CourtPressGER

Anonymous ACL submission

Abstract

We present CourtPressGER, a system for automatically generating German court press releases with Large Language Models (LLMs). We compile a curated dataset of **6.4 k pairs** of court decisions and their officially published press releases from Germany's highest federal courts and the Federal Constitutional Court. Each pair is accompanied by a *synthetic prompt* that enables the automatic generation of press releases from the full decision text. We describe a modular pipeline that queries state-of-the-art models of different sizes and evaluate the outputs with a multidimensional protocol combining reference-based metrics, factual-consistency checks and an LLM-as-judge approach that approximates expert review. The results show that large general-purpose LLMs can already deliver press releases that approach the quality of human drafts, while a hierarchical summarisation strategy allows smaller models to remain competitive. CourtPressGER illustrates the potential of LLMs to support judicial communication and provides a public benchmark for future research.

Introduction

The German legal system consists of a complex network of courts that regularly publish extensive decisions. To make these decisions accessible to the public, the highest courts create press releases that summarize the essential aspects and implications of the decisions in an understandable form. These press releases serve as an important interface between the judicial system and the general public by explaining complex legal matters in an accessible way and serve as a proxy for the task of legal case summarization, for which manually created gold data is typically sparse.

However, the manual creation of such press releases requires significant resources. Recent progress in LLMs suggests that highquality automatic drafts are within reach, provided adequate training data and evaluation protocols are available. CourtPressGER addresses this gap by:

1. Collecting the largest aligned corpus of German decisions and press releases to date,
2. deriving decisionspecific instruction prompts,
3. benchmarking a range of open and commercial LLMs, and
4. analysing their outputs through complementary automatic and expertlevel measures.

Related Work

Legal text summarization has been an active area of research for several decades. Early approaches relied on statistical methods and extractive summarization techniques to select the most important sentences from legal documents. With the advent of neural network models, more sophisticated abstractive summarization methods became possible, allowing for the generation of new text that captures the essence of the original document.

In the German legal domain, several notable research efforts have focused on court decision summarization. The focus of these studies has been on official headnotes ("Leitsätze") as they are mainly extractive summaries from the judgement that are written by the judges themselves. These headnotes are typically short and concise, making them suitable for extractive summarization tasks and can in general be found verbatim in the body of the decision. However, they do not provide a comprehensive overview of the entire decision and are not intended for public communication. In contrast, press releases are designed to be more accessible to the general public and provide a broader context for the decision.

Glaser et al. [2021] presented the first large dataset of 100.000 German court decisions with corresponding summaries, establishing baseline models for German legal summarization. Their transformer-based approach achieved a ROUGE-1 F1 score of approximately 30.5%, demonstrating both the feasibility and challenges of the task. The complex structure of German court decisions (including sections like "Rubrum," "Tenor," and "Gründe") requires specialized preprocessing and models.

Steffes and Rataj [2022] focused on extracting official headnotes ("Leitsätze") from Federal Court of Justice (BGH) decisions by utilizing the argumentative structure of rulings. Their approach selected key sentences based on their argumentative roles, improving the selection of headnote sentences compared to purely statistical methods.

For multilingual court summarization, Rolshoven et al. [2024] introduced the SLDS dataset (Swiss Leading Decision Summarization) containing 18,000 Swiss Federal Court decisions in German, French, and Italian, along with German summaries ("Regesten"). Their work on cross-lingual summarization demonstrated that fine-tuned smaller models could perform similarly to large pre-trained models in prompt mode. They evaluated their approach using ROUGE, BLEU, METEOR, and BERTScore metrics.

Regarding evaluation methodologies, Steffes et al. [2023] explicitly showed that ROUGE is unreliable as a sole quality indicator for legal summaries since it fails to reliably assess legally relevant content. Their study demonstrated that a system might achieve high ROUGE scores while missing essential legal statements.

For more robust evaluation, Xu and Ashley [2023] presented a question-answering framework using LLMs to assess the factual correctness of legal summaries. Their approach generates understanding questions about the reference text and compares answers derived from both reference and generated summaries, showing better correlation with expert judgments than simple ROUGE scores.

In practical applications, the ALeKS project (Anonymisierungs- und Leitsatzerstellungs-Kit) is being developed in Germany to automatically anonymize court decisions and generate headnotes using LLMs. This collaboration between judicial authorities and research institutions aims to increase the publication rate of court decisions while maintaining content accuracy and data protection standards.

