CourtPressGER

Anonymous ACL submission

Abstract

Official court press releases from Germany's highest courts are vital for bridging complex judicial rulings and the public. Prior efforts on German legal text summarization in NLP emphasize technical headnotes, often ignoring the need for citizen oriented communication. We introduce CourtPressGER, a 6.4k triple dataset of rulings, their human-drafted press releases, and synthetic contextual generation prompts for LLMs to generate comparable press releases. The resulting benchmark dataset is intended to train and evaluate LLMs in generating accurate, more readable summaries from long judicial texts. We benchmark a set of small and large LLMs on the task and evaluate model outputs via reference-based metrics, factual-consistency checks, and an LLM-as-judge approach that approximates expert review. We further conduct qualitative expert analysis and ranking. Results show that large LLMs produce near-human-quality drafts and only marginally lose performance when applied hierarchically. Smaller models require a hierarchical setup to be able to summarize long judgments, and achieve a range of scores. All models struggle with factual consistency, and the human drafted press release is consistently ranked highest.

Introduction

High-level German courts strive to make their decisions accessible to the public through press releases that summarize the essential aspects and implications of the decisions in an understandable form. Releases are directly authored by judges and contain both legal authority as well as a lay-friendly narrative, serving as an important interface between the judiciary and the general public. Distilling a case in this way also is one form of targeted legal case summarization, a legal NLP use case for which manually created gold data is typically sparse and expensive to create. resources. Rapid progress in the capabilities of LLMs suggests that high-quality automatic drafts are within reach, yet robust evaluations of legal decision summary quality are difficult, especially in languages other than English. CourtPressGER advances the state of German legal document summarization by:

1. Collecting the largest aligned corpus of German decisions and press releases to date, comprising 6.4k pairs,
2. deriving decision-specific summarization prompts,
3. benchmarking a range of open and commercial LLMs across two size groups, and
4. analyzing their performance through complementary automatic measures and manual expert assessment

Related Work

Legal-text summarization has progressed from early sentence-ranking heuristics borrowed from news [Grover et al., 2004, Polsley et al., 2016] to domain-adapted encoder–decoder transformers such as *Legal-BART* and *Legal-PEGASUS* [Chalkidis and Kampas, 2019, Zhang et al., 2020, Aumiller et al., 2022]. Recent surveys report steady ROUGE gains but emphasize three persistent challenges—extreme document length, jurisdiction-specific jargon, and the absence of factual-consistency metrics [Kanapala et al., 2019, Akter et al., 2025]. Researchers address the length issue with hierarchical encoders and chunk-merge strategies for book-length opinions [Chang et al., 2024] and Indian Supreme Court cases [Deroy et al., 2024b]. Yet expert evaluations reveal that higher ROUGE scores do not necessarily align with legal usefulness [Steffes et al., 2023], underscoring the need for multi-faceted assessment.

Datasets. Larger legal summarization corpora typically pair expert-targeted summaries with their underlying document. A distinction can be made between, first, summarizing legislation, such as BillSum (U.S. bills; Kornilova and Eidelman, 2019) and EUR-Lex-Sum (EU legislation; Aumiller et al., 2022) and, second, case/judgment summarization, such as the US-English Multi-LexSum (U.S. civil-rights cases; Shen et al., 2022, including e.g., complaints and motions) and Portuguese BrazilianBR (STF rulings; Feijo and Moreira, 2023). For German, *LegalSum* covers ∼100k rulings with legal-holding-focused *Leitsätze* [Glaser et al., 2021] from the German legal context, and Rolshoven et al. [2024] provides 57k *Regesten* from Swiss courts. Both target legal practitioners and their summaries consist of concise, technical, often extractive headnotes. To date, no corpus of significant size aligns German decisions with non-headnote-based summaries written for mixed audiences, including journalists and the general public.

Outside Germany, few resources focus on citizen-oriented summaries, e.g., TL;DR software license synopses [Manor and Li, 2019], Canadian lay summaries [Salaün et al., 2022], and argument-aware rewriting [Elaraby and Litman, 2022]. The German ALeKS project seeks to automate headnote generation but releases no code or data and again targets experts, leaving a gap in NLP benchmarks for court-citizen communication.

