A Multi-Dimensional Expert-Annotated Dataset for Evaluating LLM-Generated German News Summaries

Neha Deshpande **Stefan Hillmann** **Sebastian Möller**
Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany
{n.deshpande, stefan.hillmann, sebastian.moeller}@tu-berlin.de

Abstract

Large language models (LLMs) are increasingly used for abstractive summarization, but evaluating generated summaries remains a challenge, especially for languages other than English. Automatic metrics such as ROUGE and BERTScore often fail to capture important aspects of summary quality, whereas factuality-oriented and reference-free evaluation methods remain less established for German. To support human-grounded evaluation of German summarization, we introduce an expert-annotated dataset of German news summaries generated by six LLMs. The dataset was annotated by three native German-speaking annotators with linguistic expertise, using a structured protocol with eight fine-grained quality dimensions and pairwise preference judgments. We analyze inter-annotator agreement for both feedback types and examine how dimension-level ratings relate to overall human preferences. Our results show that pairwise preferences yield higher agreement than most dimension-level ratings. Among the eight dimensions, non-redundancy, grammaticality, and faithfulness, are judged most consistently, while referential clarity and structure & coherence show lower agreement. Preference analysis further indicates that content coverage and faithfulness are most strongly associated with human summary choice. The dataset provides a German-language resource for studying summarization evaluation, human preference behavior, and future LLM-as-a-judge benchmarking. Our dataset as well as the supporting python scripts are available here: https://git.tu-berlin.de/neha.deshpande/annotated_news_summaries

1 Introduction

The widespread adoption of Large Language Models (LLMs) has substantially increased the use of automated techniques for text-based tasks, particularly summarization. Recent advances in LLM capabilities have enabled the generation of high-quality summaries across a wide range of domains, including news articles, books, interviews, podcasts, lectures, scientific literature, medical text, and code (Zhang et al., 2025; Sharma et al., 2025; Huang et al., 2025; Xie et al., 2025). Beyond direct end-user applications, summarization is also increasingly used as an intermediate component in downstream systems, including question answering and retrieval-augmented generation (RAG) pipelines (Tang et al., 2025; Reuter et al., 2025).

Despite these advances, evaluating the quality of generated summaries remains a significant challenge. Traditional automatic metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) rely on reference summaries and often fail to capture important aspects of quality, particularly for abstractive text (Kirstein et al., 2025; Zhang et al., 2025). More recent approaches, including UniEval (Zhong et al., 2022), QuestEval (Scialom et al., 2021), and FactEval (Mamta and Cocarascu, 2025), attempt to provide reference-free evaluation. However, these approaches have been developed, trained, or evaluated largely in English-centered settings, limiting their reliability and generalizability to other languages. This highlights the need for language-specific evaluation resources beyond English.

To address these limitations, recent work has explored automatic evaluation using LLMs, often referred to as LLM-as-a-judge (Gu et al., 2025). In this setting, LLMs are used to assess summary quality and have shown promising results not only for evaluating generated summaries but also for improving them (Song et al., 2025; Stiennon et al., 2022). However, the development and validation of such methods depend on reliable evaluation data, including human annotations, human-in-the-loop feedback, or carefully constructed AI-generated datasets (Gu et al., 2025). Most research in this area has relied on English datasets (Gao and Wan,

2022; Fabbri et al., 2021; Stiennon et al., 2022). Since LLM-based evaluators are commonly assessed through their agreement with human preferences (Yamauchi et al., 2026; Li et al., 2025), their reliability in multilingual settings remains difficult to establish when high-quality human-annotated data is limited (Min et al., 2025).

To address the lack of publicly available German-language resources for summarization evaluation, we introduce an expert-annotated dataset of LLM-generated German news summaries. The dataset contains summaries generated by six LLMs and evaluated by expert annotators using two complementary forms of feedback: Likert-scale ratings across eight fine-grained quality dimensions and pairwise preference judgments.

Beyond presenting the dataset, this work investigates the reliability and usefulness of multi-dimensional human evaluation for German abstractive summarization. We analyze inter-annotator agreement, item-level rating variability, and the relationship between dimension-level scores and overall preference judgments. These analyses allow us to identify which dimensions provide more stable evaluation reliability and which are most closely associated with human summary choice. This dataset can provide a human-based foundation for future work on evaluation of German LLM-as-a-judge, synthetic feedback comparison, prompting, fine-tuning approaches, and generation of synthetic datasets (Gu et al., 2025; Song et al., 2025).

Based on this dataset, we aim to answer the following research questions:

- **RQ1:** To what extent do expert annotators provide consistent judgments when evaluating LLM-generated German news summaries using dimension-level ratings and pairwise preferences?
- **RQ2:** To what extent are expert annotations reliable across all eight quality dimensions and how does agreement vary between dimensions?
- **RQ3:** Which quality dimensions are most strongly associated with expert annotators' overall preferences between summaries?

Our main contributions are as follows: 1) We introduce a human-annotated benchmark dataset for German news summarization, comprising summaries generated by six LLMs and evaluated by expert annotators across eight fine-grained quality dimensions and pairwise preferences. 2) We assess the reliability of the annotation protocol by measuring inter-annotator agreement across dimension-level ratings and pairwise preference judgments. 3) We investigate the role of individual quality dimensions in human summary selection, identifying which dimensions are most strongly associated with preference judgments. 4) We provide practical insights into the annotation of summaries, including the process and selection of dimensions.

2 Related Work

2.1 Annotated Evaluation Datasets

Publicly available datasets for summarization evaluation have primarily focused on English. Datasets such as SummEval (Fabbri et al., 2021) and RealSumm (Bhandari et al., 2020) provide human judgments across dimensions such as coherence, consistency, fluency, and relevance, and are commonly used to analyze correlations between human ratings and automatic metrics. The SummEval dataset introduced by Fabbri et al. (2021) includes both expert and crowd-sourced annotations, with expert annotations found to better differentiate coherence, consistency, and fluency across summaries than crowd-sourced annotations. Existing datasets use different annotation schemes, including Likert-scale ratings (Fabbri et al., 2021), semantic content units (Bhandari et al., 2020), and question-answering-based evaluation (Liu et al., 2024). Other resources focus on specific aspects of summary quality, such as factual errors and hallucinations (Pagnoni et al., 2021; Subbiah et al., 2025), multi-domain factual evaluation (Lee et al., 2024), or domain-specific settings such as dialogue and meeting summarization (Gao and Wan, 2022; Kirstein et al., 2024). These datasets provide important foundations for studying human judgments and automatic metrics.

In addition to absolute quality ratings, preference-based datasets capture relative judgments between summaries. Such feedback reflects which summary humans prefer when comparing alternative outputs and has been used to align summarization systems with human preferences through reinforcement learning from human feedback (Stiennon et al., 2022) and preference-based optimization methods such as Direct Preference Optimization (Song et al., 2025).

2.2 LLM-Based Evaluation

A growing body of work has also explored automatic evaluation using LLMs and other reference-free strategies to reduce dependence on gold summaries. LLM-as-a-judge approaches use LLMs to directly assess the quality of LLM-generated outputs. For example, UniSumEval (Lee et al., 2024) uses LLM-based filtering to identify challenging hallucination-prone examples, while QA-based approaches such as SummEQual (Liu et al., 2024) use model-generated questions and answers to estimate summary quality. However, it remains unclear whether evaluation dimensions and preference patterns transfer reliably to other languages and to LLMs fine-tuned for those languages.