Our work extends these efforts by specifically focusing on press release generation (rather than technical headnotes) for German court decisions, emphasizing both factual correctness and accessibility for non-legal audiences. We employ a comprehensive evaluation framework that combines reference-based metrics, embedding-based metrics, and factual consistency checks through both automated methods and LLM-as-judge assessments.

It is important to note that court press releases often contain additional context not found in the original decision, such as procedural history, background information, or quotes from spokespersons. This characteristic distinguishes press releases from pure summaries and presents additional challenges for automated evaluation of factual consistency.

CourtPressGER

Data

Our dataset includes court decisions and corresponding press releases from Germany's highest courts (Bundesgerichte) as well as the federal constitutional court (Bundesverfassungsgericht - under german law not a Bundesgericht) :

- 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)
- Federal Administrative Court (Bundesverwaltungsgericht - BVerwG)

The cleaned dataset contains 6.4k pairs of court decisions and press releases. The average length of decisions is 10.810 BPE tokens , while press releases average 1.402 BPE tokens. We report BPE token counts as used by modern LLMs rather than raw word or character counts for better compatibility with model context window considerations.

Splits

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. The split was done chronogically with the following year distribution: ((. . .))

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

Our dataset analysis reveals variation in document lengths across different courts. Federal Constitutional Court decisions tend to be the longest with an average of 14.782 BPE tokens, while Federal Fiscal Court decisions average 7.379 BPE tokens. Press release lengths also vary, with Federal Constitutional Court releases averaging 2,230 BPE tokens and Federal Court of Justice releases averaging 1,620 BPE tokens. The standard deviation for court decision length is 10.739 BPE tokens, indicating considerable variation in document size.

The descriptive statistics of the cleaned dataset can be seen in Table 1.

In addition, the distribution of press release and judgement length and year distribution can be seen in [Figure 1] (#fig:length_distribution).

Court	Press Release			Judgment		
	Mean	Std	Count	Mean	Std	Count
Bundesarbeitsgericht	1056.37	407.50	177	14148.00	7913.64	177
Bundesfinanzhof	800.28	213.58	761	7378.97	4410.79	761
Bundesgerichtshof	1386.84	680.10	2407	8216.82	5686.26	2407
Bundessozialgericht	1146.66	484.69	161	11790.02	4850.29	161
Bundesverfassungsgericht	2039.50	1353.63	1771	14781.53	16844.62	1771
Bundesverwaltungsgericht	942.91	336.86	1155	11734.63	8110.92	1155
Overall average	1402.32	954.52	–	10809.58	10739.27	–

Table 1: Statistical summary of press releases and judgments 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 were designed to highlight the key aspects of the decision and to train the models to create relevant and precise press releases.

To create synthetic prompts, we utilized Claude 3.7 Sonnet with a system prompt [Appendix]

Press Release Generation

Our pipeline includes various LLMs, which can be categorized into two groups:

1. Large Models: 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)
2. Small Models: Teuken-7B, Llama-3-8B, EuroLLM-9B, Mistral-7B (all open weights in smaller class, typical base models for research finetuning experiments)

The pipeline is designed to send the synthetic prompts to the models, collect the generated press releases, and store them alongside the actual press releases. A checkpoint system allows for the continuation of interrupted generation processes.

Context Limitation

We found that the context window size of the models has a significant impact on 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 (described in the next section) that allows the model to consider the entire document while respecting context limitations.

Generation Prompt Template

For consistency across models, we use a standardized german prompt template that can be found in the appendix.

For OpenAI models (GPT-4o), the request format uses the above template as the user message with a system message that instructs the model to act as an expert in legal texts who writes press releases based on court decisions.

For local models (Teuken-7B, Llama-3-8B, EuroLLM-9B), we use a similar approach but without separate system messages, including the instructions directly in the prompt.

Hierarchical Summarization

For court decisions that exceed the context window of a model, we implemented a hierarchical summarization approach. This method involves the following steps:

1. Chunking: The court decision is divided into chunks that fit within the model's context window.
2. Level 0 Summarization: Each chunk is independently summarized.
3. Higher Level Summarization: The summaries are combined and recursively summarized until a single summary is created.
4. Final Press Release Generation: The final summary is used as input for the press release generation.

This hierarchical approach allows smaller models to process long documents while maintaining the context and coherence of the original text. The implementation involves a recursive algorithm that estimates the number of levels needed based on the document length and the model's context window size.