Extractive approaches to summarization have been mostly succeeded by abstractive summarization using transformers [Shukla et al., 2022, Moro and Ragazzi, 2022, Santosh et al., 2025, 2024b] and faithfulness-enhancing rerankers [Feijo and Moreira, 2023]. Cross-jurisdiction transfer of smaller models primarily trained in one jurisdiction poses a challenge Santosh et al. [2024a]. As as known, large commercial state of the art models are marketed as capable of summarizing judgments well, and the aims of this paper include qualitative expert analysis of this proposition on our press release dataset.

Evaluation. In addition to classic NLP metrics, we see newer factual metrics like QAGS (Question Answering for evaluating Generated Summaries) (Wang et al. [2020]) and FactCC (Factual Consistency Check) (Kryściński et al. [2019]). QAGS (Question Answering for evaluating Generated Summaries) generates questions from one text and then compares the answers to verify factual correctness. FactCC extracts claims from the one text and checks them against another body. A total factual consistency score is computed from these checks.

Also working on German court press releases, Steffes et al. [2023] demonstrate that ROUGE scores alone fails to capture whether legally salient content is present in a summary. Alternative protocols generate question–answer pairs from the reference or enlist large-language-models as judges, both correlating better with expert panels [Xu et al., 2021]. Current research still lacks (i) German press-release data, (ii) long-context benchmarks in German legal datasets, and (iii) holistic evaluation beyond ROUGE.

Our contribution. **CourtPressGER** addresses these gaps by releasing 6.4k aligned triplets of federal-court decisions, their official press releases, and a synthetic contextual prompt describing the structure of the press release to an LLM. We benchmark six open-source and commercial LLMs using overlap-, embedding- and entailment-based metrics, and validate automatic scores against expert spot-checks. The corpus and baselines establish the first citizen-oriented benchmark for German judicial communication. By supplying a public dataset and a multi-dimensional evaluation suite centered on citizen-oriented summaries, CourtPressGER complements prior resources focused on technical headnotes and narrow evaluation settings, opening a new avenue for research on transparent court communication.

CourtPressGER

Data

Our dataset includes 6.432 court decisions and corresponding press releases from Germany's highest courts from the years 1995 to 2023: Federal Labor Law Court (Bundesarbeitsgericht - BAG), Federal Fiscal Court (Bundesfinanzhof - BFH), Federal Court of Justice (Bundesgerichtshof - BGH), Federal Social Court (Bundessozialgericht - BSG), Federal Constitutional Court (Bundesverfassungsgericht- BVerfG) and the Federal Administrative Court (Bundesverwaltungsgericht - BVerwG).

Splits & Statistics

For our experiments, we divided the dataset into training, validation, and test splits in an 72.2/11.6/16.3 ratio. The training set contains 4643 pairs, while the validation set contains 744 test sets contain 1045 pairs. We decided to split chronologically because otherwise the distribution shifts incurred by rotating press office personnel over time would not be captured in the data split, leading to a potential overestimation of performance on unseen data. Descriptive statistics of the cleaned dataset can be seen in Table 1. Using EuroBERT tokenizer, we see an average length of decisions of 10.810 tokens and press releases with an average 1.402 tokens with high standard deviations of 10.739 tokens for judgements and 955 tokens for press releases.

	Press Release			Judgment		
Court	Mean	Std	Count	Mean	Std	Count
BAG	1056.37	407.50	177	14148.00	7913.64	177
BFH	800.28	213.58	761	7378.97	4410.79	761
BGH	1386.84	680.10	2407	8216.82	5686.26	2407
BSG	1146.66	484.69	161	11790.02	4850.29	161
BVerfG	2039.50	1353.63	1771	14781.53	16844.62	1771
BVerwG	942.91	336.86	1155	11734.63	8110.92	1155
Overall avg	1402.32	954.52	–	10809.58	10739.27	–

Table 1: Press releases and judgments statistics by court

Experimental Setup

Synthetic Prompts

For each decision-press release pair, we generated synthetic prompts through the Anthropic API (Claude Sonnet 3.7) to serve as input for LLMs to generate press releases. These prompts are intended to enable LLMs, when benchmarked, to generate precise text that match the reference press release in length, intended audience, and topical coverage. This way, a comparison is more meaningful than if a single generic generation prompt is used across all datapoints, which stem from many different authors.