2.3 Multilingual and German Summarization Evaluation

Despite progress in summarization evaluation, multilingual settings remain underexplored. Datasets such as SEAHORSE (Clark et al., 2023) extend evaluation to multiple languages, but often rely on simplified annotation schemes, limited model coverage, or broad categorical judgments. As a result, there is still limited evidence on how fine-grained evaluation dimensions behave across languages.

This gap is particularly apparent for German, where comprehensive summarization evaluation resources with expert human annotations remain scarce. Our work addresses this gap by introducing a German news summarization evaluation dataset with summaries generated by six LLMs and annotated across eight fine-grained quality dimensions and pairwise preference judgments. This setup allows us to analyze the reliability of individual evaluation dimensions and their relationship to overall human summary preferences.

3 Methodology

3.1 Overview

This section describes our data preparation pipeline, followed by summary generation using multiple LLMs, and the annotation framework used to collect human judgments.

3.2 Data Collection and Pre-processing

We collected 2,917 German-language news articles published over three months in 2025 by DIE ZEIT through its online platform ZEIT ONLINE[1]. Since all source articles come from a single news outlet, the dataset may reflect outlet-specific writing style, topic selection, and editorial conventions. The articles vary substantially in length, ranging from 24 to 4,388 words, with an average length of 812.34 words.

The collected articles were pre-processed to ensure consistency and readability. Text normalization was applied to correct encoding issues commonly observed in German text, particularly those affecting umlauts (e.g., ä, ö, ü) and the Eszett (ß), which may appear in inconsistent or corrupted forms due to encoding mismatches (e.g., UTF-8 vs. Latin-1). Missing or inconsistent titles were completed through additional scraping or manual curation.

Each article was then assigned a topic label using GPT-3.5-Turbo[2], based on a predefined set of news categories, including politics, health, economy, society, culture, environment, law, science and technology, sports, transport, digital media, and other categories such as money, education, real estate, travel, and miscellaneous topics. As an informal quality check, we manually inspected the automatically assigned topic labels for approximately 20–25 randomly selected articles. The inspected labels appeared broadly appropriate, and we observed no obvious systematic errors in this sample. The resulting topic distribution is shown in Appendix A.1.

3.3 Summary Generation

For the annotation process, we generated summaries for each news article using six LLMs of varying sizes and capabilities, including both German-specific and multilingual models. All models, except GPT-3.5-Turbo, were accessed via Ollama[3], a local model inference platform. The Ollama models were run using quantized versions. The models used and their respective quantization configurations are listed in Table 1.

The generation parameters were kept consistent across the models (temperature = 0, top_p = 1, max_tokens = 512) to ensure deterministic results. The summaries were constrained to a maximum of five sentences. A single standardized prompt was used across all models to ensure comparability. The prompt was designed to produce concise news summaries covering key information and is provided in Appendix A.4. A total of 17,502 summaries were

[1] https://www.zeit.de/index

[2] https://developers.openai.com/api/docs/models/gpt-3.5-turbo

[3] https://ollama.com/

Model	Description	Params
sauerkraut[4]	German instruction-tuned model (Q4_0)	12B
german-leo[5]	German-adapted Mistral-based model (Q4_K_M)	7B
wiedervereinigung[6]	German fine-tuned model (Q4_0)	7B
llama-3b[7]	Multilingual instruction-tuned model (Q4_K_M)	3B
llama-1b[8]	Multilingual instruction-tuned model (Q8_0)	1B
GPT-3.5-Turbo[9]	Proprietary multilingual model	–

Table 1: Overview of the models used for summary generation, including model type, parameter size, and quantization format. Q4 and Q8 indicate approximately 4-bit and 8-bit weight quantization, respectively, while K_M denotes a mixed K-quantization scheme designed to balance model size and output quality (Kurt, 2026).

generated for 2917 articles.

3.4 Annotation Framework

3.4.1 Selection of Dimensions

We define eight evaluation dimensions for assessing summary quality, drawing on prior work in multi-dimensional summarization evaluation and adapting it to the German news domain. Existing evaluation frameworks commonly consider dimensions such as coherence, consistency, fluency, and relevance. For example, SummEval (Fabbri et al., 2021) evaluates summaries along coherence, consistency, fluency, and relevance using a 5-point Likert scale, while Stiennon et al. (2022) consider coherence, accuracy, and overall quality using a 7-point scale. Similarly, DialSummEval (Gao and Wan, 2022) includes dimensions such as fluency, coherence, relevance, and consistency. More recent work such as UniSumEval (Lee et al., 2024) focuses on factuality-related dimensions, highlighting the importance of evaluating factual information in LLM-generated text.

In addition to these commonly used dimensions, Conroy and Dang (2008) introduce linguistic quality dimensions such as grammaticality, non-redundancy, referential clarity, focus, and structure and coherence. These dimensions are relevant for summarization evaluation because they capture aspects of readability, clarity, organization, and repetition, which can influence higher-level judgments of summary quality.

For our annotation framework, we combine content-related and linguistic dimensions to capture different aspects of German news summaries. *Faithfulness* and *Content Coverage* assess whether the summary preserves factual information and includes important content from the source article. *Grammaticality*, *Non-redundancy*, *Referential Clarity*, and *Structure & Coherence* capture linguistic quality and readability. *Focus* corresponds closely to the notion of relevance used in prior work (Fabbri et al., 2021; Gao and Wan, 2022), while *Complexity* is included to account for sentence construction, readability, and structural complexity, which are particularly relevant in German due to flexible word order and complex syntactic constructions.

Building on the reviewed literature, we define the following eight dimensions for our annotation framework:

1. **Grammaticality**: The summary should not contain datelines, formatting artifacts, capitalization errors, non-idiomatic phrasing, or ungrammatical sentences.

2. **Non-redundancy**: The summary should avoid unnecessary repetition of facts, phrases, or entities.

3. **Referential Clarity**: Pronouns and noun phrases should clearly refer to identifiable entities, and the roles of these entities should be understandable.

4. **Focus**: All sentences in the summary should be relevant to the article and should avoid unrelated or off-topic information.

5. **Structure and Coherence**: The summary should present information in a logically structured and coherent manner, where sentences connect naturally.

6. **Content Coverage**: The summary should capture the essential information from the article so that the main story can be understood without reading the original text.

7. **Faithfulness**: The summary should be factually correct with respect to the source article and should not contain hallucinated information.

[4] https://ollama.com/cyberwald/sauerkrautlm-nemo-12b-instruct

[5] https://ollama.com/bengt0/em_german_leo_mistral

[6] https://ollama.com/mayflowergmbh/wiedervereinigung

[7] https://ollama.com/library/llama3.2:3b

[8] https://ollama.com/library/llama3.2:1b

[9] https://developers.openai.com/api/docs/models/gpt-3.5-turbo

8. **Complexity**: The style and sentence structure should be natural and appropriate in complexity, avoiding overly simplistic or overly complicated phrasing.

3.4.2 Annotator Recruitment

We recruited three native German-speaking linguistics students from a pool of ten applicants. The annotators were selected through an interview process involving trial annotations with a think-aloud component, which allowed us to assess their understanding of the task and the evaluation dimensions. They were compensated at €15 per hour and also helped refine the dimension definitions described above.

3.4.3 Task Design

We collected two types of annotations: (1) *dimension-level ratings* and (2) *pairwise preferences*. For the dimension-level task, annotators rated each summary on a 7-point Likert scale across eight predefined quality dimensions. For each instance, an article and two randomized candidate summaries were shown side-by-side in a web interface, and each summary was evaluated independently.

