DETECT:
Determining Ease and Textual Clarity of German Text Simplifications

Maria Korobeynikova[UZH] **Alessia Battisti**[UZH] **Lukas Fischer**[UZH] **Yingqiang Gao**[†][UZH]

[UZH]Department of Computational Linguistics, University of Zurich, Switzerland

`{maria.korobeynikova, yingqiang.gao}@uzh.ch`

Abstract

Current evaluation of German automatic text simplification (ATS) relies on general-purpose metrics such as SARI, BLEU, and BERTScore, which insufficiently capture simplification quality in terms of *simplicity*, *meaning preservation*, and *fluency*. While specialized metrics like LENS have been developed for English, corresponding efforts for German have lagged behind due to the absence of human-annotated corpora. To close this gap, we introduce DETECT, the first German-specific metric that holistically evaluates ATS quality across all three dimensions of *simplicity*, *meaning preservation*, and *fluency*, and is trained entirely on synthetic large language model (LLM) responses. Our approach adapts the LENS framework to German and extends it with (i) a pipeline for generating synthetic quality scores via LLMs, enabling dataset creation without human annotation, and (ii) an LLM-based refinement step for aligning grading criteria with simplification requirements. To the best of our knowledge, we also construct the largest German human evaluation dataset for text simplification to validate our metric directly. Experimental results show that DETECT achieves substantially higher correlations with human judgments than widely used ATS metrics, with particularly strong gains in meaning preservation and fluency. Beyond ATS, our findings highlight both the potential and the limitations of LLMs for automatic evaluation and provide transferable guidelines for general language accessibility tasks.

🤗 Dataset   Code

1 Introduction

Automatic text simplification (ATS) aims to improve language accessibility for diverse target group users, including language learners, persons with cognitive disabilities, and persons with low literacy levels (Gao et al., 2025). While German ATS research has advanced with the emergence of multilingual LLMs and new parallel datasets (Ebling et al., 2022; Stodden et al., 2023), automatic evaluation relies on metrics like BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), and BERTScore (Zhang et al., 2020), which focus on Ngram overlap or embedding similarity rather than directly measuring the core criteria of simplification quality: *simplicity*, *meaning preservation*, and *fluency*, leading to weak correlations with human assessments (Alva-Manchego et al., 2021). Recent English-specific metrics like LENS (Maddela et al., 2023) have addressed this challenge by training models on human-rated quality scores based on core criteria, achieving stronger alignment with human judgments than traditional metrics. However, no equivalent metric exists for German (Stodden, 2024a).

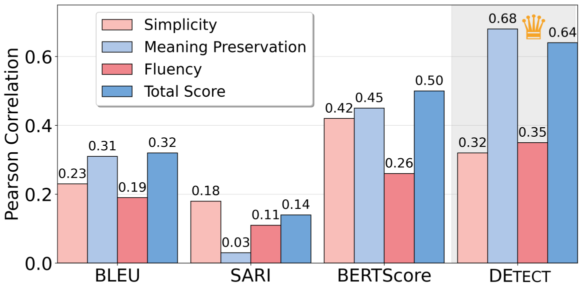


Figure 1: Correlation of standard ATS metrics and DETECT with human judgments across *simplicity*, *meaning preservation*, *fluency*, and the total score. Standard ATS metrics (BLEU, SARI, BERTScore) show limited alignment, while DETECT achieves higher correlations, particularly for *meaning preservation* and overall score.

To address this research gap, we introduce DETECT, the first German-specific ATS evaluation metric designed to directly assess text simplification quality across the three core dimensions mentioned above. Unlike LENS, which relies on human annotations, we leverage LLMs as synthetic eval-

---

[†] Corresponding author.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2852–2882
March 24-29, 2026 ©2026 Association for Computational Linguistics

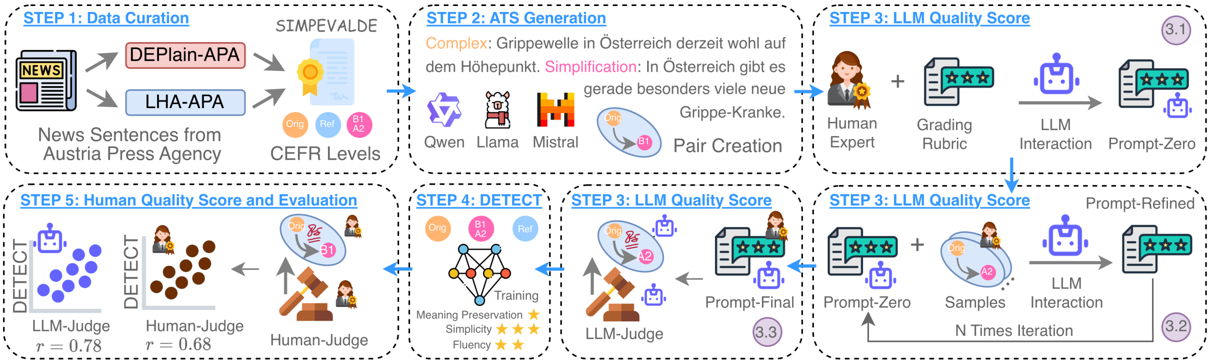


Figure 2: Pipeline of DETECT. The framework consists of four steps: (1) **Data curation**. Construction of SIMPEVALDE from Austrian Press Agency news sentences aligned across CEFR levels (DEPlaIn-APA and LHA-APA); (2) **ATS generation**. LLMs produce multiple simplifications for each complex sentence; (3) **LLM quality scoring**. Open-source LLMs are prompted as judges to assign ATS quality scores using a prompt developed through iterative human expert feedback on the original LENS prompt; (4) **DETECT**: A feed-forward neural network is trained to predict *simplicity*, *meaning preservation* and *fluency* from the synthetic LLM scores; (5) **Evaluation**. DETECT is validated against human quality judgments using a simplified RANK & RATE protocol (Maddela et al., 2023) and compared to standard ATS metrics.

uators to generate training data for this learnable metric. We construct SIMPEVALDE, a benchmark dataset combining existing German corpora with outputs from six ATS models, annotated with quality scores from three open-source LLMs. Additionally, we explore LLM-based refinement of grading criteria to improve evaluation consistency. Our work addresses three key research questions (RQs) in the ATS context:

RQ1 Can quality-based evaluation metrics be adapted for German without human-annotated data?

RQ2 Can LLMs provide reliable supervision for training learnable evaluation metrics?

RQ3 Can LLMs systematically refine evaluation rubrics to improve judgment consistency?

With DETECT, we aim at proposing general guidelines for language accessibility tasks when leveraging LLMs in a Human-in-the-Loop setting, thereby contributing to a deeper understanding of how synthetic evaluation pipelines can complement human judgment and foster the development of more reliable, cross-linguistic evaluation frameworks.

2 DETECT: Learnable Evaluation Metric for German Text Simplification

The methodology of DETECT is designed to closely follow the LENS framework (Maddela et al., 2023), but it replaces human-annotated data with LLM-generated synthetic data and refines the grading rubric through an iterative LLM-Human feedback loop. It constructs a synthetic supervision pipeline in four main steps (see Figure 2):

- STEP 1: We compile a parallel corpus of complex sentences and reference simplifications, denoted SIMPEVALDE.
- STEP 2: For each complex sentence in SIMPEVALDE, six automatic text simplifications are generated using different LLMs.
- STEP 3: Each generated simplification is scored along three dimensions: *simplicity*, *meaning preservation*, and *fluency*, by prompting three open-source LLMs as judges. These scores are then used to train DETECT. Specifically:
  - *Prompt-Zero* is created by prompting GPT-4o (OpenAI Team, 2024) to address identified limitations in the existing ATS instructions and is iteratively refined through a multi-round process using diagnostic examples and expert feedback to produce *Prompt-Final*;
  - Three distilled LLMs are prompted with *Prompt-Final* to provide quality scores for *simplicity*, *meaning preservation* and *fluency* for the entire dataset.
- STEP 4: The training set is used to fine-tune DETECT, implemented as a RoBERTa-based (Liu et al., 2019) feed-foward neural network (FFNN) to predict LLM-derived quality scores from the similarity between a text simplification, its complex source, and references.
- STEP 5: The test set of SIMPEVALDE is also

manually scored by human annotators according to the same criteria, following a simplified version of the RANK & RATE protocol introduced in LENS (Maddela et al., 2023).Then the trained DETECT model is evaluated on the test set by measuring its correlation with both human and LLM quality scores, as well as with established ATS metrics such as SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2020).

2.1 SIMPEVALDE

The SIMPEVALDE dataset is designed to mirror the structure of $\text{SIMPEVAL}_{\text{PAST}}$ (100 training sentences) and SIMPEVAL_{2022} (60 test sentences) introduced by LENS, where the target simplification level was not explicitly defined. Our manual analysis indicates that the simplifications in these datasets achieve an average Flesch Reading Ease (FRE; Flesch (1948)) score of 67, which has been reported to roughly correspond to CEFR level B1 (Council of Europe, 2020; Natova, 2021). While FRE scores cannot reliably predict CEFR levels, particularly outside of English, our analysis of A2- and B1-level sentences in the APA datasets suggests an approximate correspondence.

Building on this observation, we turn to existing German resources for ATS, which provide material at CEFR levels B1 and A2. Two variants are available: (i) LHA-APA (Spring et al., 2021), consisting of 1,828 automatically aligned Complex-Simple pairs at CEFR B1/A2, and (ii) the smaller DEPLAIN-APA (Stodden et al., 2023), which is manually aligned but lacks complex sentences. Given the importance of gold-standard references for training our metric, we attempt to merge DEPLAIN-APA with exact matches from LHA-APA. This procedure produces only 26 complete matches, and further attempts to expand alignments for the B1/A2 levels reveal many incorrect pairs.Manual inspection confirms that the APA corpus, originally constructed for document-level simplification, poses challenges for sentence-level alignment. Even correctly matched sentences often fail to correspond fully, with simplifications either omitting key information from the complex sentence or introducing unrelated content (see Appendix A.1 for examples).

To address this, we design a systematic filtering approach: adjusted BERTScore is used to identify high-similarity matches across the LHA-APA corpus, followed by a manual review of candidate pairs. While this strategy successfully eliminates low-quality pairs, i.e., 219 out of 532 randomly selected low-similarity cases prove to be entirely incorrect, it ultimately yields only 44 additional gold-standard pairs.

To reach a size of 160 rows, we therefore relax the multi-reference requirement by manually including single-reference matches (38 Complex-Simple pairs at CEFR B1 and 19 at CEFR A2) and incorporate test set pairs from DEPLAIN-APA, where both sentences fully preserve the same meaning.

The final dataset is then stratified across training and test sets based on match type and simplification strategy, following the LENS categories of *delete*, *split*, and *paraphrase* operations; the distributions can be found in Appendix A.3. We use an empirical algorithm to identify the simplification strategy, shown as follows:

Algorithm 1 Identify simplification strategy

```
if SentenceReductionRatio $> 1$ then
    Simplification Strategy $\leftarrow$ split
else if SentenceReductionRatio $\leq 1$ and WordReductionRatio $< 1$ then
    Simplification Strategy $\leftarrow$ delete
else
    Simplification Strategy $\leftarrow$ paraphrase
end if
```

We show two random examples for each simplification strategy identified by this algorithm in Appendix A.2 and present its statistical validity later in Section 3. We show the overall statistics of SIMPEVALDE and the comparison with those in the original LHA-APA and DEPLAIN-APA datasets in Appendix A.3 .

2.2 Generating Text Simplifications

To generate ATS outputs for evaluation, we follow the two-category approach of LENS by selecting six high-performing LLMs: general-purpose instruction-tuned models and task-specific fine-tuned models.

For the instruction-tuned models, we include both German-specific systems (LeoLM; Björn Plüster and Christoph Schuhmann (2023) and DiscoLlama; DiscoResearch Team (2024)) and multilingual variants (Qwen2-7B; Qwen Team (2024) and Llama3-8B; Llama Team (2024)). All instruction-tuned models are prompted using a German translation of the LENS prompt, augmented with few-shot examples balanced across three simplification strategies. This design follows prior findings that few-shot prompting outperforms zero-shot settings (Ryan et al., 2023).

For the task-specific models, we use mBART-DEPLAIN and mT5-DEPLAIN (Stodden et al., 2023), both fine-tuned on the DEPLAIN-APA dataset underlying SIMPEVALDE. This setup avoids the need for additional training.

All models are run with their default parameters, except for LeoLM, where adjustments to `temperature` and `top_p` were required to mitigate nonsensical generations observed during preliminary trials. Full configuration details and prompts are provided in Appendix B.

2.3 Quality Score Labeling with LLMs

Given the resource constraints that made full human evaluation infeasible, and motivated by the rationale to test the capacity of LLMs for this task, we rely on the LLM-as-a-Judge (Zheng et al., 2023) approach to produce quality score annotations for Complex-Simple pairs. Specifically, we adapt the LENS grading rubric into an LLM-compatible prompt through a Human-in-the-Loop (Chang et al., 2024) process with a high-capacity model. This prompt is then applied to three smaller distilled models (Distil-Llama-8B, Distil-Qwen-7B, and Zephyr-7B), whose outputs are averaged to yield the final *LLM-Judge* score.

The original LENS rubric employs five discrete rating levels (0, 25, 50, 75, 100) to jointly assess *simplicity*, *meaning preservation*, and *fluency*. However, this design poses several challenges for LLM-based evaluation: (1) vague definitions of core criteria such as "*simplicity*" and the associated simplification levels; (2) an implicit assumption of parallel progression across all dimensions, despite evidence that *simplicity* and *meaning preservation* often trade off against each other (Cripwell et al., 2024); and (3) hidden assumptions that require explicit clarification when translated into prompts for LLMs. These limitations hinder LLMs from consistently applying the rubric and exacerbate ambiguity in edge cases.