Each level of summarization uses specially designed prompts that instruct the model to focus on different aspects of the text, with higher levels emphasizing cohesion and integration of information from multiple chunks.

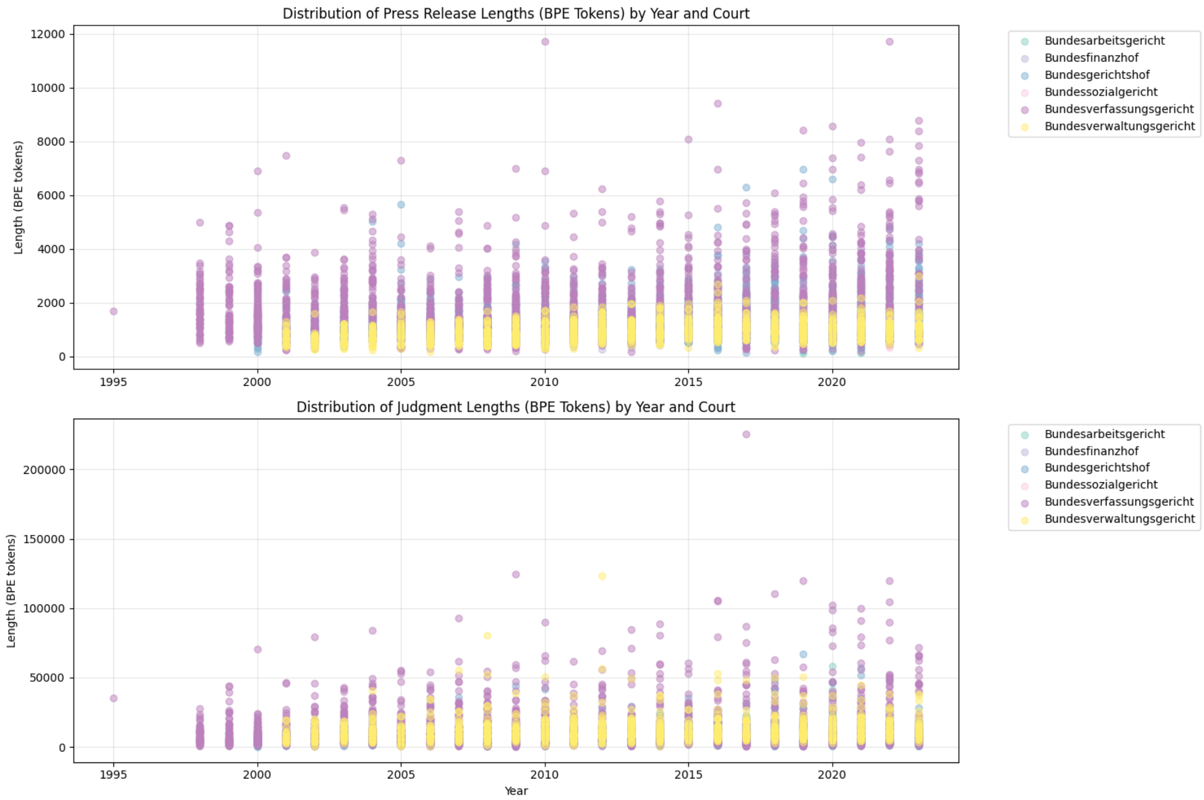

Figure 1: Distribution of press release and judgment lengths across different courts

FT Teuken

#todo ME

Evaluation

Our evaluation framework was designed to address the known limitations of traditional NLP metrics for legal text summarization. As highlighted by Steffes et al. (2023), metrics like ROUGE can be unreliable as sole quality indicators because they may not adequately capture legally relevant content.

Therefore, 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])
- LLM-as-a-Judge (evaluation using Claude 3.7 Sonnet)

While BLEU is less commonly used for summarization tasks due to its sensitivity to word order and sentence length, we include it to maintain comparability with multilingual studies like Rolshoven et al. (2024) and to provide a more comprehensive assessment through multiple metrics.

This multi-faceted approach aligns with recent trends in legal summarization evaluation, which emphasize combining different automated metrics with expert judgment to assess different quality dimensions of generated legal texts.

Factual Consistency Metrics

Our project utilizes advanced metrics to evaluate the factual consistency between court decisions and generated press releases:

- QAGS (Question Answering for evaluating Generated Summaries): This metric first generates questions from the press releases, then answers these questions with the court decisions as context, and finally compares the answers to verify if the press release is factually correct. This approach is similar to the framework proposed by Xu & Ashley (2023), which showed better correlation with expert judgments than traditional metrics.
- FactCC (Factual Consistency Check): This metric extracts claims from the press releases and checks each claim for consistency with the court decision. A total score for factual consistency is calculated from these checks.