For the preference task, annotators selected their preferred summary from the same pair. To ensure that the pairwise preference dataset covered comparisons between all summary generation models listed in Table 1, we used a stratified sampling strategy based on model pairs. Each model-pair combination was treated as a separate stratum, and samples were retained from each stratum. This resulted in a final comparison set in which all model-pair combinations are represented. In addition, the left–right position of summaries was randomized so that summaries from the same model did not consistently appear in the same position, reducing potential position bias. Model identities were not shown to annotators during the task.

3.4.4 Annotation Setup and Procedure

To facilitate the annotation process and ensure controlled data collection, we developed a web application hosted on the university server. The summary annotation interface is shown in Figure 1, with a full screenshot provided in Appendix 3.

Each article was associated with six generated summaries, one from each model. To reduce repeated reading effort, the interface grouped summaries of the same article across consecutive pages. Each page displayed the full source article together with two summaries, so annotators could evaluate multiple summaries after reading the article once while still having the source text available for reference. The interface also allowed annotators to skip items and return to them later, resume from a previous serial number, and go back to correct earlier annotations.

Annotators were given detailed annotation guidelines in a separate document, including definitions for each evaluation dimension and rating descriptions for scores 1, 4, and 7. A shortened version of these guidelines is provided in Appendix A.2.

The annotation process was conducted over approximately two months. During this period, three calibration meetings were held to align interpretations of the evaluation dimensions and to clarify future annotation decisions.

3.4.5 Annotation Statistics

Due to practical constraints such as annotation time and cost, only a subset of the generated summaries described in Section 3.3 was annotated. Table 2 summarizes the annotation coverage for the two feedback types: *dimension-level ratings* and *pairwise preference judgments*. It reports the number of annotated units with at least one, at least two, and all three annotators.

Annotation Type	**≥1**	**≥2**	**3**
Dimension-level ratings	2,256	1,470	780
Preference judgments	1,128	735	390

Table 2: Annotation coverage by feedback type and number of annotators.

4 Results and Analysis

This section first reports inter-annotator agreement for dimension-level ratings and pairwise preference judgments. We also analyzed which dimensions are associated with pairwise preferences, and performed some additional analysis about the relationships among the quality dimensions, and of the summary length and dimension-level ratings.

4.1 Inter-annotator Agreement

To assess annotation reliability, we calculate inter-annotator agreement (IAA) using Krippendorff's α for ordinal dimension-level ratings and nominal pairwise preference judgments. Krippendorff's α is suitable for corpus annotation tasks because it supports multiple annotators, missing values, and

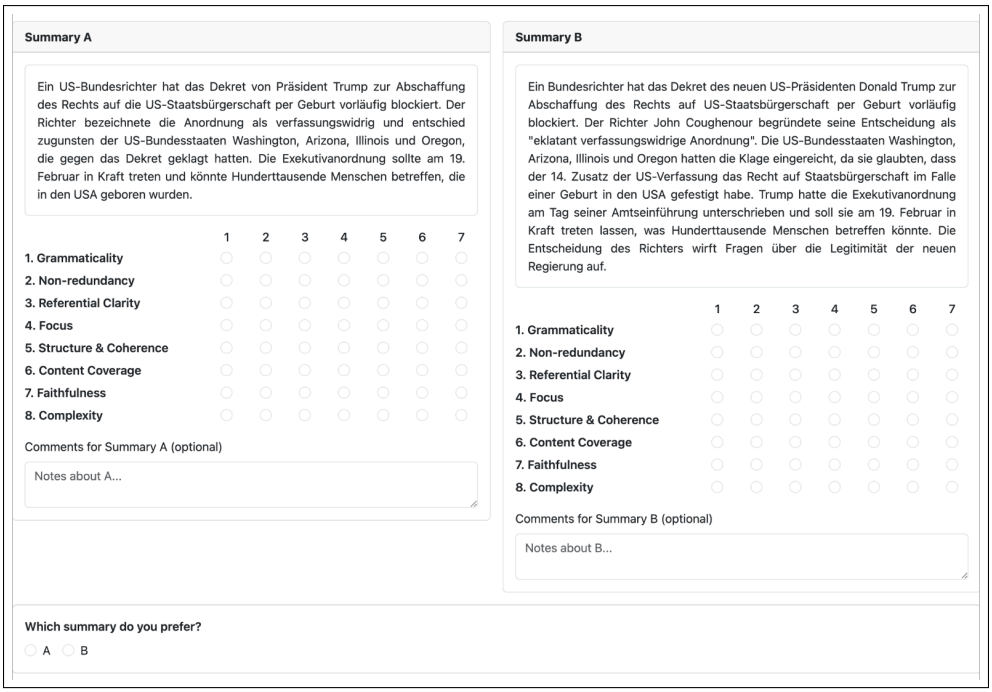

Figure 1: Annotation interface used for collecting dimension-level ratings and pairwise preferences.

different measurement levels (Krippendorff, 2018; Artstein and Poesio, 2008). Agreement was calculated separately for units annotated by at least two annotators and for the subset annotated by all three annotators for both types of feedback.

4.1.1 Dimension-level Ratings

Table 3 shows the agreement calculated separately for units annotated by at least two annotators and for the subset annotated by all three annotators for both feedback types. The highest agreement is observed for dimensions such as *Non-redundancy* (α = 0.712) and *Grammaticality* (α = 0.682), while dimensions such as *Structure & Coherence* (α = 0.252) and *Referential Clarity* (α = 0.273) show substantially lower agreement. A similar pattern holds for units annotated by all three annotators, indicating that the agreement trends remain stable across both subsets.

4.1.2 Pairwise Preferences

Inter-annotator agreement for pairwise summary preferences was measured using Krippendorff's α for nominal data. As shown in Table 4, agreement is high both for items annotated by at least two annotators ($\alpha = 0.738$) and for those annotated by all three annotators ($\alpha = 0.731$).

4.1.3 Selected Annotated Subset

Because the agreement results for units annotated by at least two annotators and those annotated by all three annotators led to broadly similar conclusions (Tables 3 and 4), the remaining analyses use units with at least two annotations. This retains a larger sample while ensuring that every included unit has multiple judgments. Table 5 summarizes the size and text lengths of this subset, while Table 6 provides summary-length statistics for each summarization model.

Dimension	α ($N \geq 2$)	α ($N = 3$)
Structure & Coherence	0.252	0.205
Referential Clarity	0.273	0.279
Focus	0.468	0.434
Complexity	0.405	0.456
Content Coverage	0.540	0.561
Faithfulness	0.638	0.642
Grammaticality	0.682	0.664
Non-redundancy	**0.712**	**0.706**

Table 3: Inter-annotator agreement (Krippendorff's α) across dimensions. Results are shown for units annotated by at least two annotators ($N \geq 2$, 1,470 units) and for units annotated by all three annotators ($N = 3$, 780 units).

Subset	**# Items**	**Krippendorff's α**
$\geq$2 annotators	735	0.738
3 annotators	390	0.731

Table 4: Agreement for pairwise preferences measured using Krippendorff's α for nominal data.

Text type	N	Mean words	Range
Articles	284	782.49	66–3,453
Summaries	1,470	132.94	6–373

Table 5: Number and length of source articles and generated summaries in the selected analysis subset. Range denotes the minimum–maximum word count.