To overcome this, we propose evaluating each criterion separately and aggregating the results with a weighted formula for the total score, with S = *simplicity*, MP = *meaning preservation*, and F = *fluency*:

$$\text{Total Score} = \begin{cases} \min(\mathrm{MP}, \mathrm{S}, \mathrm{F}), & \text{if } \min(\mathrm{MP}, \mathrm{S}, \mathrm{F}) < 25; \\ 0.4\mathrm{MP} + 0.4\mathrm{S} + 0.2\mathrm{F}, & \text{otherwise.} \end{cases}$$

Following prior findings that *simplicity* and *meaning preservation* are more failure-prone than fluency (Cripwell et al., 2024), we assign double weight to the former relative to the latter. In addition, we ground our prompt optimization in concrete German guidelines for Easy Language (German: *Leichte Sprache*; Bredel and Maaß (2016)), incorporating explicit assumptions, for example, allowing models to assign any score between 0 and 100 rather than restricting them to five discrete levels. Prompt refinement is carried out with the assistance of GPT-4o (OpenAI Team, 2024) as an iterative process:

- **Initial Prompt Creation**: GPT-4o generates *Prompt-Zero* based on the original rubric and its identified limitations;
- **Iterative Refinement**: Four diagnostic examples (meaning omission, added complexity, fluent but incorrect meaning, and unchanged input/output) are used to reveal weaknesses of the current prompt, with human feedback guiding revisions and discussions for the optimized prompt;
- **Final Prompt Generation**: LLM scores and human feedback are integrated;
- **Few-shot Enhancement**: Six additional examples graded by GPT-4o are incorporated to distill knowledge into smaller models.

All prompts, examples used, GPT-4o and human feedback, and the final few-shot examples can be found in Appendix C. Our inspection of the refined prompt and the few-shot examples confirms that our approach successfully addresses most of the identified issues, providing clearer definitions of the evaluation criteria and improving the handling of edge cases.

To select the three models for the final prompt and compute the *LLM-Judge* score, we analyzed a set of DeepSeek distilled models (Guo et al., 2025), including Distil-Qwen-1.5B, Distil-Qwen-7B, Distil-Qwen-14B, Distil-Llama-8B, and Zephyr-7B (Tunstall et al., 2024). The analysis focused on correlations between the scores of smaller models ($\leq$8B) and the larger 14B model in order to assess whether the smaller models could serve as reliable proxies and whether they could produce sufficiently diverse outputs. Details of this analysis are provided in Appendix D.

Our results show that Distil-Qwen-7B achieves correlations between $r = 0.51$ and $r = 0.87$ with Distil-Qwen-14B across the individual evaluation criteria. From ($\leq$8B) models, Zephyr-7B exhibits lower diversity in its scores, it still reaches correlations between $r = 0.41$ and $r = 0.58$ with Distil-Qwen-7B. Across all models, the highest correlations are consistently observed for *meaning preservation* and the lowest for *fluency*.

Based on these findings, we selected Distil-Llama-8B, Distil-Qwen-7B, and Zephyr-7B to calculate the final *LLM-Judge* scores, aggregated as the arithmetic mean of their individual output quality scores.

2.4 DETECT Training and Validation

DETECT builds on the LENS architecture, which concatenates embeddings of the complex sentence, simplified output, and references, along with their dot products and differences. These representations are then fed into a neural network. We introduce three key modifications. First, the model is adapted to predict three separate scores, i.e., *simplicity*, *meaning preservation*, and *fluency*, instead of the single composite score used in LENS. Second, we incorporate German-specific embeddings WECHSEL (Minixhofer et al., 2022) in place of the multilingual RoBERTa-XML (Liu et al., 2019) encoder used by LENS. Third, to address the considerably smaller training dataset, we tune hyperparameters such as the learning rate, dropout, and hidden layer size to reduce overfitting.

We explore several DETECT variants with different architectural and training configurations, with each variant building on insights from the previous run, all trained on five epochs and evaluated on the validation set using Pearson correlation and Spearman rank. The final model is selected based on the best validation performance across these metrics. To support multi-criteria prediction, the loss function is defined as the arithmetic mean of the three output scores. We report the configurations of different DETECT variants in Table 9 in Appendix A.

2.5 Human and Automatic Evaluation

We evaluate the predictions of DETECT on the SIMPEVALDE test set against two reference measures: (1) averaged LLM quality scores (*LLM-Judge*), which capture direct model accuracy using the same criteria on which the model was trained, and (2) human quality scores (*Human-Judge*), which provide an independent benchmark based on human judgment.

Human scores were collected using a customized Google Sheets tool implementing a simplified RATE & RANK procedure (Maddela et al., 2023). Evaluators first classified each simplification by strategy (*split*, *paraphrase*, *delete*), then rated *simplicity*, *meaning preservation*, and *fluency* according to the same criteria used in the LLM-grading prompt. Finally, they verified score consistency via a separate ranking sheet (see Appendix E for details). Three voluntary native German-speaking experts independently graded all 360 test pairs.

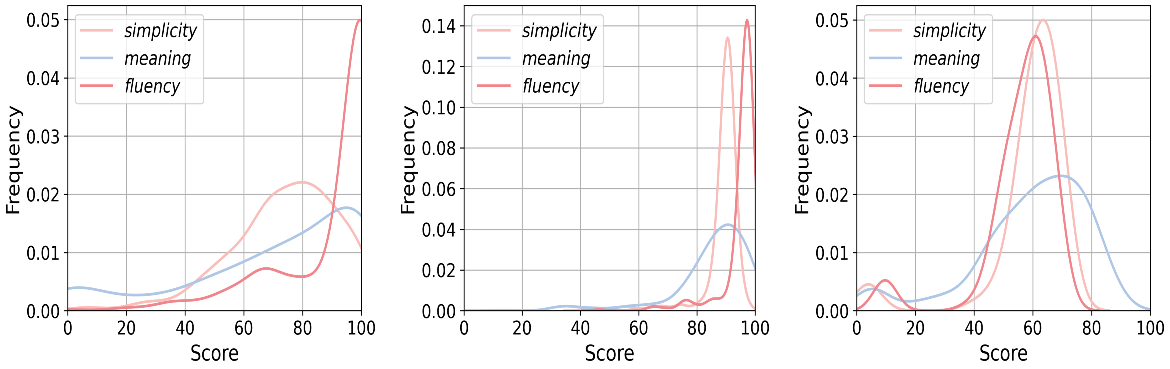
Finally, we benchmark DETECT against standard ATS evaluation metrics (i.e., BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), BERTScore Precision (Zhang et al., 2020)) by comparing their correlations with both *LLM-Judge* and *Human-Judge*.

3 Results

***Human-Judge* evaluation.** Three expert graders annotated all 360 test-set examples for *simplicity*, *meaning preservation*, and *fluency*. Inter-rater agreement is substantial, with Krippendorff's α (Krippendorff, 1970) of 0.75 for the total score, i.e., considerably higher than the $\alpha = 0.32$ reported for the SIMPEVAL$_{2022}$ grading in LENS. Agreement is strongest for *meaning preservation* ($\alpha = 0.78$), while *simplicity* and *fluency* showed lower levels ($\alpha = 0.37$ and $\alpha = 0.38$, respectively). For each criterion, the *Human-Judge* score is the mean of three graders' ratings, from which the total score is computed.

As shown in Figure 3, *meaning preservation* exhibits the widest spread, with only one-third of examples receiving the maximum score and a local peak at 0, whereas *fluency* is heavily right-skewed, with more than half of the examples achieving the maximum rating of 100. Correlation analysis shows that *meaning preservation* aligns most strongly with the total score ($r = 0.88$), followed by *fluency* ($r = 0.61$) and *simplicity* ($r = 0.56$). The weakest relationship is between *simplicity* and *meaning preservation* ($r = 0.25$), consistent with the findings of Cripwell et al. (2024).

***LLM-Judge* Evaluation.** The three *LLM-Judges* show lower inter-rater agreement ($\alpha = 0.54$) compared to human graders ($\alpha = 0.75$). As with human scoring, *meaning preservation* achieves the highest agreement ($\alpha = 0.60$), followed by *fluency*

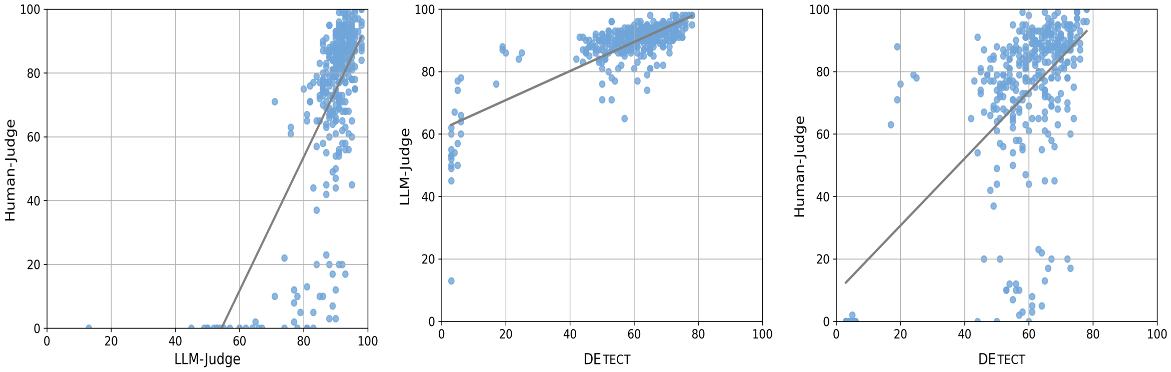


(a) *Human-Judge*

(b) *LLM-Judge*

(c) DETECT

Figure 3: Distributions of *Human-Judge*, *LLM-Judge* and DETECT by criterion



(a) *Human-Judge* vs. *LLM-Judge*

(b) *LLM-Judge* vs. DETECT

(c) *Human-Judge* vs. DETECT

Figure 4: Scatter plots of total *Human-Judge*, *LLM-Judge* and DETECT scores.

($\alpha = 0.45$) and *simplicity* ($\alpha = 0.41$). *LLM-Judge* score distributions shown in Figure 3 are narrower and more Gaussian-shaped than human ratings, with compressed interquartile ranges across all criteria, as well as higher median values for *simplicity* and *meaning preservation*.

Correlation analysis shown in Table 1 shows that *meaning preservation* aligns most strongly with human ratings ($r = 0.77$), followed by *fluency* ($r = 0.45$), and *simplicity* ($r = 0.28$). Despite the lower two scores, the total score achieves a correlation of $r = 0.72$. Figure 4 shows that while *Human-Judge* and *LLM-Judge* scores correlate strongly when both assign high total scores, a distinct cluster emerges where *LLM-Judge* assigns scores between 40 and 100 but human ratings remain below 20. Additionally, unlike in human scores, all intra-criteria correlations for *LLM-Judge* are high from $r = 0.82$ between *simplicity* and *meaning preservation* to $r = 0.94$ between *simplicity* and *fluency*, in contrast to the weak correlation of $r = 0.25$ observed between *simplicity* and *meaning preservation* in the former.

DETECT training and model selection. Table 9 in Appendix A reports the validation results for all DETECT variants, among which the DETECT_multi_wechsel_reduced model achieves the best performance, with a Spearman rank correlation of 0.68 and a Pearson correlation of 0.87, and is therefore selected as the final model for evaluation. Following LENS, predicted scores are rescaled to the range [0, 100] by mapping them to the percentage of the area under a normal curve with a mean of 0 and a standard deviation of 1.

Table 1: Person correlations (r) between DETECT, *Human(-Judge)* and *LLM(-Judge)*.

Comparison	***Simplicity***	***Meaning P.***	***Fluency***	**Total Score**
Human vs. *LLM*	0.28	0.77	0.45	0.72
Human vs. DETECT	0.32	0.68	0.35	0.64
DETECT vs. *LLM*	0.82	0.80	0.78	0.80

Comparison of DETECT with *LLM-Judge*. As shown in Table 1, DETECT exhibits strong correlations with *LLM-Judge* scores (used for training the model) across all criteria: $r = 0.78$ for *fluency*, $r = 0.80$ for *meaning preservation*, $r = 0.82$ for *simplicity*, and $r = 0.80$ overall.Visual inspection of the total-score scatter plot in Figure 4 shows distinct low- and high-score clusters of DETECT vs. *LLM-Judge* scores, resultant of the bi-modal distribution of DETECT scores seen in Figure 3; the fitted correlation line intersects both clusters, showing agreement between the two scorers at both ends of the scale. However, agreement differs in concentration between these regions: instances assigned scores of 20 or lower by DETECT correspond to a range of *LLM-Judge* scores betwen 15 and 90, whereas instances assigned scores of 60 or higher by DETECT consistently receive *LLM-Judge* scores of 70 or above, despite a larger number of observations in this region.

Comparison of DETECT with *Human-Judge*. The correlation between DETECT predictions and human scores, reflecting the real performance of the model, mirrors the relationships observed between *LLM-Judge* and human outputs. *Meaning preservation* achieves the strongest correlation with human scores at $r = 0.68$, while *simplicity* shows the weakest performance at $r = 0.32$, and *fluency* demonstrates a similar correlation at $r = 0.35$. The combined total score shows strong correlation with human judgment at $r = 0.64$. The scatter plot in Figure 4 shows overall agreement of high scores assigned by DETECT with human scores; however, the majority of the instances with low total scores as graded by humans are not graded as such by DETECT with only a few exceptions. Conversely, instances where DETECT assigns a low score while *Human-Judge* does not are rare. Scatter plots for individual criteria (See Figure 9 in Appendix E) demonstrate that *meaning preservation* scores, while having the highest correlation with human judgment, exhibit the markedly highest number of instances of low human score and high DETECT score.