Press Release Generation

Our pipeline then sends the synthetic prompts and the case to the models, collect the generated press releases, and stores them alongside the actual press releases. The

models we leverage for this generation are GPT-4o (mainstream and economical closed source model at time of experiments), Llama-3-70B (large & SotA open weights model at time of running experiments), Teuken-7B, Llama-3-8B, EuroLLM-9B, Mistral-7B (all open weights in smaller class, typical base models for research finetuning experiments)

Context Limitation

As expected, the context window size of the models limits their ability to generate high-quality press releases. Models with larger context windows (e.g., GPT-4o with a theoretical limit of 128k tokens, though in our implementation we used the API with a practical limit of 64k tokens) can process the entire court decision at once, while smaller models require document chunking and hierarchical summarization approaches. For decisions that exceed the context window of a model, we implemented a hierarchical summarization approach that allows the model to consider the entire document while respecting context limitations.

Hierarchical Summarization

Because lengthy German court rulings often exceed LLM context windows, we employ a hierarchical summarization pipeline inspired by Chang et al. [2024] to progressively condense text rather than relying on a single pass. While incremental chunk-by-chunk updating preserves detail, it can induce coherence errors for large documents [Chang et al., 2024]. In contrast, hierarchical merging systematically integrates partial synopses, yielding more coherent summaries.

A key advantage of our multi-level procedure is its *abstractive* nature. Following Deroy et al. [2024a], we note that generative legal-domain models outperform extractive baselines on metrics such as ROUGE, METEOR, and BERTScore for Court cases. It condenses complex legal reasoning into more readable, information-dense summaries. By generating chunk-wise abstractive synopses, we want to preserve legally salient entities and mitigate hallucination risks [Deroy et al., 2024a] through controlled chunk sizes and cross-chunk aggregation.

Practically, we split each decision into paragraphs, then merge them until reaching a token threshold compatible with the chosen LLM. Each chunk undergoes a "Level-0" summary to capture key arguments, reasoning, and legal points, after which partial summaries are recursively merged in a structured tree-like fashion.

Instruction-Tuned TEUKEN-7B: Joint Hierarchical Summarization and Press-Release Generation

Anticipating a negative impact of hierarchical summarization over single-pass summarization, we want to explore how far a smaller model can be improved when specifically trained for this task in a hierarchical fashion, as we have a training set available. We instruction-tune the open-source TEUKEN-7B model via a *two-stage* approach: first generating faithful level-by-level summaries of long rulings and then rewriting the highest-level summary into a press release.

Stage 1: Hierarchical summarisation. For every paragraph chunk c_i of a decision (max. token budget $4\,096$), we provide an *explicit instruction* indicating the desired abstraction level. Reference summaries are produced with `Llama-3-70B-Instruct`, yielding **35,k** chunk–instruction pairs. We adopt the *stacked* SFT scheme of Pareja et al. [2024], training TEUKEN-7B on all hierarchy levels simultaneously, which generalizes better than sequential tuning.

Stage 2: Press-release generation. We retain each final hierarchical summary s^{final} and combine it with the synthetic prompt Section and the gold press release $p^{\star}$. The model is then further trained with a constant learning-rate schedule to map the concatenated input $(\text{prompt} + s^{\text{final}})$ to $p^{\star}$.

Training set-up. Unless stated otherwise, we follow Pareja et al. [2024], reducing the effective batch size to 512 and training for four epochs over the mixed corpus, with early stopping on validation perplexity.

Evaluation

We developed a comprehensive evaluation approach using multiple complementary metrics: ROUGE (Lin [2004]), BLEU (Papineni et al. [2002]), METEOR (Banerjee and Lavie [2005]), BERTScore (Zhang et al. [2020]), QAGS (Question Answering for evaluating Generated Summaries) (Wang et al. [2020]), FactCC (Factual Consistency Check) (Kryściński et al. [2019]) and LLM-as-a-Judge.

Factual Consistency Metrics

Because press releases can include context not explicitly stated in the court decisions, QAGS and FactCC metrics may flag such information as inconsistent, potentially lowering scores for otherwise high-quality press releases. We partially address this through our LLM-as-a-Judge approach and the human evaluation process, which better distinguishes contradictory information from benign additional context.

LLM-as-a-judge

We use Claude 3.7 Sonnet to evaluate the generated press releases based on completeness, clarity, structure and comparison to the reference. The metric provides numerical ratings (1-10) and calculate an overall score across all evaluation criteria. The model was selected for this task due to its strong performance in understanding complex legal texts in multiple languages as well as its selection for synthetic prompt generation which made it a natural choice for evaluation.