Model	N	Mean words	Range
GPT-3.5	261	82.10	46–176
LLaMA-1B	231	192.20	6–314
LLaMA-3B	254	104.51	19–316
German-Leo	238	68.66	12–266
Sauerkraut	256	171.72	34–373
Wiedervereinigung	230	185.85	54–250

Table 6: Summary-length statistics by model in the selected analysis subset. N denotes the number of unique item–model summaries. Range denotes the minimum–maximum word count.

4.2 Preference Analysis

We use logistic regression to analyze which quality dimensions are associated with pairwise preference decisions, using dimension-rating differences between the chosen and rejected summaries as predictors (Peng et al., 2002). The analysis uses the 735 pairwise comparison tasks with at least two annotators (Table 2), rather than relying on singly annotated comparisons.

This subset includes 345 tasks with two selections and 390 tasks with three selections, yielding 1,860 annotator-level preference selections. Since multiple selections can correspond to the same comparison, we report cluster-bootstrap 95% confidence intervals based on resampling comparison tasks.

For each dimension, we report odds ratios (OR; e^{β}), confidence intervals, and p-values. OR values above 1 indicate that larger rating differences for that dimension are associated with higher odds that a summary is preferred, after accounting for the remaining dimensions.

Table 7 shows that all dimensions are significantly associated with preference decisions. *Content Coverage* and *Faithfulness* have the largest odds ratios, suggesting that annotators primarily prefer summaries that capture relevant information while remaining factually accurate.

As a robustness check, we also fit a comparison-level model using the 671 majority-agreement comparisons, each collapsed to one row. The strongest associations remained *Content Coverage* (OR = 3.44), *Complexity* (OR = 3.39), and *Faithfulness* (OR = 2.79), while *Referential Clarity* was no longer significant.

Dimension	OR	95% CI	p
Content Coverage	**2.37**	[2.11, 2.75]	<0.0001
Faithfulness	**2.12**	[1.92, 2.42]	<0.0001
Complexity	**1.98**	[1.75, 2.29]	<0.0001
Referential Clarity	**1.91**	[1.49, 2.50]	<0.0001
Grammaticality	**1.61**	[1.42, 1.85]	<0.0001
Focus	**1.54**	[1.33, 1.82]	<0.0001
Non-redundancy	**1.53**	[1.33, 1.82]	<0.0001
Structure & Coherence	**1.37**	[1.22, 1.56]	<0.0001

Table 7: Logistic regression relating annotator-level pairwise preferences to dimension-rating differences. OR = odds ratio; CI = cluster-bootstrap 95% confidence interval based on 1,000 bootstrap samples.

Supplementary analyses.

We also examined relationships among quality dimensions and tested whether model-wise rating differences were robust to variation in summary length. The dimensions were positively but not uniformly correlated and the model differences remained significant after adjustment for summary length. The complete results are reported in the Appendix A.9.

5 Discussion

This section discusses the main findings in relation to the research questions introduced earlier. Overall, the results show that expert human evaluation is a feasible basis for German abstractive summarization evaluation, but that reliability depends on both the type of feedback and the quality dimension being assessed.

5.1 RQ1: Reliability of Annotator Judgments

The results show that expert annotators can provide reasonably consistent judgments for LLM-generated German news summaries, but the level of consistency depends on the type of feedback. Pairwise preferences showed high agreement, indicating that annotators were particularly consistent when making relative judgments between two summaries. This suggests that pairwise comparison is a robust setup for collecting overall quality judgments.

Overall, the two feedback types serve complementary purposes: pairwise preferences provide more consistent overall judgments, whereas

dimension-level ratings offer deeper insight into individual summary quality.

5.2 RQ2: Agreement Across Dimensions

Overall, annotation reliability varied substantially across the eight dimensions, indicating that some dimensions were assessed consistently while others remained more subjective. The inter-annotator agreement, computed using Krippendorff's α, was strongest for *Non-redundancy*, *Grammaticality*, *Faithfulness*, and *Content Coverage* (Table 3). These dimensions are comparatively less subjective, as annotators can identify repetition, grammatical errors, factual inconsistencies, or missing key content with less ambiguity. In contrast, *Focus*, *Referential Clarity*, and *Structure & Coherence* showed lower agreement, suggesting stronger dependence on individual reading strategies and expectations about summary organization.

Following Krippendorff's interpretation guidelines, values between 0.67 and 0.80 can support tentative conclusions, while values below 0.67 should be interpreted with caution (Krippendorff, 2018). Thus, *Non-redundancy* and *Grammaticality* fall within the tentative reliability range, while *Faithfulness* approaches this range. However, these thresholds should be interpreted in relation to the annotation task. Prior work in computational linguistics shows that agreement can be lower for complex discourse-level annotation tasks, where category boundaries are difficult to define (Artstein and Poesio, 2008). This also aligns with datasets such as SummEval (Fabbri et al., 2021) and DialSummEval (Gao and Wan, 2022), where agreement varies substantially across quality dimensions, ranging from 0.397 for *Relevance* to 0.796 for *Consistency* in SummEval and from 0.133 to 0.493 across dimensions in DialSummEval.

These results suggest that the annotation scheme provides stronger reliability for dimensions grounded in more observable textual properties, while dimensions involving text organization and reader expectations remain more difficult to assess consistently.

5.3 RQ3: Predictors of Preference

The preference analysis (Section 4.2) shows that *Content Coverage* and *Faithfulness* are most strongly associated with summary choice. This suggests that the annotators primarily preferred summaries that captured important information while avoiding factual errors. In the context of news summarization, this is expected, as a useful summary should preserve the main content of the article while remaining faithful to the source.

Complexity also showed a strong association with preference, indicating that readability may influence summary choice. However, given its lower inter-annotator agreement, this result should be interpreted cautiously. Other dimensions, such as *Grammaticality*, *Non-redundancy*, *Focus*, and *Structure & Coherence*, were also associated with preference, but their smaller effects suggest that they may function more as baseline quality requirements than as the main drivers of summary choice.

5.4 Additional Insights

The supplementary analyses reported in Appendix A.9 show that the quality dimensions were positively but not uniformly correlated. *Focus*, *Faithfulness*, and *Complexity* showed the strongest relationships, while *Content Coverage* was more weakly related to several dimensions. This suggests that the dimensions may capture some related aspects but are not interchangeable.

Longer summaries generally received lower ratings, although *Content Coverage* showed a small positive adjusted relationship with length. The differences between summarization models remained significant after adjusting for length, indicating that summary length alone did not explain the differences in model ratings.

5.5 Insights from the Annotation Process

Annotator comments revealed recurring issues across models, including quotation artifacts, tense errors, gender inconsistencies, and occasional factual inaccuracies. These issues often explained lower ratings on specific dimensions or weaker performance in pairwise comparisons. We also observed a model-specific issue with *Sauerkraut*, which unexpectedly switched to English in two instances despite the task requiring German summaries. This is problematic in a monolingual evaluation setting, as it affects usability and may influence judgments across dimensions.

The annotation process required several refinements, including clarification of dimension definitions, discussion of ambiguous cases, and regular meetings with annotators. This highlights the importance of careful task design, clear guidelines, and iterative calibration.