Comparison with Standard ATS Metrics. Table 2 shows that DETECT substantially outperforms standard ATS metrics across nearly all evaluation dimensions, achieving the strongest correlations with human judgments in almost every category, as well as for the total score. The advantage is most pronounced for *meaning preservation*, where DETECT reaches $r = 0.68$, clearly surpassing BERTScore ($r = 0.48$), BLEU ($r = 0.31$), and SARI ($r = 0.04$). For *fluency*, DETECT again leads with $r = 0.35$, compared to BERTScore ($r = 0.31$), BLEU ($r = 0.19$), and SARI ($r = 0.11$). *Simplicity* represents the weakest relative dimension, where DETECT achieves a lower correlation ($r = 0.32$) than BERTScore ($r = 0.42$), still outperforming BLEU ($r = 0.23$) and SARI ($r = 0.18$). Overall, DETECT achieves a correlation of $r = 0.64$ with human judgments, outperforming BERTScore ($r = 0.55$), BLEU ($r = 0.32$), and SARI ($r = 0.14$). Further analysis (see Table 16 in Appendix E) confirms that DETECT achieves the highest correlations with human judgments across all simplification strategies (*paraphrase*, *delete*, *split*) and all criteria. Among these, split-based simplifications yield the strongest correlations for every metric, including DETECT. The model does, however, show some degradation relative to its *LLM-Judge* training data, particularly for delete- and paraphrase-based simplifications in *meaning preservation*, and for delete-based *fluency*.

Table 2: Pearson correlations (r) between automatic metrics and human evaluation scores across criteria. Best values are highlighted in **bold**, second-best values are underlined.

Metric	***Simplicity***	***Meaning P.***	***Fluency***	**Total Score**
BLEU	0.23	0.31	0.19	0.32
SARI	0.18	0.04	0.11	0.14
BERTScore	**0.42**	0.48	0.31	0.55
DETECT[*]	0.32	**0.68**	**0.35**	**0.64**

4 Discussion

DETECT exhibits strong alignment trained on LLM-generated quality scores, demonstrating effective adaptation even under the constraints of smaller datasets and fewer references compared to the original LENS framework:

RQ1: Metric Performance and Adaptability. Performance varies across evaluation criteria, with *meaning preservation* achieving the highest correlation ($r = 0.68$), followed by *fluency* ($r = 0.35$) and *simplicity* ($r = 0.32$). Despite an over-representation of paraphrase-based simplifications in the dataset, the metric performs particularly well

---

[*] For metrics predicting separate scores for *simplicity*, *meaning preservation* and *simplicity*, correlation is shown between each criteria and its metric, e.g., human *simplicity* score vs. DETECT *simplicity* score.

on split-based cases and worst on paraphrasing, likely reflecting the greater semantic complexity of paraphrase evaluation under reduced model capacity; a more balanced dataset could further substantiate this interpretation. Overall, DETECT consistently outperforms standard ATS metrics for *meaning preservation* and *fluency* across simplification strategies. We observe that the predictions mapped against human scores tend to form high- and low-quality clusters, showing stronger agreement with human judgments for high-quality outputs, while predictions for low-quality outputs are less reliable. Moreover, because DETECT rarely assigns intermediate scores, it is not well suited for fine-grained ranking among similarly scored candidates. Additionally, because the outputs are rescaled for comparability, they no longer correspond directly to the discrete levels of the original rubric, which reduces interpretability. Finally, the model is trained and evaluated exclusively on news-domain data, leaving its generalization to other domains unclear. Future work should explore adapting the model on data from other domains, as well as incorporating more intermediate-quality outputs and a more balanced distribution of simplification strategies, to improve fine-grained ranking, interpretability, and generalization.

RQ2: LLM-Based Annotation Reliability. LLM-generated quality scores provide reasonably reliable supervision signals depending on criterion, showing correlations with human judgments ranging from $r = 0.77$ for *meaning preservation* to $r = 0.28$ for *simplicity*. Error analysis highlights several limitations: (1) instability across repeated evaluations of identical inputs; (2) occasional German-specific misinterpretations, likely due to the underlying models being trained primarily on English and Chinese data (Guo et al., 2025); and (3) incomplete adherence to grading instructions, particularly regarding the penalization of hallucinated content. Moreover, LLMs exhibit artificially high intra-criterion correlations ($r > 0.8$), in contrast to human evaluators ($r = 0.25$) between *simplicity* and *meaning preservation*, suggesting limited sensitivity to the independence of criteria, even when providing seemingly coherent rationales. The reliance on distilled models may have amplified these issues, indicating that larger and more capable models could offer more stable and instruction-following behavior, thereby improving the reliability of synthetic data.

RQ3: Rubric Refinement Impact. The inter-rater agreement on the total score in our experiment with the reviewers using *Prompt-Final* ($\alpha = 0.75$) is substantially higher than the one reported in LENS with reviewers using the original criteria that the LLM was instructed to improve upon. ($\alpha = 0.32$). While these figures are not directly comparable due to the different sizes of the dataset, our observed improvement in the scoring done by LLMs during the prompt development phase, paired with observed integrations of the identified issues that the LLM was instructed to incorporate in *Prompt-Final* provides indication that LLMs can play a role in enhancing rubric clarity for both human and machine evaluators. Nonetheless, the analysis of LLM prompt feedback highlights remaining ambiguities, particularly in defining what qualifies as “important information” in deletion-based simplifications and in determining how to assess one criterion when others fail completely, pointing to areas where even refined rubrics require further clarification. An important limitation in evaluating our approach is that to reduce the computational load, the LLMs that were used for *LLM-Judge* are considerably smaller than the model used to develop *Prompt-Final*. Given known prompt sensitivity across model scales, future work should assess whether rubric refinements derived from larger models consistently improve grading quality for smaller architectures.

5 Related Work

ATS aims to transform complex text into simpler variants while preserving meaning and grammaticality (Alva-Manchego et al., 2021). Originally rule-based, ATS progressed through statistical machine translation (Alva-Manchego et al., 2020) and neural sequence-to-sequence models (Hochreiter and Schmidhuber, 1997), with transformer architectures becoming dominant for both sentence- and document-level simplification (Nisioi et al., 2017; Zhang and Lapata, 2017; Gonzales et al., 2021). Recent advances in LLMs enable zero- and few-shot simplification, which often outperforms task-specific models on English benchmarks (Kew et al., 2023; Maddela et al., 2023). In German, ATS work ranges from rule-based systems implementing *Leichte Sprache* guidelines (Suter et al., 2016) to neural approaches, with mBART-based adaptations prevailing as the state-of-the-art (Gonzales et al., 2021; Anschütz et al., 2023; Carrer et al., 2024;

Stodden, 2024b). Emerging multilingual LLMs such as BLOOM (Scao et al., 2022) show promise for German in zero- and few-shot regimes (Stodden, 2024b).

Sentence-level ATS requires parallel corpora of complex-simple pairs obtained through manual or automatic alignment (Nikolov and Hahnloser, 2019). Such resources often involve many-to-many mappings that complicate supervision (Ebling et al., 2022; Stodden et al., 2023). A central stream of work relies on the Austrian Press Agency (APA) news with CEFR B1/A2 targets, including APA-RST, APA-LHA, and DEPLAIN variants (Säuberli et al., 2020; Spring et al., 2021; Stodden et al., 2023; Hewett, 2023). Other corpora include the Simple German Corpus targeting *Leichte Sprache* (Toborek et al., 2023), as well as document-level collections such as *20 Minuten* and *Klexikon* (Gonzales et al., 2021; Aumiller and Gertz, 2022). Consolidation efforts like De-Lite (Jablotschkin et al., 2024) aim to unify and deduplicate sources while normalizing complexity labels. Web-derived datasets provide larger coverage but are heterogeneous, whereas professionally simplified texts are of higher quality but less accessible (Battisti et al., 2020; Ebling et al., 2022).

Evaluation of ATS typically combines human and automated assessment. Human evaluation focuses on *simplicity*, *meaning preservation*, and *fluency* (Alva-Manchego et al., 2020), though inter-rater reliability remains fragile, particularly for simplicity under absolute scales (Stodden, 2021; Alva-Manchego et al., 2021). Furthermore, the definition of "simplicity" remains highly audience-dependent, shaped by goals in accessibility, domain adaptation, or language learning (Stajner, 2021; Garimella et al., 2022). Automatic metrics are widely used despite their shortcomings. BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) rely on N-gram overlap, while BERTScore (Zhang et al., 2020) leverages contextual embeddings but may favor conservative or ungrammatical outputs. Readability measures such as Flesch Reading Ease (FRE; Flesch (1948)) and Flesch-Kincaid Grade Level (FKGL; Kincaid et al. (1975)) are often reported but assess *simplicity* only. Structure-aware or reference-free approaches such as SAMSA (Sulem et al., 2018), QuestEval (Scialom et al., 2021), and MeaningBERT (Beauchemin et al., 2023) target structural edits or semantic preservation. Recent learnable metrics trained on human judgments, such as LENS (Maddela et al., 2023) that models continuous quality scoring based on scores grading *simplicity*, *meaning preservation*, and *fluency* directly. For German, EASSE-DE (Stodden, 2024a) adapts tokenization and metric choices, though LENS-style learnable metrics was not applied. Parallel research has also modeled German sentence complexity directly from human labels (Naderi et al., 2019; Mohtaj et al., 2022b,a).

Recent work further explores LLMs for evaluation and grading. Instruction-tuning and reinforcement learning from human feedback (RLHF) align models with instructions and preferences (Wei et al., 2022a; Ouyang et al., 2022; Touvron et al., 2023), while distillation transfers these capabilities to smaller, more cost-efficient models (Li et al., 2024). Prompt engineering, including example-based prompts (Zhou et al., 2023), chain-of-thought prompting (Wei et al., 2022b), and automatic prompt or rubric induction (Honovich et al., 2023), can substantially improve zero- and few-shot evaluation (Gupta et al., 2024). Within the paradigm of LLM-as-a-Judge (Zheng et al., 2023), structured prompting methods such as G-Eval (Liu et al., 2023) increase agreement with expert judgments and scale effectively to complex, multi-criteria assessments (Gu et al., 2024).

6 Conclusion

We introduced DETECT, the first German-specific learnable metric for evaluating text simplification across *simplicity*, *meaning preservation*, and *fluency*. Developed by adapting the English LENS framework with synthetic LLM generated quality scores, DETECT fills the research gap in German text simplification evaluation and consistently outperforms BLEU, SARI, and BERTScore, with particularly strong results in *meaning preservation*. The refined grading rubric derived from LLM feedback also improved inter-rater agreement and reduced the manual effort required for supervision, showing that rubric refinement can enhance consistency for both human and automated evaluation. While some challenges remain, such as confounding factors, LLM instability, linguistic ambiguities specific to German, and residual rubric uncertainty, our study demonstrates the feasibility of creating learnable evaluation metrics through human guided synthetic supervision and provides a scalable foundation for future research on text simplification assessment for language accessibility.

Limitations

Our work has several limitations. DETECT relies on synthetic supervision from LLM-generated annotations, which can be unstable across runs, conflate criteria, and occasionally misinterpret German-specific constructions. The training and the evaluation is restricted to the news domain, limiting generalizability to other genres, such as educational or medical texts. In addition, the model tends to cluster outputs into broad quality groups rather than providing fine-grained rankings, and the score rescaling procedure reduces interpretability relative to the original rubric. Finally, this metric concerns only sentence-level simplification and is not directly applicable to document-level simplification. Our future work would therefore explore more robust annotator models, broader domain coverage, and rubric refinements that better balance interpretability with predictive power.

Acknowledgments

This work was supported by the Swiss Innovation Agency Innosuisse, Flagship Inclusive Information and Communication Technology (IICT), funding no. PFFS-21-47. We thank Sarah Ebling, Annette Rios, Patricia Scheurer, and Anja Ryser for their valuable contributions to the study.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German Dataset for Joint Summarization and Simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

David Beauchemin, Horacio Saggion, and Roger Khoury. 2023. MeaningBERT: Assessing Meaning Preservation between Sentences. *Frontiers in Artificial Intelligence*, 6:1223924.

Björn Plüster and Christoph Schuhmann. 2023. LAION LeoLM: Linguistically Enhanced Open Language Model. `https://huggingface.co/LeoLM/leo-mistral-hessianai-7b-chat`. Accessed: 2025-04-04.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*. Duden.

Luisa Carrer, Andreas Säuberli, Martin Kappus, and Sarah Ebling. 2024. Towards Holistic Human Evaluation of Automatic Text Simplification. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 71–80.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. 2022. mBART: Multidimensional Monotone BART. *Bayesian Analysis*, 17(2).

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating Document Simplification: On the Importance of Separately Assessing Simplicity and Meaning Preservation. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)@ LREC-COLING 2024*, pages 1–14.

DiscoResearch Team. 2024. Continued Pretraining on Llama3 8B for German. `https://huggingface.co/DiscoResearch/Llama3-DiscoLeo-Instruct-8B-v0.1`. Accessed: 2025-09-25.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic Text Simplification for German. *Frontiers in Communication*, 7:706718.

Rudolf Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Yingqiang Gao, Kaede Johnson, David Froehlich, Luisa Carrer, and Sarah Ebling. 2025. Evaluating the Effectiveness of Direct Preference Optimization for Personalizing German Automatic Text Simplifications for Persons with Intellectual Disabilities. *arXiv preprint arXiv:2507.01479*.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text Simplification for Legal Domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Annette Rios Gonzales, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A New Dataset and Efficient Baselines for Document-Level Text Simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Rahul Madhavan, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, and Saravan Rajmohan. 2024. CARMO: Dynamic Criteria Generation for Context-Aware Reward Modelling. *arXiv preprint arXiv:2410.21545*.

Freya Hewett. 2023. APA-RST: A Text Simplification Corpus with RST Annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 173–179.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Or Honovich, Uri Shaham, Samuel Bowman, and Omer Levy. 2023. Instruction Induction: From Few Examples to Natural Language Task Descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1935–1952.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-Lite-A New Corpus of Easy German: Compilation, Exploration, Analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, LÃl'lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, TimothÃl'e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.

J. Peter Kincaid, Richard P. Fishburne, Robert L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report 8-75, Naval Technical Training Command Millington TN Research Branch.

Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and psychological measurement*, 30(1):61–70.