Human Evaluation

In addition to above metrics, we had two cases per court per model reviewed by two human annotators, includ-

Model	R1	B1	MTR	BERT	FCC	QAGS	LJ_Fact	LJ_Compl	LJ_Clar	LJ_Struc	LJ_Ref	LJ_Tot	Human Avg
gpt 4o_hier	0.3584	0.2275	0.1836	0.7711	0.4915	0.2637	**8.1070**	**7.0885**	**8.7451**	**8.4076**	**6.8414**	**7.8379**	**4.5**
llama 3_3 70B_hier	**0.3746**	**0.2327**	**0.1931**	0.7730	0.4987	**0.2863**	7.3417	6.3637	8.1545	7.6200	5.9002	7.0760	4.714
eurollm 9B_hier	0.2800	0.1856	0.1451	0.7459	0.5065	0.1875	4.9739	4.4255	6.4043	6.6876	3.5435	5.2070	7.143
llama 3 8B_hier	0.2927	0.1829	0.1472	0.7373	0.5082	0.2289	5.2780	4.5405	6.3069	6.4295	3.7751	5.2660	6.429
mistral v03_hier	0.3571	0.2304	0.1871	**0.7777**	**0.5122**	0.2386	5.5376	4.9653	5.5578	5.2447	3.7370	5.0085	7.714
teuken_hier	0.1630	0.0794	0.0781	0.6600	0.5051	0.1607	3.0635	2.1606	4.2356	4.4077	1.8269	3.1388	10.214
teuken_inst_sft_hier	0.2865	0.2181	0.1758	0.8	0.4893	0.12	3.0758	2.0058	4.1411	4.9223	1.5835	3.1457	9.857
gpt 4o_full	0.3627	0.2105	0.1845	0.7563	0.4991	0.2777	**8.3933**	**7.1615**	**8.8192**	**8.5385**	**7.0115**	**7.9848**	**4.0**
llama 3_3 70B_full	**0.3823**	**0.2248**	**0.1986**	**0.7691**	**0.5082**	0.2898	8.1721	6.8661	8.6333	8.1552	6.6603	7.6974	4.071
mistral v03_full	0.3612	0.2126	0.1901	0.7465	0.5021	**0.3252**	6.9612	5.7141	7.1395	6.8110	5.0271	6.3306	5.857

Table 2: Press release comparison on hierarchical and full judgements

ing a licensed German attorney, who blindly ranked the generated press releases as well as the reference press release according to preference from 1 (best) to 11 (worst).

The annotators also provided feedback on the narrative coherence and usability (i.e. whether it is close to publishable) of the generated press release, and whether the it contains extraneous information that was not contained in the judgment. The latter not only covers hallucinations of LLMs, but also factually correct enriched information. Note that even the reference press release regularly contains dates, case numbers and the social context of the case that might not be represented in the document itself.

Results

Based on our evaluation, we present the results organized by evaluation type (hierarchical vs. full document processing) and model in Table 2. We structured our analysis to examine reference-based metrics, embedding-based metrics, factual consistency metrics, evaluation through LLM-as-judge and human scores.

The full-text condition reveals the upper bound a model can reach when context is not truncated, whereas the hierarchical setting approximates a local-deployment scenario. GPT-4o and Llama-3-70B are statistically tied on most automatic metrics, yet human-style LLM judging clearly prefers GPT-4o.

Note that we evaluated Mistral_v03 also on the full ruling text even though it's context is limited to 32k tokens. In our experiments, 1% of documents needed to be truncated for evaluation in this narrower context, which we find to be negligible noise.

Discussion

These results are consistent with findings from Glaser et al. [2021], who reported ROUGE-1 scores of around 30.5% for their best models on German court decision summarization. Our best models exceed this performance slightly, which may be attributed to the advancement in LLMs since their study.

Our findings confirm the intuitive trade-off between model capacity and inference cost: large models (*GPT 4o, Llama 3 70B*) heavily outperform smaller ones on fidelity, completeness and clarity, but the differential shrinks when hierarchical summarisation is used.

LLM-as-a-judge results align excellently with expert feedback collected, sending a strong signal for applicability of this evaluation method on legal case summarization. While human annotators still rank the reference press release highest (as they are still easily spotted by experts), generated press releases from larger models show high practical usability with minor edits.

These results demonstrate that while larger models generally produce press releases that are more factually correct, complete, clear, and well-structured, the hierarchical summarization approach allows smaller models to produce reasonably good summaries, particularly in terms of clarity and structure. Interestingly, the improvement from hierarchical summarisation to full summarisation is marginal for the largest models.