At the same time, the cognitively demanding task of repeatedly reading news articles and summaries

from the same set of models may have introduced certain biases. Although model identities were hidden and candidate summaries were presented in randomized order, some models produced recognizable output styles and characteristic summary lengths over repeated annotations. Table 6 shows that GPT-3.5 and German-Leo produced shorter summaries on average, while LLaMA-1B produced longer summaries with a wider length range. GPT-generated summaries performed consistently well (Appendix A.6), while other models showed recurring issues. Example summaries for the same news article from three different models are shown in Appendix A.7, along with their average annotator ratings across all eight dimensions. These recurring generation patterns may have enabled annotators to infer model identity and should therefore be considered in future evaluation designs.

Finally, the results also have implications for metric selection. Since *Content Coverage* and *Faithfulness* are the strongest predictors of human preference, German news summarization metrics should ideally capture both informational completeness and factual consistency. Metrics based mainly on lexical overlap or surface similarity may miss these aspects, especially for abstractive summaries that differ in wording from the source article.

6 Conclusion and Future Work

In this work, we introduced an expert-annotated dataset for German news summarization evaluation, combining multi-dimensional quality ratings with pairwise preference judgments. Unlike prior English-centered resources, our dataset focuses on German LLM-generated news summaries and supports analyses of annotation reliability and human preference. In our dataset, pairwise preferences showed higher agreement than dimension-level ratings, suggesting that relative comparisons may be easier to apply consistently in this setting.

Our analysis shows that reliability varies across dimensions. Annotators agreed more on linguistic and factual aspects, such as grammaticality, non-redundancy, and faithfulness, while discourse-level aspects, such as referential clarity and structure, were more subjective and showed lower agreement. We also found that content coverage and faithfulness are most strongly related to overall human preference. Pairwise preference judgments achieved higher agreement overall in our dataset, suggesting that relative comparisons may be easier to annotate than dimension-level ratings in this setting.

In future work, the dataset can support experiments with LLMs, including validation of LLM-as-a-judge approaches and fine-tuning with supervised learning, contrastive learning (Zhang et al., 2022), or preference-based optimization methods such as DPO (Rafailov et al., 2024). Because the dataset contains two complementary forms of human feedback, it may also be useful for developing summarization models that better align with human judgments.

Second, the dataset could be expanded through additional annotations, new domains, and broader summarization settings. Evaluating newer LLMs would also help identify emerging error patterns and determine whether the current quality dimensions remain sufficient or should be refined to reflect changes in model behavior.

Limitations

This work has several limitations. First, expert annotation is costly and time-intensive, which limits the size of the annotated dataset and the number of annotators involved. Second, the dataset is restricted to the German news domain, so the findings may not directly generalize to other domains, such as legal texts or medical summarization. Third, the source texts are relatively short news articles, and the results may not apply to longer documents or multi-document summarization settings.

The selection of LLMs used to generate the summaries represents an additional limitation. Although the dataset includes outputs from multiple models, it could be expanded by incorporating a broader range of models, including larger and more recent state-of-the-art systems. This would increase the diversity of the dataset.

Finally, the annotations collected in this study are limited to ratings across eight quality dimensions and pairwise preference judgments. Other annotation formats, such as error-span identification, hallucination-span annotation, error-type labels, or question-answering-based evaluation, were not explored.

Acknowledgments

We acknowledge ZEIT ONLINE for providing the news article data used in this work. This research was supported by the German Federal Ministry of Research, Technology and Space (BMFTR).

References

Ron Artstein and Massimo Poesio. 2008. "Survey Article: Inter-Coder Agreement for Computational Linguistics". *Computational Linguistics*, 34(4):555–596.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. "Re-evaluating Evaluation in Text Summarization". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

John M. Conroy and Hoa Trang Dang. 2008. "Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality". In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK. Coling 2008 Organizing Committee.

Alexander R. Fabbri, Wojciech KryõciÑski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Mingqi Gao and Xiaojun Wan. 2022. "DialSummEval: Revisiting Summarization Evaluation for Dialogues". In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Zhenyu Huang, Xianlai Chen, Yunbo Wang, Jincai Huang, and Xing Zhao. 2025. A survey on biomedical automatic text summarization with large language models. *Information Processing Management*, 62(5):104216.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025. CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization. *Journal of Artificial Intelligence Research*, 82:313–365.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024. "What's under the hood: Investigating Automatic Metrics on Meeting Summarization". In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6709–6723, Miami, Florida, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Uygar Kurt. 2026. Which Quantization Should I Use? A Unified Evaluation of llama.cpp Quantization on Llama-3.1-8B-Instruct. *Preprint*, arXiv:2601.14277.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. "UniSumEval: Towards Unified, Fine-grained, Multi-dimensional Summarization Evaluation for LLMs". In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. "From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge". In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

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

Junyuan Liu, Zhengyan Shi, and Aldo Lipani. 2024. "SummEQuAL: Summarization Evaluation via Question Answering using Large Language Models". In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 46–55, Bangkok, Thailand. Association for Computational Linguistics.

Mamta and Oana Cocarascu. 2025. FactEval: Evaluating the Robustness of Fact Verification Systems in the Era of Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10647–10660, Albuquerque, New Mexico. Association for Computational Linguistics.

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. "Towards Multi-dimensional Evaluation of LLM Summarization across Domains and Languages". In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. "Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics". In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M. Ingersoll. 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1):3–14.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Preprint*, arXiv:2305.18290.

Markus Reuter, Tobias Lingenberg, Ruta Liepina, Francesca Lagioia, Marco Lippi, Giovanni Sartor, Andrea Passerini, and Burcu Sayin. 2025. "Towards Reliable Retrieval in RAG Systems for Large Legal Datasets". In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 17–30, Suzhou, China. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. "QuestEval: Summarization Asks for Fact-based Evaluation". In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kanta Prasad Sharma, Mohd Shukri Ab Yajid, J. Gowrishankar, Rohini Mahajan, Anas Ratib Alsoud, Abhilasha Jadhav, and Devendra Singh. 2025. A Systematic Review on Text Summarization: Techniques, Challenges, Opportunities. *Expert Systems*, 42(4):e13833. E13833 EXSY-Aug-24-3783.R2.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. "Learning to Summarize from LLM-generated Feedback". In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 835–857, Albuquerque, New Mexico. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia B. Chilton, and Kathleen McKeown. 2025. STORYSUMM: Evaluating Faithfulness in Story Summarization. *Preprint*, arXiv:2407.06501.

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Zhuang Li. 2025. QQSUM: A Novel Task and Model of Quantitative Query-Focused Summarization for Review-Based Product Question Answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20810–20831, Vienna, Austria. Association for Computational Linguistics.

Tao Xie, Yuanyuan Kuang, Ying Tang, Jian Liao, and Yunong Yang. 2025. Using LLM-supported lecture summarization system to improve knowledge recall and student satisfaction. *Expert Syst. Appl.*, 269(C).

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. 2026. "An Empirical Study of LLM-as-a-Judge: How Design Choices Impact Evaluation Reliability". In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 167–176, San Diego, California, USA. Association for Computational Linguistics.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2025. A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. *ACM Comput. Surv.*, 57(11).

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022. "Contrastive Data and Learning for Natural Language Processing". In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *CoRR*, abs/1904.09675.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. *Preprint*, arXiv:2210.07197.