Yiheng Li, Yulong Qian, Yuxuan Zhang, Yihan Liu, Junyi Yin, and Jiaguang Sun. 2024. A Survey on Knowledge Distillation and its Application in Large Language Models. *arXiv preprint arXiv:2402.13116*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Llama Team. 2024. Llama 3: Open foundation and instruction models. `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`. Accessed: 2025-04-26.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A Learnable Evaluation Metric for Text Simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective Initialization of Subword Embeddings for Cross-lingual Transfer of Monolingual Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022a. Overview of the GermEval 2022 Shared Task on Text Complexity Sssessment of German Text. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 1–9.

Salar Mohtaj, Babak Naderi, Sebastian Möller, Faraz Maschhur, Chuyang Wu, and Max Reinhard. 2022b. A Transfer Learning Based Model for Text Readability Assessment in German. *arXiv preprint arXiv:2207.06265*.

Babak Naderi, Sara Mohtaj, Kristof Ensikat, and Sebastian Möller. 2019. Subjective Assessment of Text Complexity: A Dataset for German Language. *arXiv preprint arXiv:1904.07733*.

Ivanka Natova. 2021. Estimating CEFR Reading Comprehension Text Complexity. *The Language Learning Journal*, 49(6):699–710.

Nikola I Nikolov and Richard Hahnloser. 2019. Large-Scale Hierarchical Alignment for Data-driven Text Rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 844–853.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring Neural Text Simplification Models. In *Proceedings of The 55th Annual Meeting of The Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

OpenAI Team. 2024. GPT-4o: Omni-modal AI Model. `https://openai.com/index/hello-gpt-4o/`. Large multimodal language model.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Qwen Team. 2024. Qwen2: The Next Generation of Qwen Models. `https://huggingface.co/Qwen/Qwen2-7B-Instruct`. Accessed: 2025-04-26.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting Non-English Text Simplification: A Unified Multilingual Benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927.

Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking Data-driven Automatic Text Simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with Reading Difficulties (READI)*, pages 41–48.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Thomas Scialom, Louis Martin, Jacopo Staiano, Eric Villemonte de La Clergerie, and Benoît Sagot. 2021. Rethinking Automatic Evaluation in Sentence Simplification. *arXiv preprint arXiv:2104.07560*.

Nicolas Spring, Annette Rios Gonzales, and Sarah Ebling. 2021. Exploring German Multi-Level Text Simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349.

Sanja Stajner. 2021. Automatic Text Simplification for Social Good: Progress and Challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Regina Stodden. 2021. When the Scale is Unclear-Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*.

Regina Stodden. 2024a. EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 107–116.

Regina Stodden. 2024b. Reproduction of German Text Simplification Systems. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 1–15.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic Structural Evaluation for Text Simplification. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 685–696.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based Automatic Text Simplification for German. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2023. A New Aligned Simple German Corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11393–11412.

Hugo Touvron, Louis Martin, Kevin Stone, and 1 others. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct Distillation of LM Alignment. In *First Conference on Language Modeling (COLM 2024)*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Fine-tuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations (ICLR 2022)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanzhi Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR 2020)*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

A The SimpEvalDE Dataset

A.1 Inappropriate Training Examples

Table 3 shows text simplifications we have excluded as gold-standard training samples.

Complex	Simplification	Issue
Insgesamt mussten bereits mehr als 500 Menschen behandelt werden, neun lagen auf der **Intensivstation**. *(In total, more than 500 people had to be treated, nine were in intensive care.)*	Hunderte Menschen wurden **in Sicherheit gebracht**. Mehr als 500 Menschen mussten von Ärzten behandelt werden. *(Hundreds of people were brought to safety. More than 500 people had to be treated by doctors.)*	The simplification is unfaithful to the complex text, introducing fabricated information about people being brought to safety.
Rund 15 Jahre zuvor hatte es in Indien nur noch etwa **1,400 Exemplare gegeben**. *(Around 15 years earlier, there had been only about 1,400 specimens in India.)*	Vor rund 15 Jahren gab es in Indien nur noch **1,400 Tiger**. *(Around 15 years ago, there were only 1,400 tigers left in India.)*	The simplification introduces extra hallucinated information about tigers that cannot be inferred from the complex sentence.

Table 3: Example simplifications from the APA-LHA dataset that are correctly aligned to the complex texts, but are unfaithful or add extra information.

A.2 Text Simplification Strategies

Table 4 gives examples of text simplifications using different strategies (*delete*, *split*, *paraphrase*).

A.3 Statistics of SimpEvalDE

Table 5 and Table 6 present the lexical statistics of training and test set of SimpEvalDE, respectively.

Table 7 compares the lexical measures of APA-LHA and DEplain-APA dataset.

Table 8 shows the simplification strategies in the training and test set of SimpEvalDE.

B Configurations and Prompt Details

We select mainstream open-source instruction-tuned and task-specific fine-tuned LLMs and their German variants as the backbone of ATS models.

B.1 Instruction-Tuned Models

LeoLM Björn Plüster and Christoph Schuhmann (2023)

- Model: German-specific Mistral (Jiang et al., 2023) fine-tuned on diverse German instruction-following data.
- Configuration: `top_p` = 0.95, `temperature` = 0.6 (adjusted due to repetition issues).

Table 4: Examples of complex-simple pairs with their assigned simplification strategies, determined using the holistic Algorithm 1. In the last case, the algorithm assigns *delete*, while manual verification suggests *paraphrase*; both assignments are considered acceptable.

Complex	Simplification	Strategy
Das hat Bildungsminister Heinz Faßmann (ÖVP) bei einer Pressekonferenz am Freitag bekanntgegeben. *(Education Minister Heinz Faßmann (ÖVP) announced this at a press conference on Friday.)*	Das hat der Bildungs-Minister Heinz Faßmann gesagt. *(Education Minister Heinz Faßmann said this.)*	Delete
Grippewelle in Österreich derzeit wohl auf dem Höhepunkt. *(Flu wave in Austria currently probably at its peak.)*	In Österreich gibt es gerade besonders viele neue Grippe-Kranke. *(In Austria there are currently particularly many new flu patients.)*	Paraphrase
Wenn die Temperaturen hoch sind, die Sonne viel scheint und es wenig regnet produzieren die Algen in dem See besonders viel Beta Carotin - und das Wasser wird pink. *(When temperatures are high, the sun shines a lot and it rains little, the algae in the lake produce particularly much beta carotene - and the water turns pink.)*	Wenn die Sonne viel scheint und es heißwird, produzieren die Algen den Farbstoff Beta Carotin. Durch diesen Farbstoff und dem vielen Salz wird das Wasser dann pink. *(When the sun shines a lot and it gets hot, the algae produce the pigment beta carotene. Through this pigment and the much salt, the water then turns pink.)*	Split
Man kann nur an die Menschen appellieren, sich keine Übermengen an Medikamenten verschreiben zu lassen. *(One can only appeal to people not to have excessive amounts of medication prescribed.)*	Man soll sie sich wie immer ganz normal verschreiben lassen. *(One should have them prescribed normally as always.)*	Delete/ Paraphrase

- HF Endpoint: 🤗 `LeoLM/leo-mistral-hessianai-7b-chat`

DiscoLlama DiscoResearch Team (2024)

- Model: German-specific instruction-tuned variant of Llama 3 (Grattafiori et al., 2024) designed for high-quality German text generation.
- Configuration: Default settings.
- HF Endpoint: 🤗 `DiscoResearch/Llama3-DiscoLeo-Instruct-8B-v0.1`

Qwen-7B Qwen Team (2024)

- Model: A multilingual instruction-tuned Qwen model.
- Configuration: Default settings.
- HF Endpoint: 🤗 `Qwen/Qwen2-7B-Instruct`

Llama-8B Llama Team (2024)

- Model: A multilingual instruction-tuned Llama model.
- Configuration: Default settings.
- HF Endpoint: 🤗 `meta-Llama/Meta-Llama-3-8B-Instruct`

Table 5: Statistics of lexical measures for complex texts, CEFR B1 and A2 simplifications in the training set of SimpEvalDE. Q1, Q2, Q3 denote the first quartile, the median, and the third quartile, respectively.

Statistic	Complex		B1			A2		
	num_words	**avg_word_length**	**num_words**	**avg_word_length**	**sentence_count**	**num_words**	**avg_word_length**	**sentence_count**
Count	100.00	100.00	48.00	48.00	48.00	100.00	100.00	100.00
Mean	13.43	6.69	12.85	6.36	1.27	9.96	5.78	1.26
Std.	4.72	1.09	5.34	1.59	0.45	6.28	1.32	0.48
Min.	6.00	4.62	5.00	3.80	1.00	1.00	4.00	1.00
Q1	10.00	6.00	9.00	5.20	1.00	6.00	4.67	1.00
Q2	12.00	6.54	11.50	5.85	1.00	10.00	5.87	1.00
Q3	17.00	7.07	16.00	7.21	2.00	14.00	6.86	1.00
Max.	28.00	10.45	28.00	12.00	2.00	27.00	9.50	3.00

Table 6: Statistics of lexical measures for complex texts, CEFR B1 and A2 simplifications in the test set of SimpEvalDE.

Statistic	Complex		B1			A2		
	num_words	**avg_word_length**	**num_words**	**avg_word_length**	**sentence_count**	**num_words**	**avg_word_length**	**sentence_count**
Count	60.00	60.00	29.00	29.00	29.00	60.00	60.00	60.00
Mean	13.62	6.75	12.10	6.19	1.14	10.02	5.78	1.30
Std.	5.31	1.16	5.26	1.47	0.35	8.38	1.56	0.65
Min.	6.00	4.40	5.00	4.43	1.00	1.00	4.00	1.00
Q1	10.00	5.89	8.00	5.17	1.00	2.50	4.00	1.00
Q2	12.00	6.72	11.00	6.20	1.00	9.00	5.73	1.00
Q3	17.00	7.38	15.00	6.60	1.00	14.00	6.79	1.00
Max.	28.00	10.14	26.00	10.40	2.00	40.00	10.33	4.00

Table 7: Statistics of lexical measures for complex texts, CEFR B1 and A2 simplifications in the original APA-LHA dataset and DEplain-apa test set.

(a) APA-LHA Dataset.

Statistic	num_words_compl	avg_word_length_compl
Count	1828	1828
Mean	17.36	6.39
Std.	8.01	1.16
Min.	3	3.50
Q1	11	5.67
Q2	17	6.29
Q3	23	6.92
Max.	42	17.20

(b) DEplain-apa Test Set.

Statistic	B1_words	B1_length	A2_words	A2_length
Count	1231	1231	1231	1231
Mean	10.57	6.40	10.80	6.16
Std.	4.68	1.56	5.04	1.56
Min	2.00	3.71	1.00	3.47
Q1	7.00	5.38	7.00	5.22
Q2	10.00	6.11	10.00	5.86
Q3	13.00	7.00	13.00	6.80
Max.	45.00	16.67	48.00	26.00

Table 8: Count of simplification strategies across CEFR levels in the train and test sets of SimpEvalDE.

Subset	Strategy	Complex-B1-A2	Complex-B1	B1-A2	Complex-A2	Total
Train	*delete*	5	5	3	5	18
	paraphrase	15	12	24	4	55
	split	6	5	14	2	27
Test	*delete*	4	3	2	3	12
	paraphrase	7	9	14	2	32
	split	3	3	9	1	16

Table 9: Model configuration parameters and validation performance for DEtect.

Variant	Dropout	LR	Pre-trained Model	Hidden Sizes	Val. Pearson	Val. Spearman
DEtect_multi	0.10	3e-5	xml-roberta-base	[2304, 768]	0.86	0.64
DEtect_multi_reg	0.20	1e-5	xml-roberta-base	[2304, 768]	0.86	0.66
DEtect_multi_reg_wechsel	0.20	1e-5	roberta-base-wechsel-german	[2304, 768]	0.87	0.67
DEtect_multi_wechsel_reduced	0.10	1e-5	roberta-base-wechsel-german	[128, 64]	0.86	0.68

B.2 Task-Specific Fine-Tuned Models

mBART-DEplain Stodden et al. (2023)

- Model: Based on mBART (Chipman et al., 2022) for multilingual text-to-text generation. Fine-tuned on DEplain-apa dataset targeting CEFR

A2 level of text simplification. Previously used in German ATS research by Gonzales et al. (2021), Carrer et al. (2024), and Stodden (2024b).

- Configuration: Default settings.
- HF Endpoint: 🤗 `DEplain/trimmed_mbart_sents_apa`

mT5-DEplain Stodden et al. (2023)

- Model: Based on mT5 (Xue et al., 2021) architecture for multilingual text-to-text generation. Fine-tuned on DEPLAIN-APA dataset targeting CEFR A2 level of text simplification. Previously used by Stodden (2024b) for reproduction studies.
- Configuration: Default settings.
- HF Endpoint: 🤗 DEplain/mt5-DEplain-APA

B.3 ATS Generation Prompt

We used the following prompt to generate text simplifications. We omitted here the English translation of this prompt for less verbosity, however, we have written all LLM-as-a-Judge prompts in English to facilitate the Human-in-the-Loop process.

Bitte formulieren Sie den folgenden komplexen Satz um. Sie können dies tun, indem Sie komplexe Wörter durch einfachere Synonyme ersetzen (d. h. paraphrasieren), unwichtige Informationen löschen (d.h. zusammenfassen) und/oder einen langen komplexen Satz in mehrere einfachere Sätze aufteilen. Der/Die endgültige(n) vereinfachte(n) Satz(e) muss/müssen grammatikalisch korrekt und flüssig sein und die Hauptgedanken des ursprünglichen Satzes beibehalten, ohne dessen Bedeutung zu verändern.