While our fine-tuned Teuken model showed some improvement over the base, it still performs far below larger models, suggesting that parameter count remains a decisive factor for this complex task.

Conclusions

Our comprehensive evaluation of the CourtPressGER system demonstrates that modern LLMs can effectively generate German court press releases, with performance varying according to model size and architecture. Key findings include:

Model size matters: Larger models consistently outperform smaller models across all evaluation metrics.

Hierarchical summarization is effective: Our hierarchical approach enables smaller models to process long documents while maintaining reasonable quality.

Factual consistency challenges: Even the best models struggle with perfect factual consistency, indicating room for improvement.

Language-specific models: European-language-specific models like EuroLLM can show competitive performance for their size compared to larger multilingual models.

Limitations

We found several limitations of our approach:

1. Evaluation metrics: Our use of QAGS and FactCC metrics, which were developed and validated on English datasets, introduces uncertainty when applied to German legal texts. Future work should explore German-specific factual consistency metrics.
2. LLM-as-judge vs. human evaluation: While our LLM-based evaluation provides valuable insights, it serves as a proxy for human expert evaluation and would benefit from further validation through targeted expert reviews.
3. Additional context in press releases: Court press releases often contain contextual information not present in the original decision, which can confound factual consistency metrics.
4. Divergence from Rolshoven et al. findings: Unlike [Rolshoven et al., 2024], who found that fine-tuned smaller models could approach the performance of larger models, our results show a clear advantage for larger models. This difference may be attributed to our focus on press releases rather than technical summaries ("Regesten"), the different nature of our dataset, or the specific characteristics of German federal court decisions.

Ethics Statement

All data originate from publicly available court websites. Personal names are already anonymised by the courts. Our dataset is released excluding any confidential meta-data.

References

Mousumi Akter, Erion Cano, Erik Weber, Dennis Dobler, and Ivan Habernal. 2025. A comprehensive survey on legal summarization: Challenges and future directions. *arXiv preprint arXiv:2501.17830*.

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*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.

Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. *Preprint*, arXiv:2310.00785.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024a. Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization. *Preprint*, arXiv:2407.12848.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024b. Ensemble methods for improving extractive summarization of legal case judgements. *Artificial Intelligence and Law*, 32(1):231–289.

Mohamed Elaraby and Diane Litman. 2022. Argle-galsumm: Improving abstractive summarization of legal documents with argument mining. *Preprint*, arXiv:2209.01650.

Diego de Vargas Feijo and Viviane P Moreira. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1):91–113.

Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021. Summarization of German Court Rulings. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 180–189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claire Grover, Ben Hachey, Ian Hughson, and Buccleuch Place. 2004. The holj corpus: Supporting summarisation of legal texts. In *In Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the Factual Consistency of Abstractive Text Summarization. *Preprint*, arXiv:1910.12840.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries.

Laura Manor and Junyi Jessy Li. 2019. Plain english summarization of contracts. *arXiv preprint arXiv:1906.00424*.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11085–11093.

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

Aldo Pareja, Nikhil Shivakumar Nayak, Hao Wang, Krishnateja Killamsetty, Shivchander Sudalairaj, Wenlong Zhao, Seungwook Han, Abhishek Bhandwaldar, Guangxuan Xu, Kai Xu, Ligong Han, Luke Inglis,

and Akash Srivastava. 2024. Unveiling the Secret Recipe: A Guide For Supervised Fine-Tuning Small LLMs. *Preprint*, arXiv:2412.13337.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

Luca Rolshoven, Vishvaksenan Rasiah, Srinanda Brügger Bose, Matthias Stürmer, and Joel Niklaus. 2024. Unlocking Legal Knowledge: A Multilingual Dataset for Judicial Summarization in Switzerland. *Preprint*, arXiv:2410.13456.

Olivier Salaün, Aurore Troussel, Sylvain Longhais, Hannes Westermann, Philippe Langlais, and Karim Benyekhlef. 2022. Conditional abstractive summarization of court decisions for laymen and insights from human evaluation. In *Legal Knowledge and Information Systems*, pages 123–132. IOS Press.

T. Y. S. S. Santosh, Youssef Farag, and Matthias Grabmair. 2025. Coperlex: Content planning with event-based representations for legal case summarization. *Preprint*, arXiv:2501.14112.

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

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

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*.

Bianca Steffes, Piotr Rataj, Luise Burger, and Lukas Roth. 2023. On evaluating legal summaries with rouge. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 457–461.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. *Preprint*, arXiv:2004.04228.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254.