A Appendix

A.1 Distribution of News Article Topics

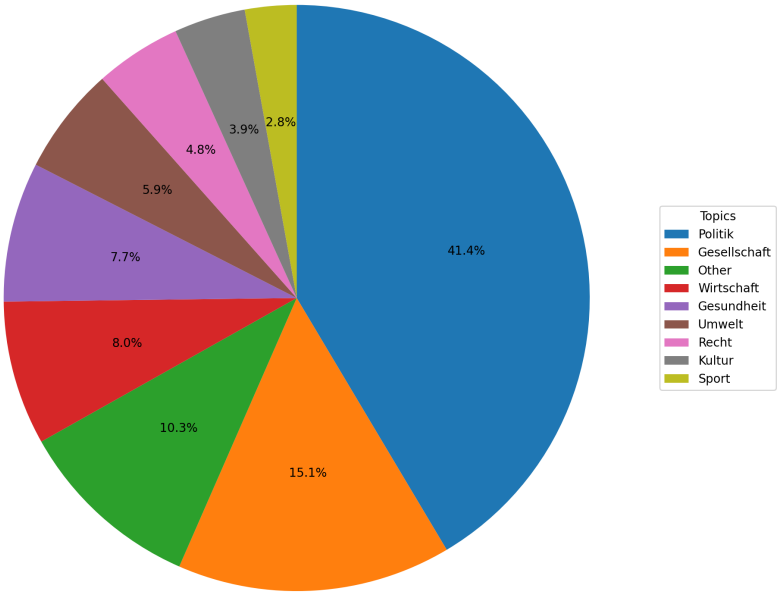

Figure 2: Distribution of news article topics in the dataset, based on categories automatically assigned using GPT-3.5-Turbo.

A.2 Annotation Guidelines

Annotators rated each summary on a scale from 1 (poor) to 7 (excellent). The following guidelines illustrate how the scale was interpreted during annotation.

Annotation Rating Guidelines

Grammaticality
7 – Excellent: No grammatical errors.
4 – Mixed: A few small errors, but understandable.
1 – Poor: Many errors; hard to read.

Non-redundancy
7 – Excellent: No repetition.
4 – Mixed: Some repetition, but not distracting.
1 – Poor: Lots of repetition; summary feels bloated.

Referential Clarity
7 – Excellent: All references are clear.
4 – Mixed: One or two unclear references.
1 – Poor: Several unclear references; confusing.

Focus
7 – Excellent: Every sentence is on-topic.
4 – Mixed: One off-topic or irrelevant detail.
1 – Poor: Several off-topic details.

Structure and Coherence
7 – Excellent: Flows smoothly; ideas connect logically.
4 – Mixed: Mostly coherent, but a small jump exists.
1 – Poor: Hard to follow; sentences do not connect.

Content Coverage
7 – Excellent: All key information included.
4 – Mixed: Misses one or two important details.
1 – Poor: Misses many important points.

Faithfulness

7 – Excellent: Fully factual, no hallucinations.
4 – Mixed: Minor factual imprecision.
1 – Poor: Several factual errors or misleading content.

Complexity
7 – Excellent: Clear, concise, natural flow; varied sentences.
4 – Mixed: Slightly too simple or wordy but understandable.
1 – Poor: Very simplistic or overly complex/confusing.

A.3 Annotation Interface

Figure 3 shows the web interface used for the annotation of summaries.

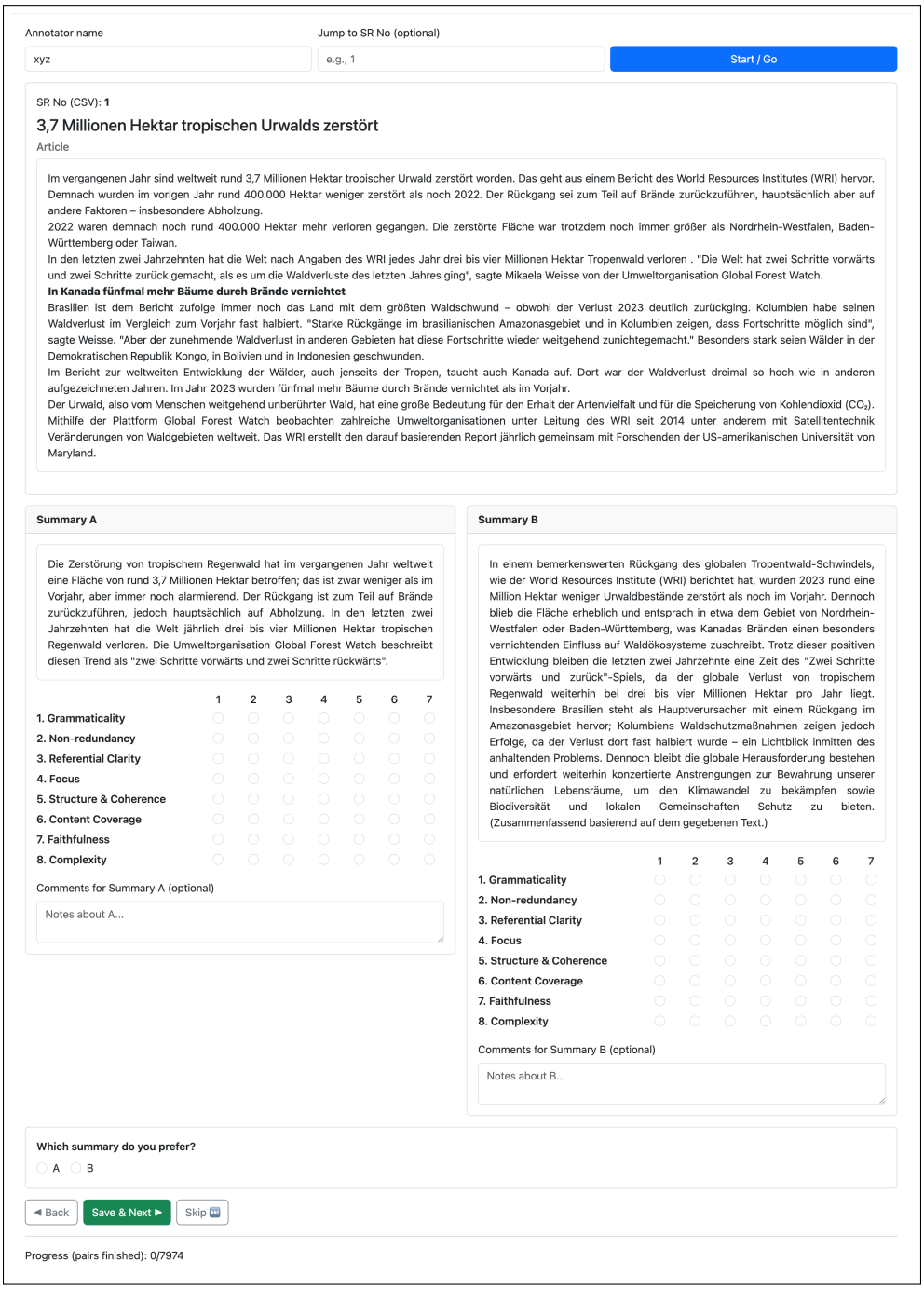

Figure 3: Screenshot of the annotation interface used for summary evaluation.

A.4 Prompt Used for Summary Generation

The following prompt was used to generate summaries with the language models. The placeholder `{{ARTICLE}}` was replaced with the full article text.

Prompt used for summarization

Du bist Redakteur einer großen Onlinezeitung. Deine Aufgabe ist es, einen Artikel zusammenzufassen. Konzentriere dich dabei auf die Kerninhalte des Artikels. Verfasse einen prägnanten ersten Satz ohne Nebensatz. Schreibe genau fünf Sätze.