Ein einfacher Satz im Deutschen sollte:

- Einfache Wörter und gebräuchliche Wörter verwenden (seltene oder lange Substantive durch ein Synonym ersetzen oder zusammengesetzte Wörter mit einem Bindestrich trennen, z. B. Atomabkommen -> Atom-Abkommen). Schwierige Begriffe vermeiden und gegebenenfalls erläutern.
- Einfache Satzstrukturen mit möglichst wenigen Nebensätzen, koordinierten Sätzen und Genitivkonstruktionen verwenden.
- Informationen in Klammern (z. B. Angela Merkel (die ehemalige Bundeskanzlerin)) oder in verschachtelten Kommas (z. B. Angela Merkel, die ehemalige Bundeskanzlerin, . . .) vermeiden.
- Aktiv statt Passiv verwenden.
- Plusquamperfekt statt Plusquamperfekt verwenden
- Metaphern, Initialen/Abkürzungen, Sonderzeichen, Prozentsätze oder große Zahlen vermeiden
- Zahlen (vier) durch Ziffern (4) ersetzen
- Fremdwörter vermeiden, außer sie sind sehr bekannt
- Positive statt negative Sätze verwenden und gegebenenfalls direkte Rede mit Wörtern wie “du” oder “sie” verwenden.

`{five_shot}`

Nur die vereinfachte Version ausgeben, ohne Einleitung oder Erklärung.

Eingabe: `{text}`

Ausgabe:

C LLM Grading Rubric Development

C.1 Instruction to Generate *Prompt-Zero*

We used the following prompt to generate the initial Prompt-Zero for LLM-as-a-Judge evaluation.

I need you to rewrite the following grading rubric so that it can be used as a prompt for a large language model to evaluate the quality of a simplified sentence(s) compared to its original version and fixes the issues outlined below.

A complex sentence can be simplified by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. condensing) and/or splitting a long complex sentence into several simpler sentences. The final simplified sentence(s) must be grammatically correct, flow smoothly and retain the main ideas of its original counterpart without changing its meaning.

Current Rubric:

Rate the simplified sentence(s) based on how well it simplifies the original sentence using the following rubric:

- 100 - Only when the sentence is fully simplified, entirely fluent, and preserves the core meaning of the original sentence.
- 75 - The sentence is somewhat simpler, mostly fluent, and the meaning is close to the original sentence.
- 50 - The sentence is simpler, somewhat fluent, and the meaning is similar to the original sentence.
- 25 - The sentence is equivalently simple, still has some fluency, but the meaning is lost.
- 0 - The sentence is completely unreadable.

Issues with the Current Rubric:

1. Lack of clear definitions - The rubric does not explicitly define what makes a sentence simple, fluent, or meaningfully preserved.
2. Vague language: Words like “somewhat” and “similar” lack precise meaning and could lead to

differing interpretations by different evaluators. A good example of a grading criteria level could be: "75 - Mostly simple but may retain minor complexity, such as one or two long words or slightly complex phrasing."

3. No separation of criteria - The rubric does not evaluate simplicity, meaning preservation, and fluency independently.

4. Limited scoring flexibility - The model should be able to assign any score between 0 and 100, not just fixed levels.

5. Simplicity should not be about how well it is simplified compared to the original but how simple it is in general (i.e. if the original is already simple, it does not need to be simplified).

6. Meaning preservation does not account for situations where additional information is introduced not present in the original.

7. It is not clearly stated that the simplification can consist of two sentences if the splitting of a complex sentence simplifies reading.

Definitions of Key Criteria:

1. Simplicity: Measures how easy to understand the simplified sentence is, irrespective of the complexity of the original. A simple sentence in German should:

   - Use simple and common words (replace rare or long nouns with a synonym or split compound words using a hyphen, e.g. Atomabkommen -> Atom-Abkommen). Avoid difficult terms and explain if necessary.
   - Use simple sentence structures with minimal subordinate clauses, coordinated clauses, and genitive constructions.
   - Avoid including information in parentheses (e.g. Angela Merkel (die ehemalige Bundeskanzlerin)) or in nested commas (for example Angela Merkel, die ehemalige Bundeskanzlerin, ...).
   - Use active as opposed to passive voice.
   - Use past perfect instead of past simple.
   - Avoid metaphors, initials/abbreviations, special characters, percentages, or big numbers.
   - Replace numbers written as words (vier) with digits (4).
   - Avoid foreign terms unless they are very well known.
   - Use positive rather than negative sentences and direct speech with words like "du" or "Sie" where appropriate.

2. Meaning Preservation: Measures whether the simplified output adequately expresses the meaning of the original sentence, perhaps omitting the least important information.

3. Fluency: Measures grammatical correctness and fluency of the output.

How to Rewrite the Rubric:

1. Each criterion (Simplicity, Meaning Preservation, Fluency) should be scored separately from 0 to 100.

2. Before providing the grading levels for each criterion, the criterion should be defined in detail, especially simplicity, and then grading levels provided.

3. The grading rubric should have clearly defined levels (100, 75, 50, 25, 0), with concise and objective descriptions at each level that specify what makes a sentence highly simplified, moderately simplified, barely simplified, etc. Ensure that two different evaluators would not interpret the criteria differently.

4. The model should be instructed to assign a continuous score between 0-100.

The output format should be structured as follows:

Feedback: Explain strengths and weaknesses for each criterion, referencing original and simplified.

Score: Simplicity: {X}, Meaning Preservation: {Y}, Fluency: {Z}.

Now, generate the revised grading instruction prompt for a large language model following these guidelines and addressing the issues.

C.2 *Prompt-Zero*

We present the initial *Prompt-Zero* as follows:

Task Overview:

Evaluate the quality of a simplified sentence or set of sentences by comparing it to its original version. The evaluation should focus on three main criteria: Simplicity, Meaning Preservation, and Fluency. Assign a separate score for each criterion on a scale from 0 to 100, and provide feedback highlighting strengths and weaknesses. The simplified version can consist of one or more sentences if splitting enhances readability.

Criteria Definitions and Grading Levels:

1. Simplicity
   - Definition: Simplicity measures the ease of understanding the simplified sentence, irrespective of the original sentence's complexity. Key factors influencing simplicity include vocabulary, sentence structure, and clarity.
   - Guidelines for Simplicity:
     - Use straightforward and common words, replacing rare or compound words with simpler synonyms.
     - Maintain simple sentence structures with few subordinate clauses and minimal use of coordination and genitive forms.
     - Present information without parentheses or nested commas.
     - Favor active voice and past perfect tense over passive voice and simple past tense.

  - Avoid metaphors, initials, special characters, percentages, big numbers, and complex terms.
  - Replace numbers with digits and use positive sentences with direct speech (using “du” or “Sie” appropriately).
- Grading Levels:
  - 100: Exceptionally simple language, very easy to understand, adheres strictly to simplicity guidelines.
  - 75: Mostly simple with minor complexities, such as infrequent longer words or slightly complex structures.
  - 50: Moderate simplicity, noticeable complexity in vocabulary or structure, room for improvement.
  - 25: Complexity levels similar to the original, little effort towards simplification.
  - 0: Highly complex and difficult to understand, fails to simplify the original sentence.

2. Meaning Preservation

   - Definition: Meaning preservation evaluates whether the simplified version accurately conveys the core meaning of the original sentence, omitting only the least important details without introducing additional, unrelated information.
   - Grading Levels:
     - 100: Fully maintains the original meaning, with no significant omissions or extraneous additions.
     - 75: Mostly preserves the core message, with some minor omissions or slight additions.
     - 50: Meaning is generally preserved but with notable omissions or the introduction of some unrelated information.
     - 25: Basic meaning is lost, with significant omissions or many added details not present in the original.
     - 0: Fails to convey the original meaning, with major discrepancies or irrelevant information.

3. Fluency

   - Definition: Fluency assesses grammatical correctness and the natural flow of the simplified sentence. A fluent sentence reads smoothly, without awkward phrasing or grammatical errors.
   - Grading Levels:
     - 100: Completely grammatically correct, flows naturally and reads smoothly.
     - 75: Mostly fluent with minor grammatical errors or slight awkwardness.
     - 50: Moderately fluent, with noticeable errors impacting reading flow.
     - 25: Fluency is compromised, with frequent grammatical mistakes or clumsy phrasing.
     - 0: Completely lacks fluency, with critical grammatical issues making the text unreadable.

Evaluation Output Format:

- Feedback: Describe the strengths and weaknesses of the simplified sentence in comparison to the original. Use specific examples from the text to support the assessment within each criterion.

- Score:

  - Simplicity: {X}
  - Meaning Preservation: {Y}
  - Fluency: {Z}

Example Output:

- Feedback:

  - Simplicity: The `{simplified_sentence}` sentence uses approachable vocabulary, but includes one complex compound word.
  - Meaning Preservation: The main idea is preserved in `{simplified_sentence}`, but some minor details were omitted.
  - Fluency: `{simplified_sentence}` reads smoothly except for one awkward phrase in the middle.

- Score:

  - Simplicity: 85
  - Meaning Preservation: 70
  - Fluency: 80

This comprehensive rubric and evaluation format ensures clear and objective assessment of sentence simplification, allowing for varied interpretations without losing consistency across different evaluators.

C.3 Iterative Prompt Refinement

Example 1

Complex: Wenn es gelinge, die Todeszahl durch Eindämmungsmaßnahmen auf 100.000 zu begrenzen, dann haben wir alle zusammen einen guten Job gemacht, sagte Trump am Sonntagabend im Weißen Haus. *(If we manage to limit the number of deaths to 100,000 through containment measures, then we have all done a good job together, Trump said on Sunday evening in the White House.)*

Simplified: Trump sagte auch: Wenn nicht mehr als 100.000 Menschen sterben, dann haben wir einen guten Job gemacht. *(Trump also said: If no more than 100,000 people die, then we have done a good job.)*

***LLM-Judge* (GPT-4o):**

Evaluation of Sentence Simplification

- Feedback:

  - Simplicity: The `{simplified_sentence}` sentence achieves a relatively high level of simplicity by removing some of the more complex phrasing found in the

`{complex_sentence}`. For instance, it uses straightforward vocabulary, such as replacing "Eindämmungsmaßnahmen" and "im Weißen Haus" with simpler phrasing. However, it could further simplify by using more everyday language for "sterben" such as "umkommen" if the context allows.
  - Meaning Preservation: The core meaning of the `{complex_sentence}` is largely preserved in the `{simplified_sentence}`. The simplification retains the primary message of Trump's statement regarding the death toll limit as an indicator of success. However, some contextual details, like the venue "im Weißen Haus" and the specific timing "am Sontagabend", are omitted. While these details may not be crucial, their removal somewhat changes the context.
  - Fluency: The `{simplified_sentence}` sentence is mostly fluent, with a clear grammatical structure and a smooth flow. There are no significant grammatical errors, and the sentence reads naturally. The use of direct speech is correctly integrated, maintaining coherence and readability.

- Score:
  - Simplicity: 90
  - Meaning Preservation: 80
  - Fluency: 95

The simplified sentence excels in simplicity and fluency, making it accessible and easy to read. While it mostly preserves the essential meaning, it could better maintain the context by considering the inclusion of key details present in the original. Overall, it is an effective simplification.

***Human-Judge*:**

Simplicity: The sentence is simpler indeed, although the sentence is somewhat long. I would grade 80.

Meaning Preservation: I agree with your feedback that there is some loss of information. Notably the omission of "containment measures" erases the information that people are dying due to COVID-19 and not another reason, which does somewhat alter the understanding. I would grade 60.

Fluency: I agree with your feedback but "sagte auch" is a bit confusing. I would give about 90.

Example 2

Complex: Liverpool steht wie im Vorjahr im Finale der Fußball-Champions-League. *(Liverpool is in the final of the Champions League, just like last year.)*

Simplified: Durch den überraschend hohen Sieg steht Liverpool jetzt aber im Finale der Champions League. *(Thanks to the surprisingly high victory, Liverpool is now in the final of the Champions League.)*

***LLM-Judge* (GPT-4o):**

Evaluation of Sentence Simplification

- Feedback:
  - Simplicity: The `{simplified_sentence}` sentence introduces slightly more complexity compared to the `{complex_sentence}`. The `{complex_sentence}` is already quite simple, with succinct phrasing. The inclusion of the phrase "durch den überraschend hohen Sieg" in the `{simplified_sentence}` introduces additional information, potentially increasing complexity without simplifying the original context. There is a shift to a more elaborate construction instead of the straightforward statement in the `{complex_sentence}`.
  - Meaning Preservation: The `{simplified_sentence}` sentence retains the general idea that Liverpool is in the final of the Champions League. However, it introduces new information about an "überraschend hohen Sieg", which was not present in the `{complex_sentence}`. This addition could mislead readers about the context or suggest a narrative not provided in the `{complex_sentence}`. The simplicity of the original indication of repetition from "wie im Vorjahr" is lost, altering the sequence or context regarding the team's progression to the final.
  - Fluency: The `{simplified_sentence}` sentence maintains fluency with grammatically correct structure and clear communication. It reads smoothly, but the introduction of the phrase "jetzt aber" could slightly disrupt the directness found in the `{complex_sentence}`. The sentence still flows naturally, although it becomes slightly less straightforward due to the additional context.

- Score:
  - Simplicity: 60
  - Meaning Preservation: 50
  - Fluency: 85

The `{simplified_sentence}` sentence struggles to improve simplicity due to the introduction of extra information, which may divert from the original's straightforwardness. While it maintains overall fluency, the added context can obscure the precise meaning and relevance intended in the `{complex_sentence}` sentence. Further simplification should aim at preserving the essence without unnecessary embellishments, focusing more directly on the primary message from the `{complex_sentence}` without additional context unless essential.

***Human-Judge*:**

Simplicity: While the sentence is lexically simpler than the original, it is a syntactically more complicated structure. Adding information not present in the original also does it make it more complicated. TO INCORPORATE IN THE FINAL RUBRIC: Even if the final sentence is not extremely complex, due to it being more complex than the original, it should be below 25.

Meaning Preservation: Hallucinating information not present in the original should be highly penalized. While it does not change the sentence meaningfully as the main point is Liverpool advancing into the finals, the grading needs to score sentences that could mislead the reader low. TO INCORPORATE IN THE FINAL RUBRIC: Therefore, introduction of any information not present in the original should make the score low - below 50.

Fluency: While it is fluent, “aber” is not necessary and makes it slightly confusing. I would score also around 85-90.