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

Appendix

AI usage

We leveraged Claude Sonnet 3.7 for coding tasks and GPT-4o for wording, shortening and Latex tasks.

Prompts

We used the following prompts for our experiments:

Synthetic prompt generation

We used the following prompt for synthetic prompt generation:

Synthetic prompt generation

Du bist ein Experte für juristische Texte und Kommunikation. Deine Aufgabe ist es, ein Gerichtsurteil und die dazugehörige Pressemitteilung zu analysieren und dann herauszufinden, welcher Prompt verwendet worden sein könnte, um diese Pressemitteilung aus dem Gerichtsurteil zu generieren, wenn man ihn einem LLM gegeben hätte.

1. Analysiere, wie die Pressemitteilung Informationen aus dem Urteil vereinfacht, umstrukturiert und Schlüsselinformationen hervorhebt
2. Berücksichtige den Ton, die Struktur und den Detaillierungsgrad der Pressemitteilung
3. Identifiziere, welche Anweisungen nötig wären, um den juristischen Text in diese Pressemitteilung zu transformieren

Erkläre NICHT deine Überlegungen und füge KEINE Meta-Kommentare hinzu. Gib NUR den tatsächlichen Prompt aus, der die Pressemitteilung aus dem Gerichtsurteil generieren würde. Sei spezifisch und detailliert in deinem synthetisierten Prompt.
Hier ist das originale Gerichtsurteil:
{court_ruling}
Und hier ist die Pressemitteilung, die daraus erstellt wurde:
{press_release}
Erstelle einen detaillierten Prompt, der einem LLM gegeben werden könnte, um die obige Pressemitteilung aus dem Gerichtsurteil zu generieren. Schreibe NUR den Prompt selbst, ohne Erklärungen oder Meta-Kommentare.

Press release generation

We used the following prompt for press release generation:

Press release generation

{prompt}
Gerichtsurteil:
{ruling}

LLM-as-a-judge

We used the following prompt for LLM-as-a-judge evaluation:

LLM-as-a-judge

Du bist ein Experte für juristische Texte und bewertst die Qualität von Pressemitteilungen für Gerichtsurteile. Bewerte die generierte Pressemitteilung anhand der folgenden Kriterien auf einer Skala von 1-10:
1. Faktische Korrektheit: Wie genau spiegelt die Pressemitteilung die Fakten aus dem Gerichtsurteil wider?
2. Vollständigkeit: Wurden alle wichtigen Informationen aus dem Urteil in der Pressemitteilung berücksichtigt?
3. Klarheit: Wie verständlich ist die Pressemitteilung für ein nicht-juristisches Publikum?
4. Struktur: Wie gut ist die Pressemitteilung strukturiert und organisiert?
5. Vergleich mit Referenz: Wie gut ist die generierte Pressemitteilung im Vergleich zur Referenz-Pressemitteilung?
Gib für jedes Kriterium einen numerischen Wert zwischen 1 und 10 an und eine kurze Begründung. Berechne abschließend einen Gesamtscore als Durchschnitt aller Einzelwerte. Gib deine Antwort im folgenden JSON-Format zurück:
{
"faktische_korrektheit": {
"wert": X, "begründung": "..."
},
"vollständigkeit": {
"wert": X, "begründung": "..."
}, "klarheit": {
"wert": X, "begründung": "..."
}, "struktur": {
"wert": X, "begründung": "..."
},
"vergleich_mit_referenz": {
"wert": X, "begründung": "..."
},
"gesamtscore": X.X
}
# Gerichtsurteil
{source}
Generierte Pressemitteilung
{generated}
Referenz-Pressemitteilung
{reference}

