1 Abstract

We presents CourtPressGER - a system for automatically generating German court press releases using Large Language Models (LLMs). We present a curated dataset with 6.4k entries of court decisions with corresponding press releases from Germany's highest courts. The dataset is enhanced with synthetic prompts that enable automated generation of press releases from court decisions. We describe a pipeline for generating press releases with various state-of-the-art models and evaluate the results using automated metrics and LLM-based evaluation approaches that simulate expert assessment. Our approach combines specialized legal language models with domain-specific techniques to produce accurate and informative press releases that adhere to journalistic and legal standards.

2 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. At the same time, recent advances in Large Language Models (LLMs) offer new possibilities for automated text generation in specialized domains. Our project CourtPressGER aims to leverage these capabilities for the automatic generation of court press releases.

The main contributions of our project are:

- The creation of a curated dataset with 6.4k entries of court decisions with corresponding press releases from Germany's highest federal courts.
- The development of synthetic prompts for each decision-press release pair that can be used to automatically generate press releases.
- The evaluation of the generated press releases using a combination of traditional metrics and LLM-based approaches, as well as qualitative output analysis.

3 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 & 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 & 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.

4 CourtPressGER

4.1 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.

4.2 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.

4.3 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 distribution of press release and judgement length and year distribution can be seen in the following table:

	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: Statistical summary of press releases and judgments by court

5 Experimental Setup

5.1 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]

5.2 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.

5.2.1 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.

5.2.2 Generation Prompt Template

For consistency across models, we use a standardized prompt template [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.

5.3 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.

5.4 FT Teuken

#todo ME

6 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 (Rouge-1, Rouge-2, Rouge-L)
- BLEU (BLEU-1 to BLEU-4)
- METEOR
- BERTScore
- QAGS (Question Answering for evaluating Generated Summaries)
- FactCC (Factual Consistency Check)
- 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.

6.0.1 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.

6.0.2 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.

7 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.

Modell	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore P	BERTScore R	BERTSco
gpt 4o	0.3584	0.1242	0.1758	0.2275	0.1280	0.0786	0.0495	0.1836	0.7835	0.7595	0
llama 3_3 70B	**0.3746**	**0.1411**	**0.1864**	**0.2327**	**0.1358**	**0.0879**	**0.0593**	**0.1931**	**0.7918**	0.7557	0
eurollm	0.2800	0.0611	0.1199	0.1856	0.0832	0.0418	0.0212	0.1451	0.7570	0.7362	0
llama 3 8b	0.2927	0.0780	0.1344	0.1829	0.0897	0.0499	0.0287	0.1472	0.7519	0.7239	0
mistral v03	0.3571	0.1218	0.1638	0.2304	0.1264	0.0780	0.0509	0.1871	0.7908	**0.7645**	**0**
teuken	0.1630	0.0213	0.0703	0.0794	0.0284	0.0105	0.0043	0.0781	0.6966	0.6303	0

Table 2: Press release comparison on hierarchical summarized judgements.

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore Precision	BERTScore Recall
gpt 4o	0.3627	0.1452	0.1918	0.2105	0.1266	0.0832	0.0559	0.1845	0.7746	0.7396
llama 3_3 70B	**0.3823**	**0.1601**	**0.1997**	**0.2248**	**0.1385**	**0.0946**	**0.0668**	**0.1986**	**0.7889**	**0.7508**
mistral v03	0.3612	0.1561	0.1844	0.2126	0.1304	0.0907	0.0660	0.1901	0.7706	0.7255

Table 3: Press release comparison on full judgements

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.

7.1 Reference-based Metrics

Our evaluation of reference-based metrics shows that larger models consistently outperform smaller models across all metrics.

According to our evaluation, the complete results for full-document processing (without hierarchical summarization) are:

For hierarchical summarization, the performance is slightly lower but follows a similar pattern:

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.

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Llama-3-70B	0.3823	0.1601	0.1997
GPT-4o	0.3627	0.1452	0.1918
Mistral-v0.3	0.3612	0.1561	0.1844

Table 4: Reference-based metrics for full-document processing

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Llama-3-70B	0.3746	0.1411	0.1864
GPT-4o	0.3584	0.1242	0.1758
Mistral-v0.3	0.3571	0.1218	0.1638
Llama-3-8B	0.2927	0.0780	0.1344
EuroLLM	0.2800	0.0611	0.1199
Teuken-7B	0.1630	0.0213	0.0703

Table 5: Reference-based metrics for hierarchical summarization

7.2 Embedding-based Metrics

BERTScore metrics, which capture semantic similarity using contextual embeddings, show similar trends to the reference-based metrics.