Richte dich stilistisch nach folgenden Werten:

Wortschatz Reichhaltig und abwechslungsreich, teilweise gehoben. Mischung aus Alltagssprache, Fachjargon und spezifischen Begriffen.

Jargon Medien- und Podcast-Spezifika. Anglizismen und Fachbegriffe. Szenejargon.

Tonfall Kritisch, aber auch humorvoll und neugierig.

Satzbau Variabel: teils kurze, prägnante Sätze, teils längere Konstruktionen. Flüssig lesbar mit gutem Rhythmus.

Struktur und Aufbau Chronologisch erzählt, eingebettet in Reflexionen und Analysen. Wechsel zwischen Ereignissen und gesellschaftlicher Einordnung.

Gliederung Klarer Einstieg, Mitte und Schluss.

Einleitung Neugierig machend.

Zielgruppe Gebildete Leser mit Interesse an Kultur, Medien, Politik und gesellschaftlichem Diskurs.

Stilmittel Metaphern, bildhafte Vergleiche, Ironie, Sarkasmus, Aufzählungen und Humor.

Tempo und Rhythmus Dynamischer Wechsel zwischen schnellen Szenen und reflektierenden Passagen.

Flüssigkeit Sehr gut lesbar durch klar strukturierte Übergänge.

Dynamik Spannungsaufbau durch chronologische Erzählung und Reflexion.

Erledige diese Aufgabe für folgenden Artikel:
`"""ARTICLE"""`

A.5 Model-wise Agreement

In addition to computing agreement by evaluation dimension, we also examined whether annotation reliability differed across summary generation models. Table 8 reports Krippendorff's α for each model and dimension. This analysis helps identify whether certain models produced summaries that were easier or more difficult for annotators to evaluate consistently.

Dimension	German-Leo	GPT-3.5	LLaMA-1B	LLaMA-3B	Sauerkraut	Wiedervereinigung
Complexity	-0.071	-0.166	0.341	0.115	0.190	0.024
Content Coverage	0.486	0.099	0.503	0.471	0.462	0.438
Faithfulness	0.371	0.175	0.699	0.433	0.491	0.229
Focus	0.173	0.057	0.429	0.275	0.353	0.320
Grammaticality	0.558	0.223	0.552	0.636	0.328	0.135
Non-redundancy	0.441	0.341	0.802	0.566	0.617	0.212
Referential Clarity	0.193	0.052	0.239	0.340	0.015	0.117
Structure & Coherence	0.115	0.029	0.327	0.120	0.134	-0.006

Table 8: Model-wise Krippendorff's α across evaluation dimensions.

A.6 Model-wise Ratings

We computed model-wise mean ratings using the 1,470 item–model units that were evaluated by at least two annotators. As shown in Table 9, GPT-3.5 obtained the highest mean rating on six of the eight

dimensions. German-Leo obtained the highest mean ratings for Referential Clarity and Structure & Coherence. The difference between German-Leo and GPT-3.5 was small for Referential Clarity, whereas it was more pronounced for Structure & Coherence.

Dimension	GPT-3.5	LLaMA-1B	LLaMA-3B	German-Leo	Sauerkraut	Wiedervereinigung
Grammaticality	**6.696**	4.312	5.483	6.511	6.380	5.428
Non-redundancy	**6.846**	5.063	6.434	6.817	6.157	6.567
Referential Clarity	6.862	6.361	6.693	**6.883**	6.861	6.679
Focus	**6.796**	5.064	5.986	6.539	5.819	4.885
Structure & Coherence	5.957	4.634	5.883	**6.301**	5.816	6.106
Content Coverage	**6.160**	4.363	4.583	4.621	5.414	4.830
Faithfulness	**6.277**	3.742	4.466	5.922	5.196	3.978
Complexity	**6.547**	3.886	5.496	5.711	5.132	4.160

Table 9: Average human ratings per model on a 1–7 Likert scale, computed over the 1,470 item–model units annotated by at least two annotators. The highest value in each row is shown in bold.

Mean ratings alone do not reveal whether lower performance results from occasional low scores or from a broader shift toward moderately lower scores. To complement the unit-level means, Figure 4 shows the distribution of individual rating responses for each model and dimension.

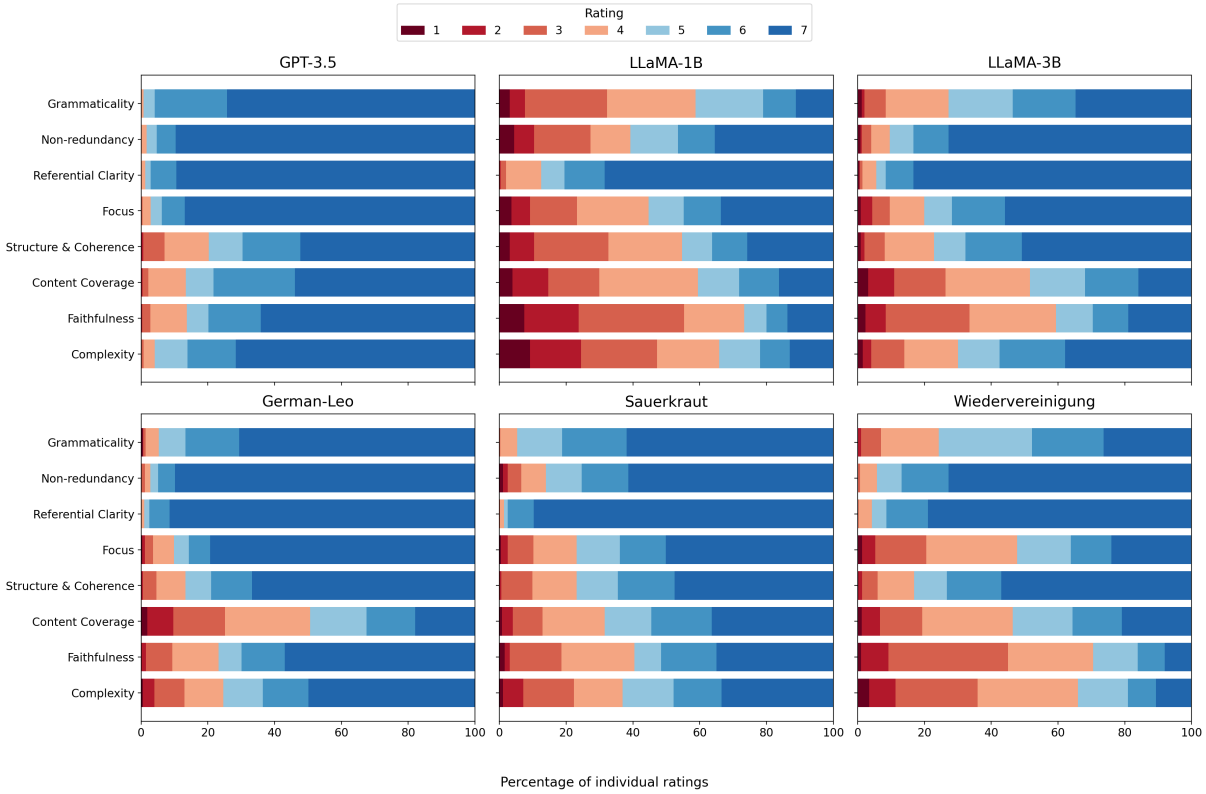

Figure 4: Distribution of human ratings by summarization model and evaluation dimension for the 1,470 item–model units annotated by at least two annotators. Each bar shows the percentage of scores from 1 to 7 for the corresponding model and dimension. The distributions are based on 3,720 individual annotation records, with one rating per dimension in each record. Dimensions are displayed in the same order as in Table 9.