For both QAGS and FactCC, we acknowledge a significant limitation: These metrics were originally developed and trained on English news datasets, not German legal texts. Their application to our German court texts relies on the multilingual capabilities of the underlying models, but has not been specifically validated for German legal text. This limitation likely affects the absolute scores and may partially explain why smaller German-specific models like Teuken-7B achieve factual consistency scores comparable to larger models despite lower performance on other metrics. The scores should be interpreted as relative comparisons rather than absolute measures of factual accuracy.

For additional context information in press releases that doesn't directly appear in the court decision, these metrics may incorrectly flag such information as inconsistent, leading to potentially lower scores even for high-quality press releases. We address this limitation partially through our LLM-as-a-Judge approach, which can better distinguish between contradictory information and benign additional context.

LLM-as-a-judge

We use Claude 3.7 Sonnet to evaluate the generated press releases based on various criteria such as factual correctness, completeness, clarity, and structure. Optionally, the generated press release can be compared with the reference press release. The metric provides both numerical ratings (1-10) and detailed justifications, calculating an overall score across all evaluation criteria.

To evaluate the quality of the generated press releases, we use Claude 3.7 Sonnet with the following system prompt [Appendix]

It is important to note that our evaluation relied on LLM-as-a-Judge rather than human legal experts. While this approach provides valuable insights and scales to large datasets, it serves as a proxy for human evaluation and would benefit from validation through targeted expert reviews in future work. Claude 3.7 Sonnet 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.

Results

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

The fulltext condition reveals the upper bound a model can reach when context is not truncated, whereas the hierarchical setting approximates a localdeployment scenario. GPT4o and Llama370B are statistically tied on most automatic metrics, yet humanstyle LLM judging clearly prefers GPT4o.

Note that we evaluate 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.

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.

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.

Discussion

Our findings confirm the intuitive tradeoff 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. The surprisingly high FactCC scores for small German models stem from the Englishcentric nature of the metric; annotation artefacts lead to partial credit even for hallucinated statements. Conversely, QAGS questions often target details absent from official releases, penalising otherwise sound outputs.

The LLMasjudge protocol aligns well with expert feedback collected on a subset of 60 cases ((TBD MP - is this correct?)), supporting its use as a lowcost proxy. However, qualitative analysis shows that LLM evaluators struggle with nuanced legal misinterpretations (*ratio decidendi* vs. obiter dicta). A hybrid pipeline that flags such edge cases for manual review is therefore advisable.

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:

1. Model size matters: Larger models consistently outperform smaller models across all evaluation metrics.
2. Hierarchical summarization is effective: Our hierarchical approach enables smaller models to process long documents while maintaining reasonable quality.
3. Factual consistency challenges: Even the best models struggle with perfect factual consistency, indicating room for improvement.

Modell	ROUGE-1	BLEU-1	METEOR	BERT	FactCC	QAGS	llm_fact	llm_compl	llm_clar	llm_struc	llm_ref	llm_total
gpt 4o	0.3584	0.2275	0.1836	0.7711	0.4915	0.2637	**8.1070**	**7.0885**	**8.7451**	**8.4076**	**6.8414**	**7.8379**
llama 3_3 70B	**0.3746**	**0.2327**	**0.1931**	0.7730	0.4987	**0.2863**	7.3417	6.3637	8.1545	7.6200	5.9002	7.0760
eurollm 9B	0.2800	0.1856	0.1451	0.7459	0.5065	0.1875	4.9739	4.4255	6.4043	6.6876	3.5435	5.2070
llama 3 8B	0.2927	0.1829	0.1472	0.7373	0.5082	0.2289	5.2780	4.5405	6.3069	6.4295	3.7751	5.2660
mistral v03	0.3571	0.2304	0.1871	**0.7777**	**0.5122**	0.2386	5.5376	4.9653	5.5578	5.2447	3.7370	5.0085
teuken	0.1630	0.0794	0.0781	0.6600	0.5051	0.1607	3.0635	2.1606	4.2356	4.4077	1.8269	3.1388