Example 3

Complex: Die finanzielle Lage der Haushalte hat sich in der Coronakrise deutlich verschärft, geht aus einer Umfrage (Sample 1.200 Personen) der Tarifvergleichsplattform durchblicker hervor. *(The financial situation of households has changed significantly during the Corona crisis, according to a survey (sample 1,200 people) by the tariff comparison platform durchblicker.)*

Simplified: Rund 38 Prozent der Haushalte verdienen weniger als vor der Corona-Krise. Rund 67 Prozent der Haushalte wollen die Fixkosten senken. *(Around 38 percent of households earn less than before the Corona crisis. Around 67 percent of households want to reduce fixed costs.)*

LLM-Judge **(GPT-4o):**

Evaluation of Sentence Simplification

- Feedback:
  - Simplicity: The `{simplified_sentence}` sentences break down the information into two distinct, shorter statements, making it easier for the reader to process. This approach aids clarity and aligns well with simplicity guidelines. However, the `{complex_sentence}` sentence’s use of parentheses and specific phrasing is removed, potentially complicating comprehension by introducing new statistics that were not in the original, which adds unnecessary complexity instead of focusing on simplification.
  - Meaning Preservation: The simplification significantly diverges from the `{complex_sentence}` by introducing new statistical information that was not present initially. The `{complex_sentence}` mentions a general deterioration in the financial situation, but the `{simplified_sentence}` sentences provide specific figures (“38 Prozent” for decreased income and “67 Prozent” intending to reduce fixed costs) that were not stated in the `{complex_sentence}`, altering the intended message and context. Thus, it fails to adequately preserve the original meaning.
  - Fluency: The `{simplified_sentence}` sentences maintain grammatical accuracy and read smoothly, with clear connections between ideas. There is no awkward phrasing or noticeable grammatical errors, ensuring that the sentences flow naturally. The clear distinction of two separate ideas contributes to enhanced readability.
- Score:
  - Simplicity: 40
  - Meaning Preservation: 20
  - Fluency: 90

While the `{simplified_sentence}` sentences attempt to present the information in a clearer manner, they introduce details not found in the `{complex_sentence}`, complicating the overall message and context. It retains high fluency due to its grammatical correctness and logical flow but struggles in simplicity due to added complexity. For an effective simplification, the focus should remain on conveying the essential meaning without adding or omitting critical information from the `{complex_sentence}` sentence.

Notes on Incorporation into Final Rubric:

1. Simplicity: Ensure added information does not increase complexity compared to the original, and score below 25 if it does.

2. Meaning Preservation: Penalize heavily (below 50) for any introduction of information not present in the original, even if it maintains core thematic elements.

3. Fluency: Maintain high scores for sentences that are grammatically correct and smoothly flowing, even when meaning preservation is not fully achieved.

***Human-Judge*:**

Simplicity: Here, the output was made significantly simpler than the original, which was a long sentence with two clauses with compound words and information inside parentheses. TO INCORPORATE IN THE FINAL RUBRIC: whether the original content is well reflected in the simplified output should only be assessed in Meaning Preservation category. The simplification splits the clauses into separate sentences in active voice, which is encouraged. While additional information was introduced, it is definitely easier to read. I agree that the concrete figures make it somewhat more complex, but it is still significantly easier than the original. I would give this simplification about 80-90.

Meaning Preservation: While the general meaning of the household's situation worsening was retained in the first sentences, additional information not present in the original (38%). Furthermore, the second sentence is completely unrelated. However, the information coming from a survey is a detail removing which does not change the meaning of the sentence. I would give this around 25.

Fluency: Splitting a complicated sentence into two is encouraged in simplification, as long as the sentences are logically connected. Here, however, the splitting of the sentences does indeed produce a non-smooth transition between ideas. I would give this about 70.

C.4 *Prompt-Final*

We present the ultimate *Prompt-Final* as follows:

Task Overview:

Evaluate the quality of a simplified German sentence or set of sentences by comparing it to its original version. The evaluation should focus on three main criteria: Simplicity, Meaning Preservation, and Fluency. Assign a separate score for each criterion on a scale from 0 to 100. The simplified version can consist of one or more sentences if splitting enhances readability.

Ways to Simplify:

A complex sentence can be simplified through:

- Paraphrasing: Replacing complex words with simpler synonyms.
- Condensing: Deleting unimportant information.
- Splitting: Dividing a long, complex sentence into several simpler sentences.

Evaluation Criteria and Definitions:

1. Simplicity
   - Definition: Simplicity measures how easy the simplified sentence is to understand, irrespective of the original sentence's complexity. This includes vocabulary, sentence structure, and clarity. Simplicity should not be assessed by comparing directly to the original but by how simple the sentence is in general. If the original is already very simple, an unchanged sentence might still score high.
   - Guidelines for Simplicity:
     - Use simple words and common words (replace rare or long nouns with a synonym or split compound words using a hyphen, e.g. Atomabkommen -> Atom-Abkommen). Avoid difficult terms and explain if necessary.
     - Use simple sentence structures with minimal subordinate clauses, coordinated clauses and genitive constructions.
     - Avoid including information in parentheses (e.g. Angela Merkel (die ehemalige Bundeskanzlerin)) or in nested commas (e.g. Angela Merkel, die ehemalige Bundeskanzlerin,...).
     - Use active as opposed to passive voice.
     - Use past perfect instead of past simple.
     - Avoid metaphors, initials/abbreviations, special characters, percentages or big numbers.
     - Replace numbers (vier) with digits (4).
     - Avoid foreign terms unless they are very well-known.
     - Use positive rather than negative sentences and direct speech with words like "du" or "Sie" where appropriate.
   - Grading Levels:
     - 100: Exceptionally simple, very easy to grasp; adheres strictly to simplicity guidelines.
     - 75: Mostly simple with minor complexities, such as infrequent longer words.
     - 50: Moderate simplicity; noticeable complexity in vocabulary or structure.
     - 25: Similar complexity level to the original; minimal simplification.
     - 0: Overly complex, difficult to understand.
2. Meaning Preservation
   - Definition: Meaning preservation evaluates whether the simplified version genuinely conveys the core meaning of the original sentence without introducing unrelated information along with omitting less important details.
   - Grading Levels:
     - 100: Fully retains the original meaning without omissions or unrelated additions.
     - 75: Mostly preserves the message, with minor omissions.
     - 50: Generally preserves meaning but with notable omissions or some unrelated information.
     - 25: Loses basic meaning with significant omissions or many irrelevant details.
     - 0: Fails to convey the original meaning accurately.
3. Fluency
   - Definition: Fluency assesses the grammatical correctness and natural flow of the simplified sentence.
   - Grading Levels:
     - 100: Fully grammatically correct and naturally flowing.
     - 75: Mostly fluent with minor grammatical errors.
     - 50: Moderately fluent with noticeable errors affecting flow.
     - 25: Many grammatical mistakes, affecting readability.

  - 0: Lacks fluency significantly; unreadable.

Evaluation Output Format:

- Feedback: Describe the strengths and weaknesses of the simplified sentence in comparison to the original. Include specific examples to support the assessment. Score cannot be less than 0 or more than 100.
- Score:
  - Simplicity: {X}
  - Meaning Preservation: {Y}
  - Fluency: {Z}

Example 1:

Complex: Coca-Cola bezeichnete die Auszeichnung hingegen als nicht gerechtfertigt und kündigte an, den Goldenen Windbeutel nicht anzunehmen.

Simplification: Coca-Cola will den Goldenen Windbeutel nicht annehmen.

Feedback:

- Simplicity: The simplification effectively reduces the complexity of the original sentence by eliminating intricate details and focusing on the core message. It replaces the more complex structure of the original with a straightforward statement. The removal of terms like "bezeichnet" and "als nicht gerechtfertigt" reduces complexity, aligning well with simplicity guidelines.
- Meaning Preservation: While the simplified sentence maintains the core meaning that Coca-Cola does not intend to accept the award (Goldenen Windbeutel), it omits some contextual details from the original. Specifically, the simplification does not convey that Coca-Cola also considered the award to be unjustified. This omission reduces the score for meaning preservation.
- Fluency: The simplified sentence is grammatically correct and flows naturally. It adheres to standard German syntax and structure, without any grammatical errors.

Score:

- Simplicity: 90
- Meaning Preservation: 50
- Fluency: 100

Example 2:

Complex: Der Lenker soll schwer verletzt und mit einem Notarzthubschrauber in ein Krankenhaus geflogen worden sein.

Simplification: Der Lkw-Fahrer wurde beim Unfall schwer verletzt und in ein Krankenhaus geflogen.

Feedback:

- Simplicity: The simplified sentence replaces "Lenker" with "Lkw-Fahrer", making the subject more specific, which can either increase or decrease simplicity depending on the assumed knowledge of the reader. The sentence structure is straightforward, eliminates the passive voice and the use of a helicopter, opting for a more direct description. It scores well in simplicity by avoiding complex constructions and by using common vocabulary.
- Meaning Preservation: The complex sentence mentions the use of an emergency doctor helicopter ("Notarzthubschrauber"), which is omitted in the simplified version. This is a minor yet specific detail that contributes to the original meaning. However, the critical information about an injury and hospital transport is preserved.
- Fluency: The simplified sentence is grammatically correct and maintains a natural flow, comparable to the original. It uses proper sentence structure, making it easy to read and understand.

Score:

- Simplicity: 85
- Meaning Preservation: 75
- Fluency: 95

Example 3:

Complex: Bei einem schweren Busunglück auf der portugiesischen Urlaubsinsel Madeira sind 29 Menschen ums Leben gekommen.

Simplification: Am Mittwoch ist es auf der Urlaubsinsel Madeira in Portugal zu eine schweren Bus-Unfall gekommen. Dabei sinde 29 Menschen gestorben.

Feedback: The simplified sentence effectively uses two distinct sentences to convey the information, enhancing readability. The use of common terms like "Unfall" (accident) instead of "Unglück" (misfortune) aligns well with the simplicity criterion. However, there are some grammatical issues, such as "sinde" instead of "sind", which affects fluency.

- Simplicity: The simplified version achieves simplicity by splitting the original sentence and using straightforward language. The sentence structure is simple enough for easy comprehension. The simplification slightly modifies the phrasing but retains simple language, fitting well within the guidelines for Simplicity.
- Meaning Preservation: The simplified sentence effectively preserves the essential meaning of the original, including the location and the occurrence of a bus accident resulting in fatalities. However, the simplification introduces the detail "Am Mittwoch", which is not present in the original, which could affect meaning preservation slightly.
- Fluency: The grammatical errors, such as "sinde" instead of "sind", detract from the fluency. Additionally, "zu eine" should be "zu einem" for grammatical correctness in German. These errors can disrupt the natural flow and readability of the text.

Score:

- Simplicity: 80
- Meaning Preservation: 70
- Fluency: 50

Example 4:

Complex: Ein Lkw mit einem Anhñger ist Montagnachmittag in Munderfing im Bezirk Braunau im Innviertel mit einem Zug zusammengestoßen und hat in der Folge ein Haus schwer beschädigt.

Simplification: In Oberösterreich fuhr ein Zug in einen Lkw. Am Montag kam es im Ort Munderfing im Bundesland Oberösterreich zu einem Unfall. Anschließend krachte der Lkw in ein Haus. Der Fahrer wurde mit einem Hubschrauber in ein Krankenhaus gebrach.

Feedback: The simplified version successfully breaks down the original sentence into several shorter sentences, enhancing readability. It effectively segregates the complex information into distinct parts, making it easier to follow along for the reader. However, the addition of new information, such as "Der Fahrer wurde mit einem Hubschrauber in ein Krankenhaus gebracht", introduces details not present in the original, impacting meaning preservation.

- Simplicity: The approach of splitting the original sentence into smaller sentences significantly aids understanding, and the use of straightforward language makes it simple. The sentences are straightforward and devoid of complex structures. The place "Oberösterreich" is used more than once, and the phrase "kam es zu einem Unfall" could be simplified further. Nevertheless, the overall simplification is good.
- Meaning Preservation: While the main events are preserved, there is a newly introduced detail about the driver, which was not mentioned in the original sentence, and it can mislead about the severity or outcome of the collision. This reduces the score for meaning preservation.
- Fluency: The sentences mostly follow grammar rules well and maintain a natural flow. There are minor issues, such as redundancy, which slightly affect overall fluency. Still, the content largely maintains a logical progression.

Score:

- Simplicity: 85
- Meaning Preservation: 50
- Fluency: 80

Example 5:

Complex: Damit sind seit Bekanntwerden des Erregers Sars-CoV-2 Ende vergangenen Jahres weltweit mehr als 28,6 Millionen Infektionen gemeldet worden.

Simplification: Des Corona-Virus seit Ausbruch gab es mehr weltweit als 28,6 Millionen Infektionen.

Feedback: The simplified sentence "Des Corona-Virus seit Ausbruch gab es mehr weltweit als 28,6 Millionen Infektionen" attempts to condense the original information. However, there are several issues in terms of simplicity, meaning preservation, and fluency.

- Simplicity: The simplified sentence is slightly simpler in using common words. It effectively reduces the complex expression "Erregers Sars-CoV-2" to "Corona-Virus". However, the phrase "Des Corona-Virus seit Ausbruch" is awkward and can confuse readers due to its incorrect grammatical construction. Additionally, the ordering and structuring make it harder to read than necessary.
- Meaning Preservation: The simplification loses critical context from the original sentence, such as the timeframe "Ende vergangenen Jahres" which sets a clear historical context for when the infections began reporting. Without this piece of information, the reader may not have a clear understanding of the full context.
- Fluency: The sentence lacks correct grammatical structure, especially evident in "Des Corona-Virus seit Ausbruch", which suffers from poor syntax. This grammatical flaw detracts from the overall fluency and makes the sentence less readable. The use of "mehr weltweit als" is also awkward and could be expressed more naturally to improve the flow.

Score:

- Simplicity: 50
- Meaning Preservation: 25
- Fluency: 25

Example 6:

Complex: Weltweit wird laut Studie immer mehr Alkohol getrunken.

Simplification: Laut einer Studie wird auf der Welt immer mehr Alkohol getrunken.