A.7 Example Generated Summaries with Ratings

Table 10 shows the summaries generated by three of the six models for the same news article along with the averaged ratings among the annotators.

Generated Summaries	Scores (avg.)
Model: German-Leo Die Zerstörung von tropischem Regenwald hat im vergangenen Jahr weltweit eine Fläche von rund 3,7 Millionen Hektar betroffen; das ist zwar weniger als im Vorjahr, aber immer noch alarmierend. Der Rückgang ist zum Teil auf Brände zurückzuführen, jedoch hauptsächlich auf Abholzung. In den letzten zwei Jahrzehnten hat die Welt jährlich drei bis vier Millionen Hektar tropischen Regenwald verloren. Die Umweltorganisation Global Forest Watch beschreibt diesen Trend als "zwei Schritte vorwärts und zwei Schritte rückwärts".	Gram: 7.00 NonRed: 7.00 RefCl: 7.00 Focus: 7.00 Coh: 5.67 Cov: 5.67 Faith: 6.00 Complx: 6.67
Model: LLaMA-3B Hier ist ein prägnanter Erstsatz und fünf Sätze, die den Artikel zusammenfassen: "Im vergangenen Jahr sind weltweit rund 3,7 Millionen Hektar tropischer Urwald zerstört worden, was einem Rückgang von 400.000 Hektar gegenüber 2022 entspricht. Der Rückschlag ist teilweise auf Brände zurückzuführen, aber hauptsächlich auf Abholzung. Die Welt verliert jeden Jahr drei bis vier Millionen Hektar Tropenwald, was ein Zeichen für die zunehmende Bedrohung unserer Planeten ist. Brasilien bleibt das Land mit dem größten Waldschwund, aber Kolumbien hat seinen Verlust im Vergleich zum Vorjahr fast halbiert. Trotz dieser Fortschritte bleibt der Waldverlust ein großes Problem, das wir nicht ignorieren können."	Gram: 6.00 NonRed: 7.00 RefCl: 6.67 Focus: 7.00 Coh: 6.00 Cov: 6.67 Faith: 6.33 Complx: 6.33
Model: GPT-3.5 Im vergangenen Jahr wurden weltweit rund 3,7 Millionen Hektar tropischer Regenwald abgeholzt, wie aus einem Bericht des World Resources Institute (WRI) hervorgeht. Obwohl dies 400.000 Hektar weniger sind als 2022, ist die abgeholzte Fläche weiterhin größer als Bundesländer wie Nordrhein-Westfalen oder Baden-Württemberg. Während der Waldverlust in Brasilien und Kolumbien zurückging, ist er in Ländern wie der Demokratischen Republik Kongo, Bolivien und Indonesien gestiegen. Auch in Kanada war der Waldverlust aufgrund von Waldbränden historisch hoch.	Gram: 7.00 NonRed: 7.00 RefCl: 7.00 Focus: 6.33 Coh: 6.67 Cov: 6.33 Faith: 6.00 Complx: 7.00

Table 10: Example generated summaries with averaged human evaluation scores across eight dimensions.

A.8 Standard Deviation Analysis

Figure 5 shows the distribution of item-level rating dispersion across the 8 dimensions used.

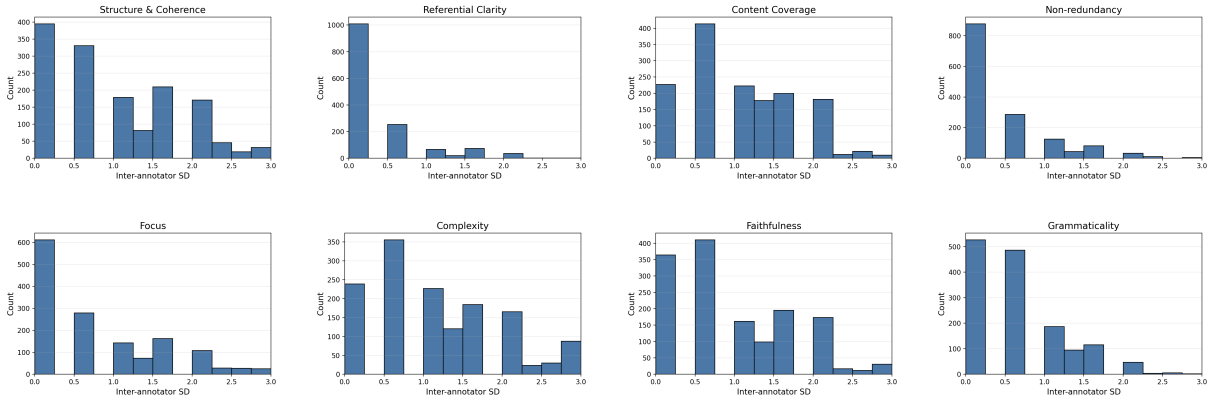

Figure 5: Distribution of item-level rating dispersion across evaluation dimensions. Each panel shows a histogram of the standard deviation of annotator scores for one dimension, computed over units with at least two annotations ($N = 1470$).

A.9 Supplementary Analyses

A.9.1 Effect of Summary Length

We examined the relationship between summary word count and each quality dimension using Spearman correlations. We then used regression models to adjust for differences between source articles and summarization models.

Longer summaries generally received lower ratings (Table 11), with the strongest negative correlations for *Complexity*, *Focus*, and *Non-redundancy*. After adjustment, length remained negatively related to seven dimensions, while *Content Coverage* showed a small positive relationship. Model differences remained significant for all eight dimensions after adjusting for length ($p < .001$), indicating that summary length alone did not explain the differences between models.

Dimension	ρ with length	Adjusted change/10 words
Grammaticality	-0.452	-0.034
Non-redundancy	-0.551	-0.097
Referential Clarity	-0.203	-0.009
Focus	-0.581	-0.080
Structure & Coherence	-0.294	-0.048
Content Coverage	-0.065	$+0.031$
Faithfulness	-0.511	-0.071
Complexity	-0.592	-0.091

Table 11: Relationship between summary length and ratings for the 1,470 item–model units annotated by at least two annotators. Spearman's ρ shows the unadjusted correlation with word count. Adjusted change gives the estimated rating difference per additional 10 words after accounting for source-article and model differences. All adjusted length relationships were statistically significant ($p < .001$).

A.9.2 Relations Between Quality Dimensions

To examine relationships among the eight quality dimensions, we computed pairwise Spearman correlations using ratings averaged within the 1,470 item–model units annotated by at least two annotators. All 28 correlations were positive and statistically significant after Benjamini–Hochberg correction, ranging from $\rho = 0.146$ to $\rho = 0.625$. The five strongest and three weakest associations are shown in Table 12.

Dimension 1	Dimension 2	ρ
Focus	Complexity	0.625
Faithfulness	Complexity	0.546
Focus	Faithfulness	0.533
Grammaticality	Faithfulness	0.506
Grammaticality	Complexity	0.476
Referential Clarity	Content Coverage	0.173
Non-redundancy	Content Coverage	0.150
Structure & Coherence	Content Coverage	0.146

Table 12: Five strongest and three weakest Spearman correlations between quality dimensions in the selected analysis subset ($N = 1{,}470$).