Model	R1	R2	RL	B1	B2	B3	B4	MTR	BP	BR	BF1	KW	ENT	Len	Fcc	FccC	QGS	Qn	LJ_Fact	LJ_Compl	LJ_Clar	LJ_Struc	LJ_Ref	LJ_Tot
openai_gpt_4o_full	0.3627	0.1452	0.1918	0.2105	0.1266	0.0832	0.0559	0.1845	0.7746	0.7396	0.7563	0.2082	0.2290	0.4572	0.4991	0.5068	0.2777	4.75	**8.3933**	**7.1615**	**8.8192**	**8.5385**	**7.0115**	**7.9848**
openai_gpt_4o_hier	0.3584	0.1242	0.1758	0.2275	0.1280	0.0786	0.0495	0.1836	0.7835	0.7595	0.7711	0.1883	0.2157	0.5114	0.4915	0.4758	0.2637	4.78	8.1070	7.0885	8.7451	8.4076	6.8414	7.8379
llama_3_3_70B_full	**0.3823**	**0.1601**	**0.1997**	0.2248	**0.1385**	**0.0946**	**0.0668**	**0.1986**	0.7889	0.7508	0.7691	**0.2198**	**0.2311**	0.4972	0.5082	0.5144	0.2898	4.87	8.1721	6.8661	8.6333	8.1552	6.6603	7.6974
llama_3_3_70B_hier	0.3746	0.1411	0.1864	**0.2327**	0.1358	0.0879	0.0593	0.1931	**0.7918**	0.7557	0.7730	0.2132	0.2158	0.5156	0.4987	0.5005	0.2863	**4.94**	7.3417	6.3637	8.1545	7.6200	5.9002	7.0760
eurollm_9B_hier	0.2800	0.0611	0.1199	0.1856	0.0832	0.0413	0.0212	0.1451	0.7570	0.7362	0.7459	0.1275	0.1229	0.5249	0.5065	**0.5290**	0.1875	4.84	4.9739	4.4255	6.4043	6.6876	3.5435	5.2070
llama_3_8B_hier	0.2927	0.0780	0.1344	0.1829	0.0897	0.0499	0.0287	0.1472	0.7519	0.7239	0.7373	0.1456	0.1444	0.4958	0.5082	0.5081	0.2289	4.90	5.2780	4.5405	6.3069	6.4295	3.7751	5.2660
mistral_v03_full	0.3612	0.1561	0.1844	0.2126	0.1304	0.0907	0.0660	0.1901	0.7706	0.7255	0.7465	0.2132	0.2074	0.4929	0.5021	0.5044	**0.3252**	4.72	6.9612	5.7141	7.1395	6.8110	5.0271	6.3306
mistral_v03_hier	0.3571	0.1218	0.1638	0.2304	0.1264	0.0780	0.0509	0.1871	**0.7918**	**0.7645**	**0.7777**	0.1884	0.1825	**0.5475**	**0.5122**	0.5189	0.2386	4.69	5.5376	4.9653	5.5578	5.2447	3.7370	5.0085
teuken_hier	0.1630	0.0213	0.0703	0.0794	0.0284	0.0105	0.0043	0.0781	0.6966	0.6303	0.6600	0.0705	0.0673	0.3553	0.5051	0.5068	0.1607	**4.94**	3.0635	2.1606	4.2356	4.4077	1.8269	3.1388
teuken_inst_sft	0.2865	0.0569	0.1164	0.2181	0.0993	0.0531	0.0312	0.1758	0.7973	0.8037	0.8	0.1332	0.1317	0.6039	0.4893	0.4542	0.12	4.54	3.0758	2.0058	4.1411	4.9223	1.5835	3.1457

Table 3: Full automatic evaluation scores (hierarchical Summaries → `_hier_`; complete Judgements → `_full_`)

R1, R2, RL	ROUGE-1/-2/-L F1
B1–B4	BLEU-1 … BLEU-4
MTR	METEOR
BP, BR, BF1	BERTScore Precision/Recall/F1
QGS, Qn	QAGS Score / Ø Questions
KW	Keyword-Overlap
ENT	Entity-Overlap
Len	Length-Ratio
Fcc, FccC	FactCC Score / Consistency
LJ_Fact	`llm_judge` fact. Corr.
LJ_Compl	LLM-as-judge Completeness
LJ_Clar	LLM-as-judge Clarity
LJ_Struc	LLM-as-judge Structure
LJ_Ref	LLM-as-judge Comparison with Reference
LJ_Tot	LLM-as-judge Total Score

Model	Avg. Rank	Ext. Info Rate	Incoherent Rate	Publishable Rate
reference_summary	1.500	0.786	0.071	1.000
openai_gpt_4o_generated_full_summary	4.000	0.643	0.000	0.714
llama_3_3_70B_generated_full_summary	4.071	0.714	0.071	0.571
openai_gpt_4o_gen_hier_summary	4.500	0.571	0.071	0.786
llama_3_3_70B_gen_hier_summary	4.714	0.714	0.000	0.571
mistral_v03_generated_full_summary	5.857	0.714	0.214	0.214
llama_3_8b_gen_hier_summary	6.429	0.714	0.071	0.214
eurollm_gen_hier_summary	7.143	0.786	0.286	0.214
mistral_v03_gen_hier_summary	7.714	1.000	0.143	0.143
teuken_gen_hier_summ_summary-press-summary-sft	9.857	0.929	0.214	0.071
teuken_gen_hier_summary	10.214	0.714	0.214	0.000