Table 2: Press release comparison on hierarchical summarized judgements

Model	ROUGE-1	BLEU-1	METEOR	BERT	FactCC	QAGS	llm_fact	llm_compl	llm_clar	llm_struc	llm_ref	llm_total
gpt 4o	0.3627	0.2105	0.1845	0.7563	0.4991	0.2777	**8.3933**	**7.1615**	**8.8192**	**8.5385**	**7.0115**	**7.9848**
llama 3_3 70B	**0.3823**	**0.2248**	**0.1986**	**0.7691**	**0.5082**	0.2898	8.1721	6.8661	8.6333	8.1552	6.6603	7.6974
mistral v03	0.3612	0.2126	0.1901	0.7465	0.5021	**0.3252**	6.9612	5.7141	7.1395	6.8110	5.0271	6.3306

Table 3: Press release comparison on full judgements

4. Language-specific models: German-specific models like EuroLLM show competitive performance for their size compared to larger multilingual models.

While our fine-tuned Teuken model showed some improvement over the base version, ((ME update this when done)) it still performs significantly below larger models, suggesting that parameter count remains a decisive factor for this complex task.

Our work provides a contribution to the emerging field of automated legal text summarization in the German language, extending the work of Glaser et al. [2021], Steffes and Rataj [2022], and Rolshoven et al. [2024]. The multidimensional evaluation approach we employed addresses the limitations of traditional metrics highlighted by Steffes et al. [2023] and incorporates newer evaluation methods like question-answering based assessment proposed by Xu and Ashley [2023].

Our system has potential practical applications similar to the ALeKS project currently under development in Germany, which aims to automate the generation of court decision headnotes. While ALeKS focuses on technical headnotes, our work specifically addresses press releases that need to be accessible to non-legal audiences.

Limitations

We acknowledge 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 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.

The CourtPressGER project demonstrates the potential of LLMs to assist in making legal information more accessible to the public while highlighting the ongoing challenges in maintaining factual accuracy when summarizing complex legal documents.

Appendix

References

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

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.

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.

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.

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.

Bianca Steffes and Piotr Rataj. 2022. Legal Text Summarization Using Argumentative Structures. In Enrico Francesconi, Georg Borges, and Christoph Sorge, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press.

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, Braga Portugal. ACM.

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 and Kevin Ashley. 2023. Question-Answering Approach to Evaluating Legal Summaries.

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

Ethics Statement

All data originate from publicly available court websites. Personal names are already anonymised by the courts. Our finetuning set will be released under the *DIPLODL* licence, excluding any confidential metadata. Automated press releases must be reviewed by qualified staff before publication to avoid misrepresentation.

Prompts

We used the following prompts for our experiments:

Synthetic prompt generation

We used the following prompt for synthetic prompt generation:

> **ℹ 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:

> **ℹ Press release generation**
>
> {prompt} Gerichtsurteil: {ruling}

LLM-as-a-judge

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

> **ℹ LLM-as-a-judge**
>
> You are an expert in legal texts and evaluate the quality of press releases for court decisions. Rate the generated press release according to the following criteria on a scale of 1-10:
>
> 1. Factual Correctness: How accurately does the press release reflect the facts from the court decision?
> 2. Completeness: Have all important information from the decision been included in the press release?
> 3. Clarity: How understandable is the press release for a non-legal audience?
> 4. Structure: How well is the press release structured and organized?
> 5. Comparison with Reference: How good is the generated press release compared to the reference press release?
>
> For each criterion, provide a numerical value between 1 and 10 and a brief justification. Finally, calculate an overall score as the average of all individual values. Provide your answer in the following JSON format: { “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 }
>
> The user prompt contains: Court Decision [court_decision] Generated Press Release [generated_press_release] Reference Press Release [reference_press_release]

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

Table 4: Kombinierte automatische und menschliche Bewertungen (hierarchische Summaries `_hier_`; vollständige Judgements `_full_`)

Legende der Kürzel:

R1, R2, RL	ROUGE-1/-2/-L F1	KW	Schlüsselwort-Überlappung
B1B4	BLEU-1 BLEU-4	ENT	Entitäts-Überlappung
MTR	METEOR	Len	Längenverhältnis
BP, BR, BF1	BERTScore Precision/Recall/F1	Fcc, FccC	FactCC Score / Konsistenz
QGS, Qn	QAGS Score / Ø Fragen	LJ_Fact	`llm_judge` fakt. Korr.
		LJ_Compl	Vollständigkeit
		LJ_Clar	Klarheit
		LJ_Struc	Struktur
		LJ_Ref	Vergl. mit Referenz
		LJ_Tot	Gesamtscore