Feedback: The simplified sentence, "Laut einer Studie wird auf der Welt immer mehr Alkohol getrunken," demonstrates some simplification, although the original sentence was already relatively straightforward. It remains clear and easy to understand, employing a similar sentence structure but with slight changes, like switching "weltweit" to "auf der Welt," which is a more common phrase and may be slightly easier for some readers to parse.

- Simplicity: The simplification retains the straightforwardness of the original sentence but doesn't significantly increase simplicity since both sentences use common vocabulary and simple structures. There are minor changes that enhance readability, but the overall complexity remains similar.
- Meaning Preservation: The meaning is entirely preserved in the simplified version. Both sentences convey the same information regarding the increase in alcohol consumption worldwide. There are no omissions or additions that alter the original meaning.

- Fluency: The sentence is grammatically correct and flows naturally, similar to the original. There are no errors affecting readability or comprehension, and it reads smoothly in German.

Score:

- Simplicity: 85
- Meaning Preservation: 100
- Fluency: 100

Now grade the following simplification:

Complex: `{complex_sentence}`

Simplification: `{simplified_sentence}`

D *LLM-Judge* Selection Analysis

D.1 Questions and Hypothesis

We choose the Deepseek-V3 (Liu et al., 2024) herd of distilled models and Zephyr (Tunstall et al., 2024) as the backbone of *LLM-Judge*. We concentrate on studying the following two questions:

- **Model scalability:** Do smaller models, such as Distill-Qwen-1.5B or Distill-Qwen-7B, perform similarly to the larger model Distill-Qwen-14B? Answering this question ensures that the smaller models can be used to reliably generate quality scores for model training.
- **Inter-model performance:** To what extent do scoring behaviors differ between different model families of a similar size, e.g., Distill-Qwen, Distill-Llama and Zephyr (i.e., a Mistral variant)? This is done to validate that the three selected models exhibit similarly high grading quality.

To investigate the above two questions, we selected the following open-source LLMs for experiments:

- **Distill-Qwen-1.5B**: 🤗 `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`
- **Distill-Qwen-7B**: 🤗 `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`
- **Distill-Llama-8B**: 🤗 `deepseek-ai/DeepSeek-R1-Distill-Llama-8B`
- **Zephyr-7B**: 🤗 `HuggingFaceH4/zephyr-7b-beta`

Each model was used to generate *fluency*, *meaning preservation*, and *simplicity* scores using the *Prompt-Final* in ten sampling runs per instance for the full training set. The resulting scores were then analyzed in two ways: by examining the variance in scores across the ten runs for the same instance (sampling variability), and by evaluating the distribution of the averaged scores across the ten runs (score diversity).

Three hypotheses were tested as a measure of model quality:

- **Hypothesis 1 (Score diversity):** Higher-performing models will exhibit greater standard deviation in their *average scores* due to an increased ability to discriminate between low-quality and high-quality texts.
- **Hypothesis 2 (Sampling stability):** Higher-performing models will exhibit lower *per-row sampling variance*, i.e., the standard deviation of scores across the ten runs of the same data row, due to improved consistency and output stability.
- **Hypothesis 3 (Score correlation):** If two models are similarly capable, their scoring behavior should be positively correlated.

D.2 Experimental Findings

1. Scalability Across Model Sizes.

- ***Hypothesis 1: Increased Model Size Increases Scoring Variability.***

Table 10: Comparison of the standard deviation of average scores for Distill-Qwen-1.5B, Distill-Qwen-7B, and Distill-Qwen-14B, grouped by evaluation criterion and model pair. The p-values from Levene's test indicate the statistical significance of variance differences between model pairs.

Criterion	**Standard Deviation**	**p-value**
Distill-Qwen-1.5B vs Distill-Qwen-7B		
Simplicity	4.47 vs 9.22	0.0006
Meaning P.	10.18 vs 16.98	<0.0001
Fluency	6.60 vs 7.55	0.0220
Distill-Qwen-1.5B vs Distill-Qwen-14B		
Simplicity	4.47 vs 13.60	<0.0001
Meaning P.	10.18 vs 25.12	<0.0001
Fluency	6.60 vs 12.74	*0.1604*
Distill-Qwen-7B vs Distill-Qwen-14B		
Simplicity	9.22 vs 13.60	<0.0001
Meaning P.	16.98 vs 25.12	<0.0001
Fluency	7.55 vs 12.74	0.0075

As shown in Figure 5, the distributions of ten-run average scores for Distill-Qwen-1.5B, Qwen-7B, and Distill-Qwen-14B indicate that larger models produce greater variability in *simplicity* and *meaning preservation*, while trends for *fluency* are less consistent (see Table 10). For *simplicity*, the median remains stable across models, but the minimum scores drop sharply (50 to 14, $p < 0.0001$), driving standard deviations from 4.47 to 13.60 ($p < 0.001$). For *meaning preservation*, distributions shift from near-Gaussian to bimodal, with extreme values (0 and 100) increasingly frequent in larger models; standard deviations rise significantly (10.18 to 25.11, $p < 0.0001$). For *fluency*, although standard deviations also increase (6.6 to 12.7, no significance), inter-quartile ranges shrink and Distill-Qwen-14B predicts 100 in over half of cases, suggesting limited effective diversity. Overall, Hypothesis 1 is well supported for *simplicity* and *meaning preservation*, but evidence for *fluency* remains inconclusive.

- ***Hypothesis 2: Higher-Performing Models Exhibit Lower Sampling Variability.***

  Table 11 reports the average standard deviation of scores across ten sampled outputs per input for different Distill-Qwen sizes. Overall, variance decreases as model size increases, supporting Hypothesis 2. The main exception is *simplicity*, where the standard deviation rises significantly from 6.5 to 8.2 ($p = 0.0001$) between Distill-Qwen-7B and Distill-Qwen-14B. For all other criteria, standard deviations decrease consistently, with the largest reduction occurring between Distill-Qwen-1.5B and Distill-Qwen-7B.

Table 11: Comparison of the mean per-row output standard deviation for Distill-Qwen-1.5B, Distill-Qwen-7B, and Distill-Qwen-14B, grouped by evaluation criterion and model pair. p-values from independent *t*-tests indicate the statistical significance of differences in mean sampling variance between models.

Criterion	**Mean Std. Deviation**	**p-value**
Distill-Qwen-1.5B vs Distill-Qwen-7B		
Simplicity	9.006 vs 6.507	0.0000
Meaning P.	15.734 vs 10.047	0.0000
Fluency	11.456 vs 6.198	0.0000
Distill-Qwen-1.5B vs Distill-Qwen-14B		
Simplicity	9.006 vs 8.203	*0.0899*
Meaning P.	15.734 vs 7.392	0.0000
Fluency	11.456 vs 4.456	0.0000
Distill-Qwen-7B vs Distill-Qwen-14B		
Simplicity	6.507 vs 8.203	0.0001
Meaning P.	10.047 vs 7.392	0.0000
Fluency	6.198 vs 4.456	0.0000

- ***Hypothesis 3: Models of Increasing Size Exhibit Positive Correlation.***

  Table 12 shows pairwise correlations between Distill-Qwen models by criterion, with all values significant at $p < 0.0001$. The shading highlights cases where one model predicts below 25 while the other does not. Distill-Qwen-1.5B and 7B correlate moderately (Pearson $r = 0.39$ to 0.64), though *meaning preservation* exhibits many cases where 7B assigns low scores that 1.5B does not. Correlations with Distill-Qwen-14B drop sharply for 1.5B (Pearson $r = 0.23$ to 0.60), again reflecting more extreme low scores from the larger model. By contrast, 7B and 14B align more strongly (Pearson $r = 0.51$ to 0.87), though 14B still assigns unusually low values for *meaning preservation* and, to a lesser degree, *fluency*. Agreement in normalized z-scores confirms this pattern: Krippendorff's α reaches 0.70/0.87/0.51 for the 7B-14B pair, compared to 0.30/0.60/0.23 for 1.5B-14B. Overall, Qwen-7B and 14B show the highest consistency across all criteria.

Table 12: Pearson (r) and Spearman (ρ) correlations for (Distill-Qwen-)1.5B, (Distill-Qwen-)7B and (Distill-Qwen-)14B scores by criterion, all p-values are 0.

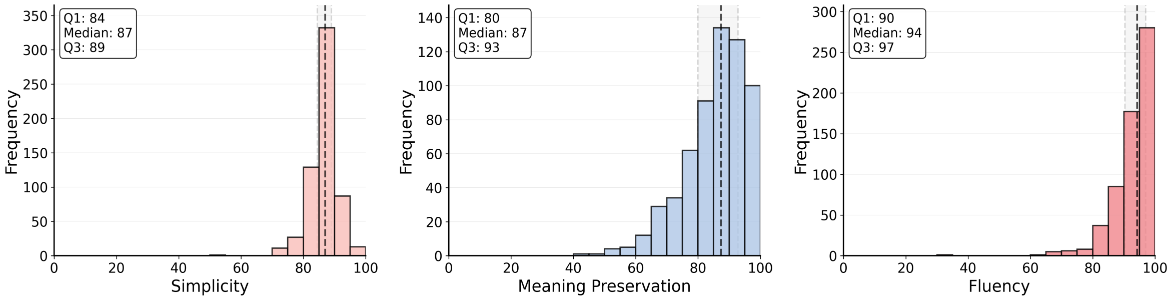
Model Comparison	**Criterion (r / ρ)**		
	Simplicity	***Meaning P.***	***Fluency***
1.5B v. 7B	0.39 / 0.32	0.64 / 0.67	0.42 / 0.38
1.5B v. 14B	0.30 / 0.21	0.60 / 0.60	0.23 / 0.27
7B v. 14B	0.70 / 0.45	0.87 / 0.79	0.51 / 0.45

2. Inter-model Comparison of Similar Sizes

- ***Hypothesis 1: Models Differ in the Variance and Distribution of Average Scores.***

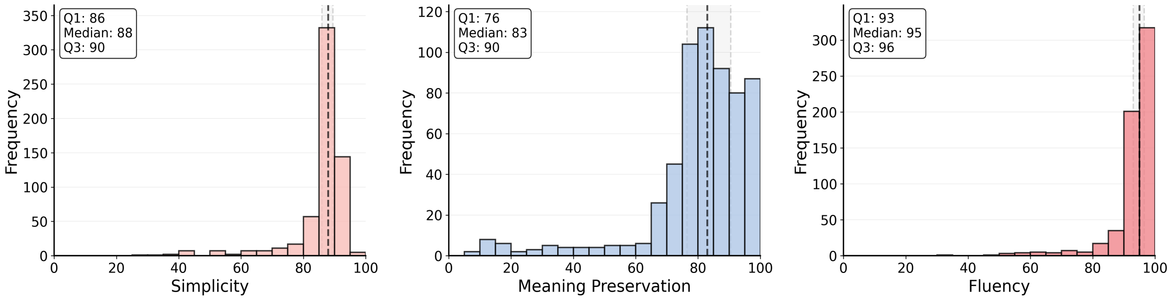
  Figure 6 compares the distributions of *simplicity*, *meaning preservation*, and *fluency* for Distill-Qwen-7B, Distill-Llama-8B, and Zephyr-7B. Distill-Llama-8B shows a stronger right skew in *simplicity* (median 90) and higher peaks in *meaning preservation* (median 91, with values extending down to 2), though its standard deviations are not significantly different from Distill-Qwen-7B. For *fluency*, the two models are visually similar, but Llama-8B has a slightly higher variance ($p < 0.1$). Zephyr-7B, by contrast, produces tighter distributions across all criteria, with higher medians (*simplicity*: 94) and lower standard deviations, reflecting reduced diversity.

i) Distill-Qwen-1.5B



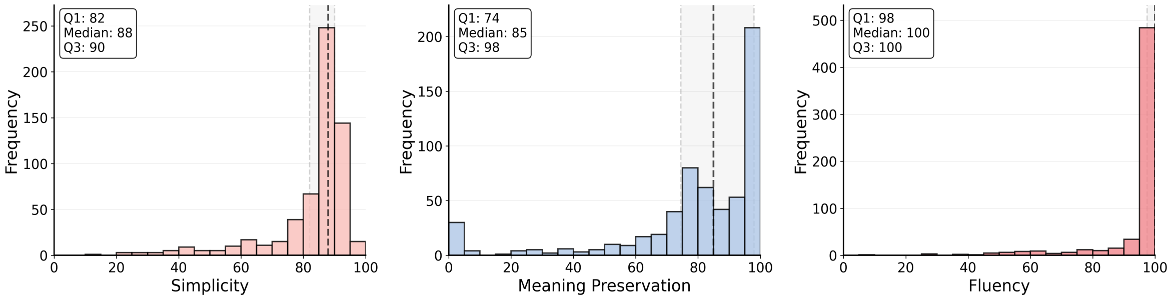
(a) Simplicity (b) Meaning Preservation (c) Fluency

ii) Distill-Qwen-7B



(d) Simplicity (e) Meaning Preservation (f) Fluency

iii) Distill-Qwen-14B



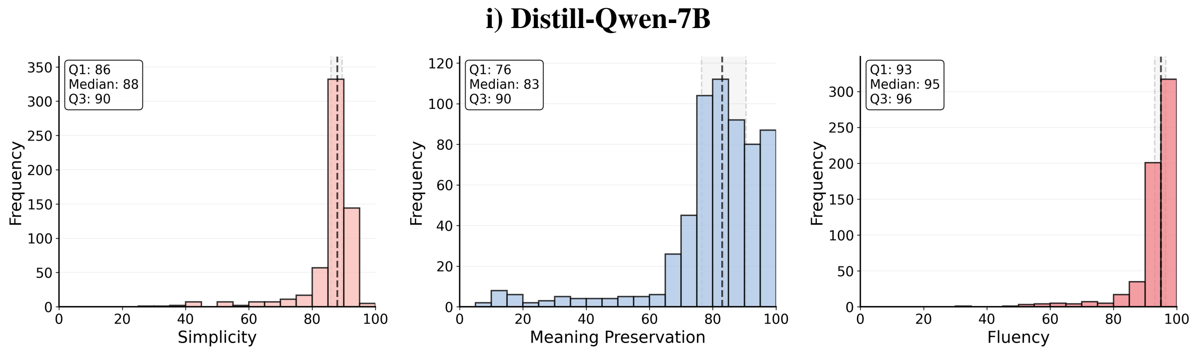
(g) Simplicity (h) Meaning Preservation (i) Fluency

Figure 5: Histograms of score distributions for i) Distill-Qwen-1.5B, ii) Distill-Qwen-7B, and iii) Distill-Qwen-14B, broken down by scoring criterion. The gray band indicates the IQR, while the black vertical dashed lines show the minimum, median, and maximum values. These values are also listed in the box in the upper left corner of each plot.