Table 4: Full human evaluation of summary metrics for each model (hierarchical Summaries → `_hier_`; complete Judgements → `_full_`)

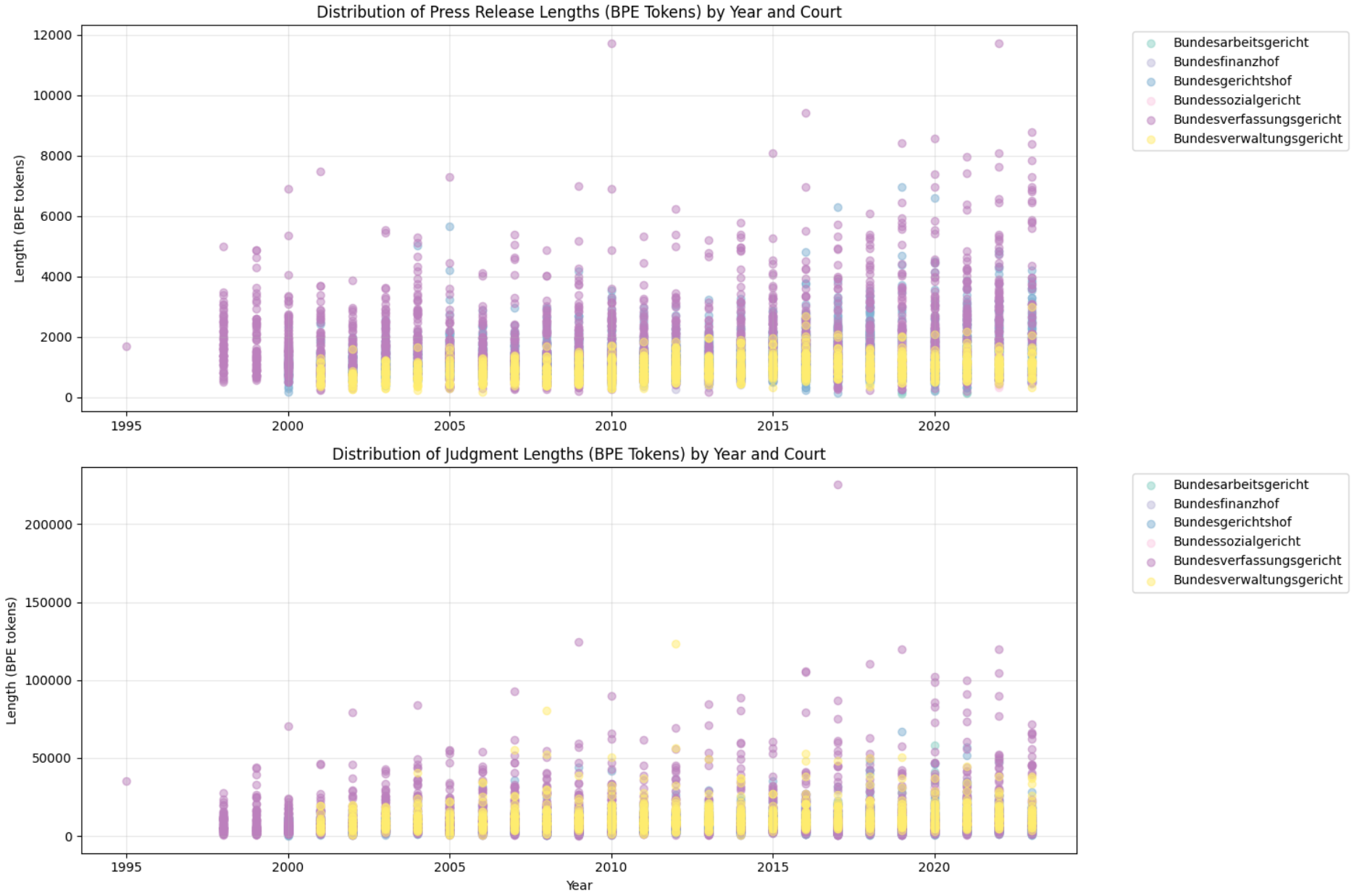

Figure 1: Distribution of PRs and Judgements by year and court