This metric is particularly relevant for legal text evaluation as noted by recent surveys, which highlight BERTScore's ability to detect semantic similarity beyond simple n-gram matching:

Model	BERTScore Precision	BERTScore Recall	BE
Llama-3-70B	0.7889	0.7508	
GPT-4o	0.7746	0.7396	
Mistral-v0.3	0.7706	0.7255	

Table 6: Embedding-based metrics for full-document processing

For hierarchical summarization:

7.3 Factual Metrics

The QAGS evaluation, which measures factual consistency through question answering (similar to the approach proposed by Xu & Ashley), shows varying degrees of factual accuracy:

For hierarchical summarization:

Model	BERTScore Precision	BERTScore Recall	BERTScore F1
Llama-3-70B	0.7918	0.7557	0.7730
GPT-4o	0.7835	0.7595	0.7711
Mistral-v0.3	0.7918	0.7645	0.7777
Llama-3-8B	0.7519	0.7239	0.7373
EuroLLM	0.7570	0.7362	0.7459
Teuken-7B	0.6966	0.6303	0.6600

Table 7: Embedding-based metrics for hierarchical summarization

Model	QAGS Score	Question Count
Llama-3-70B	0.2898	4.87
GPT-4o	0.2777	4.75
Mistral-v0.3	0.3252	4.72

Table 8: QAGS evaluation for full-document processing

Model	QAGS Score	Question Count
Llama-3-70B	0.2863	4.94
GPT-4o	0.2637	4.78
Mistral-v0.3	0.2386	4.69
Llama-3-8B	0.2289	4.90
EuroLLM	0.1875	4.84
Teuken-7B	0.1607	4.94

Table 9: QAGS evaluation for hierarchical summarization

FactCC scores, which directly evaluate the factual consistency of claims:

Model	FactCC Score	Consistency Ratio	Claim Cou
Llama-3-70B	0.5082	0.5144	5.0
GPT-4o	0.4991	0.5068	5.0
Mistral-v0.3	0.5021	0.5044	4.9

Table 10: FactCC scores for full-document processing

For hierarchical summarization, Teuken-7B achieved a FactCC score of 0.5051 with a consistency ratio of 0.5068, comparable to larger models despite its lower performance on other metrics. This surprising result likely reflects limitations in applying FactCC to German texts rather than true parity in factual consistency, as our LLM-as-Judge evaluation shows significant differences in factual correctness scores.

7.3.1 LLM-as-a-Judge

The LLM-as-a-Judge evaluation using Claude 3.7 Sonnet provides a more nuanced assessment of the generated press releases, addressing the dimensions of quality emphasized in legal summarization research:

Model	Factual Correctness	Completeness	Clarity
Llama-3-70B	8.17	6.87	8.63
GPT-4o	8.39	7.16	8.82
Mistral-v0.3	6.96	5.71	7.14

Table 11: LLM-as-a-Judge evaluation for full-document processing

For hierarchical summarization:

Model	Factual Correctness	Completeness	Clarity
Llama-3-70B	7.34	6.36	8.15
GPT-4o	8.11	7.09	8.75
Mistral-v0.3	5.54	4.97	5.56
Llama-3-8B	5.28	4.54	6.31
EuroLLM	4.97	4.43	6.40
Teuken-7B	3.06	2.16	4.24

Table 12: LLM-as-a-Judge evaluation for hierarchical summarization

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.

8 Discussion

tbd

9 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.
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, 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 & 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 & 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.

9.1 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.

10 Appendix

10.1 Ethics Statement

tbd

10.2 References

tbd

10.3 Prompts

We used the following prompts for our experiments:

10.3.1 Synthetic prompt generation

10.3.2 Press release generation

10.3.3 LLM-as-a-judge

Instructions for *ACL Proceedings

Anonymous ACL submission

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

11 Introduction

These instructions are for authors submitting papers to *ACL conferences using LaTeX. They are not self-contained. All authors must follow the general instructions for *ACL proceedings,[1] and this document contains additional instructions for the LaTeX style files.

The templates include the LaTeX source of this document (`acl_latex.tex`), the LaTeX style file used to format it (`acl.sty`), an ACL bibliography style (`acl_natbib.bst`), an example bibliography (`custom.bib`), and the bibliography for the ACL Anthology (`anthology.bib`).

12 Engines

To produce a PDF file, pdfLaTeX is strongly recommended (over original LaTeX plus dvips+ps2pdf or dvipdf). The style file `acl.sty` can also be used with luaLaTeX and XeLaTeX, which are especially suitable for text in non-Latin scripts. The file `acl_lualatex.tex` in this repository provides an example of how to use `acl.sty` with either luaLaTeX or XeLaTeX.

13 Preamble

The first line of the file must be

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{acl}
```

For the final version, omit the `review` option:

```
\usepackage{acl}
```

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.)

Please see the LaTeX source of this document for comments on other packages that may be useful.

Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LaTeX source for examples.

By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

14 Document Body

14.1 Footnotes

Footnotes are inserted with the `\footnote` command.[2]

14.2 Tables and figures

See Table 13 for an example of a table and its caption. **Do not override the default caption sizes.**

As much as possible, fonts in figures should conform to the document fonts. See Figure 1 for an example of a figure and its caption.

[1] http://acl-org.github.io/ACLPUB/formatting.html

[2] This is a footnote.

Command	Output	Command	Output
`{\"a}`	ä	`{\c c}`	ç
`{\^e}`	ê	`{\u g}`	ğ
``{\`i}``	ì	`{\l}`	ł
`{\.I}`	İ	`{\~n}`	ñ
`{\o}`	ø	`{\H o}`	ő
`{\'u}`	ú	`{\v r}`	ř
`{\aa}`	å	`{\ss}`	ß

Table 13: Example commands for accented characters, to be used in, *e.g.*, BibTEX entries.

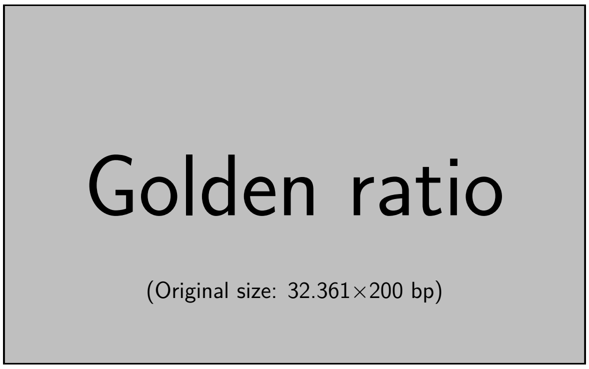

Figure 1: A figure with a caption that runs for more than one line. Example image is usually available through the mwe package without even mentioning it in the preamble.

Using the `graphicx` package graphics files can be included within figure environment at an appropriate point within the text. The `graphicx` package supports various optional arguments to control the appearance of the figure. You must include it explicitly in the LATEX preamble (after the `\documentclass` declaration and before `\begin{document}`) using `\usepackage{graphicx}`.

14.3 Hyperlinks

Users of older versions of LATEX may encounter the following error during compilation:

> `\pdfendlink` ended up in different nesting level than `\pdfstartlink`.

This happens when pdfLATEX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LATEX to 2018-12-01 or later.

14.4 Citations

Table 14 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by **?**. You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (**?**). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. **?**).

A possessive citation can be made with the command `\citeposs`. This is not a standard natbib command, so it is generally not compatible with other style files.

14.5 References

The LATEX and BibTEX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LATEX file will generate the references section for you:

```
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTEX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliography{anthology,custom}
```

Please see Section 15 for information on preparing BibTEX files.

14.6 Equations

An example equation is shown below:

$$A = \pi r^2 \tag{1}$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command.

This an example cross-reference to Equation 1.

14.7 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

15 BibTEX Files

Unicode cannot be used in BibTEX entries, and some ways of typing special characters can disrupt BibTEX's alphabetization. The recommended way of typing special characters is shown in Table 13.

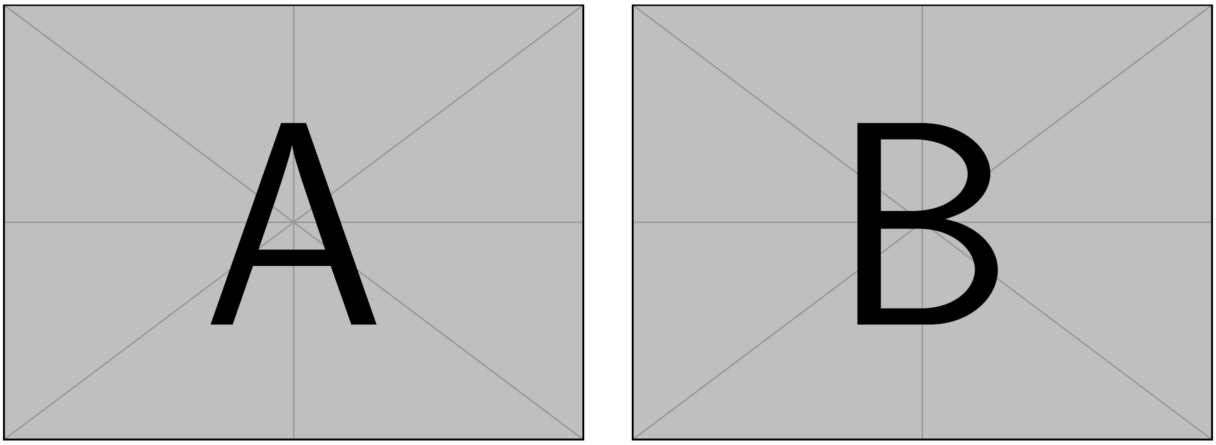

Figure 2: A minimal working example to demonstrate how to place two images side-by-side.

Output	**natbib command**	**ACL only command**
(**?**)	`\citep`	
?	`\citealp`	
?	`\citet`	
(**?**)	`\citeyearpar`	
?’s (**?**)		`\citeposs`

Table 14: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

Please ensure that BibTEX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTEX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LATEX package.

Limitations

Since December 2023, a "Limitations" section has been required for all papers submitted to ACL Rolling Review (ARR). This section should be placed at the end of the paper, before the references. The "Limitations" section (along with, optionally, a section for ethical considerations) may be up to one page and will not count toward the final page limit. Note that these files may be used by venues that do not rely on ARR so it is recommended to verify the requirement of a "Limitations" section and other criteria with the venue in question.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

A Example Appendix

This is an appendix.