of the concrete method and		
1749	the rationale of the child maintenance re-		
1750	vision.		
1751	<b>Expert Commentary:</b> The output correctly iden-		
1752	tifies the reference and shows an understanding of		
1753	the case at the outset. However, it misrepresents the		
1754	core of the calculation discussed in the judgment		
1755	and uses overly complex language. The judgment		
1756	outlines a two-step calculation for maintenance,		
1757	which is not clearly reflected here. Legally, this		
1758	makes the headnote unnecessarily complicated and		
1759	suggests a potential misunderstanding of the deci-		
1760	sion.		
1761	<b>Model 3 (Qwen2.5 14B)</b>		
1762	Art. 276 et seq. CC; calculation of the		
1763	maintenance contribution for the educa-		
1764	tion of children of unmarried parents. If,		
1765	in applying the profit-sharing calculation		
1766	method, a profit occurs during the mi-		
1767	nority of the child, it must be shared be-		
1768	tween the parent obliged to contribute		
1769	and the child according to the principle		
1770	of "big and small head," or according to		