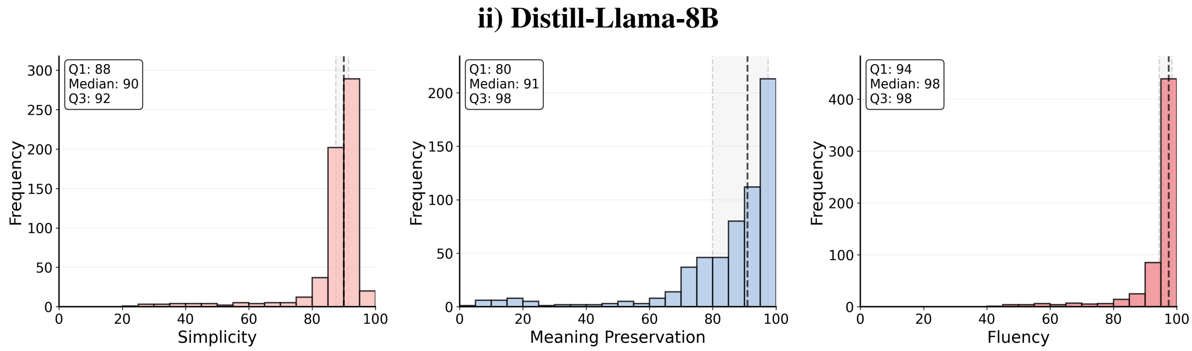
Overall, Zephyr-7B is less variable than the other two models, while differences between Distill-Qwen-7B and Llama-8B are minor and mostly not significant (see also Table 13).

- ***Hypothesis 2: Different Models have Different Sampling Variability.***

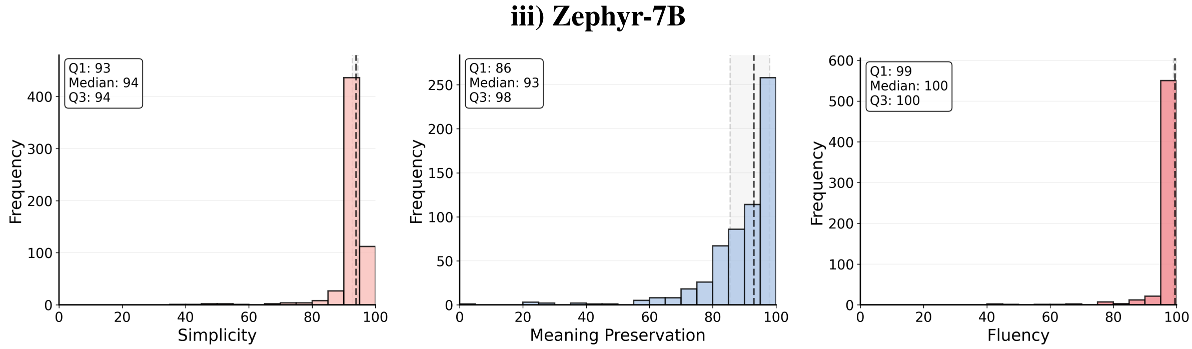
Table 14 compared the variance of ten repeated outputs per simplification pair. Distill-Qwen-7B and Distill-Llama-8B show nearly identical sampling variability across all criteria (e.g., *simplicity*: 6.51 vs. 6.49; *meaning preservation*: 10.05 vs. 10.03), with no significant differences ($p > 0.5$). Zephyr-7B, however, exhibits significantly lower variance in *simplicity* (4.01) and *fluency* (3.06) than either model ($p < 0.0001$), indicating more consistent predictions. The only exception is *meaning preservation*, where Zephyr-7B shows slightly higher variance (11.08 vs. 10.03), a marginal but significant difference ($p \approx 0.02$). Overall, Qwen-7B and Llama-8B display similar stability, while Zephyr-7B is notably steadier in *simplicity* and *fluency* but less so in *meaning preservation*.



(a) Simplicity (b) Meaning Preservation (c) Fluency



(d) Simplicity (e) Meaning Preservation (f) Fluency



(g) Simplicity (h) Meaning Preservation (i) Fluency

Figure 6: Histograms of score distributions for i) Distill-Qwen-7B, ii) Distill-Llama-8B, and iii) Zephyr-7B, broken down by scoring criterion. The yellow band indicates the IQR, while vertical ticks show the minimum, median, and maximum values. These values are also listed in the box in the upper left corner of each plot.

- ***Hypothesis 3: Models of Similar Size Exhibit Positively Correlated Scoring Behavior.***

  Table 15 shows significant positive correlations ($p < 0.0001$) across all model pairs. Distill-Qwen-7B and Distill-Llama-8B align most closely (Pearson r from 0.85 to 0.89), while correlations with Zephyr-7B are more moderate (Pearson r from 0.45 to 0.58). Disagreements mainly occur in cases where one model assigns scores below 25 and the other does not. Krippendorff's α on z-scores remains moderate to high (0.53 to 0.69), indicating consistent though weaker alignment with Zephyr-7B, comparable to correlations with larger Qwen models.

Table 15: Pearson (r) and Spearman (ρ) correlations with p-values for (Distill-)Qwen(-7B), (Distill-)Llama(-8B) and Zephyr(-7B) scores by criterion. All p-values are 0.

Model Comparison	**Criterion** (r / ρ)		
	Simplicity	***Meaning P.***	***Fluency***
Qwen vs Llama	0.85 / 0.51	0.89 / 0.77	0.76 / 0.45
Qwen vs Zephyr	0.49 / 0.36	0.58 / 0.66	0.42 / 0.37
Llama vs Zephyr	0.45 / 0.33	0.58 / 0.67	0.41 / 0.34

Table 13: Comparison of the standard deviation of average scores for Distill-Qwen-7B, Distill-Llama-8B, and Zephyr-7B, grouped by evaluation criterion and model pair. The p-values from Levene's test indicate the statistical significance of variance differences between model pairs.

Criterion	**Standard Deviation**	**p-value**
Distill-Qwen-7B vs Distill-Llama-8B		
Simplicity	9.22 vs 10.95	*0.371*
Meaning P.	16.98 vs 19.27	*0.257*
Fluency	7.55 vs 9.18	*0.097*
Distill-Qwen-7B vs Zephyr-7B		
Simplicity	9.22 vs 6.04	<0.0001
Meaning P.	16.98 vs 12.21	<0.0001
Fluency	7.55 vs 5.95	<0.0001
Distill-Llama-8B vs Zephyr-7B		
Simplicity	10.95 vs 6.04	<0.0001
Meaning P.	19.27 vs 12.21	<0.0001
Fluency	9.18 vs 5.95	<0.0001

Table 14: Comparison of the mean per-row output standard deviation for Distill-Qwen-7B, Distill-Llama-8B, and Zephyr-7B grouped by evaluation criterion and model pair. p-values from independent t-tests indicate the statistical significance of differences in mean sampling variance between models.

Criterion	**Mean Std. Deviation**	**p-value**
Distill-Qwen-7B vs Distill-Llama-8B		
Simplicity	6.51 vs 6.49	*0.9610*
Meaning P.	10.05 vs 10.03	*0.9436*
Fluency	6.20 vs 6.41	*0.5573*
Distill-Qwen-7B vs Zephyr-7B		
Simplicity	6.51 vs 4.01	<0.0001
Meaning P.	10.05 vs 11.08	0.0152
Fluency	6.20 vs 3.06	<0.0001
Distill-Llama-8B vs Zephyr-7B		
Simplicity	6.49 vs 4.01	<0.0001
Meaning P.	10.03 vs 11.08	0.0193
Fluency	6.41 vs 3.06	<0.0001

To summarize the experimental findings:

- **Correlation with larger judges**: smaller Qwen models show moderate to strong correlations with the 14B reference model, with the highest agreement consistently observed for meaning preservation and the lowest for fluency.

- **Criterion-dependent reliability**: Across all evaluated models, score reliability varies substantially by criterion; meaning preservation is the most stable dimension, while fluency exhibits compressed score ranges and lower discriminative power.

- **Model-specific behavior**: Distil-Qwen-7B achieves the strongest overall correlation with the 14B model, whereas Zephyr-7B produces less diverse but more conservative scores, indicating complementary strengths rather than a single optimal judge.

- **Implication for aggregation**: These findings motivate the use of multi-model aggregation for *LLM-Judge* scores and confirm the necessity of independent human evaluation for validating downstream metrics such as DETECT.

E DETECT vs Other Metrics by Type

Table 16 and Figure 9 show the per score type correlations between DETECT and other metrics. Each category was re-sampled to obtain 100 samples of size 45. Figure 10 shows the annotation tool we customized for this work.

Table 16: Pearson correlations of DETECT and other metrics with *Human-Judge* scores.

(a) Simplicity score correlations.

Metric	***Simplicity***		
	Paraphrase	*Delete*	*Split*
BLEU	0.225	0.194	0.415
SARI	0.072	0.223	0.357
BERTScore	0.370	0.395	0.554
LLM-Judge	0.265	0.322	0.395
DETECT	0.311	0.307	0.490

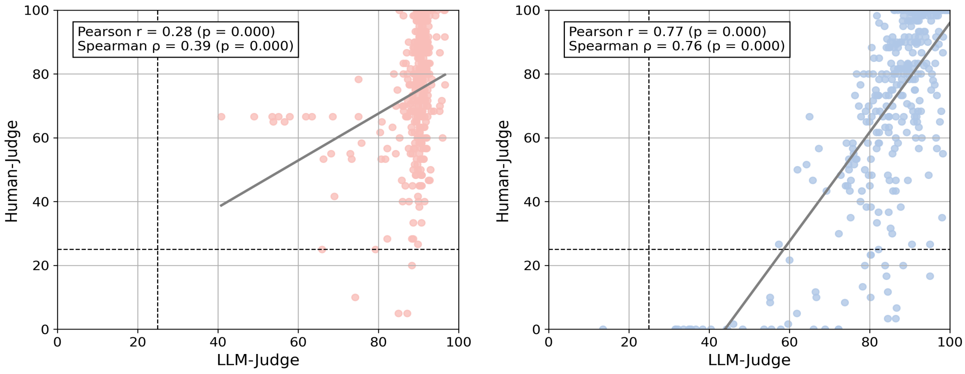
(b) Meaning preservation score correlations.

Metric	***Meaning Preservation***		
	Paraphrase	*Delete*	*Split*
BLEU	0.240	0.337	0.471
SARI	-0.017	0.256	0.276
BERTScore	0.310	0.469	0.680
LLM-Judge	0.636	0.791	0.795
DETECT	0.464	0.654	0.744

(c) Fluency score correlations.

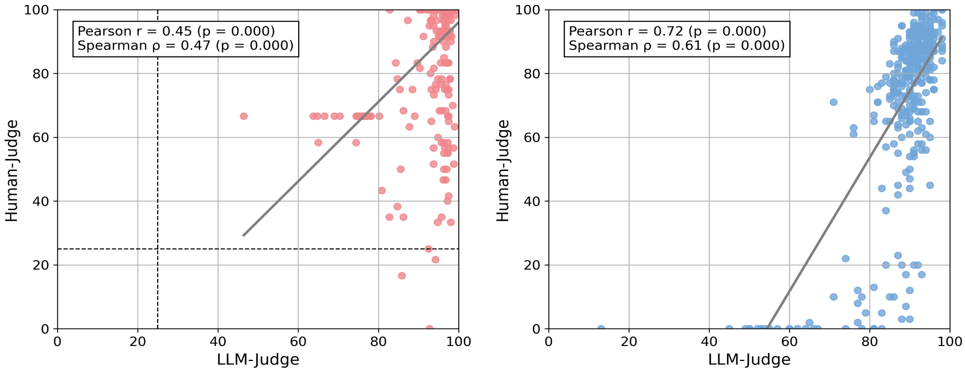
Metric	***Fluency***		
	Paraphrase	*Delete*	*Split*
BLEU	0.160	0.157	0.217
SARI	0.044	0.164	0.316
BERTScore	0.196	0.307	0.294
LLM-Judge	0.301	0.511	0.471
DETECT	0.215	0.289	0.429

i) *Human-Judge* vs. *LLM-Judge*



(a) Simplicity Score

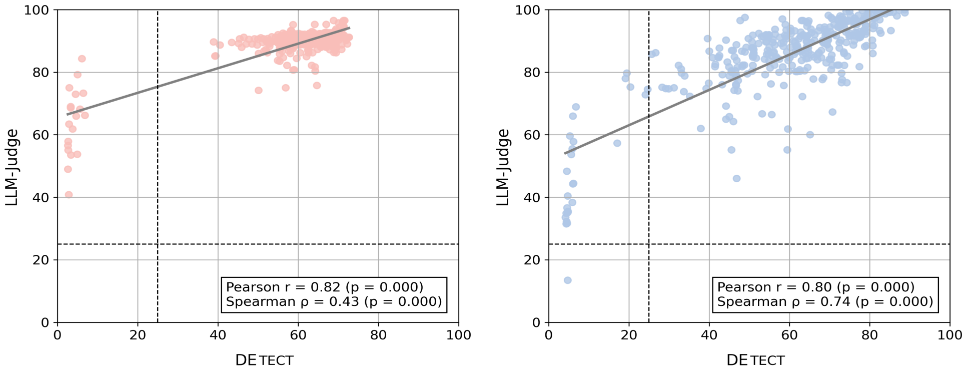
(b) Meaning Preservation Score



(c) Fluency Score

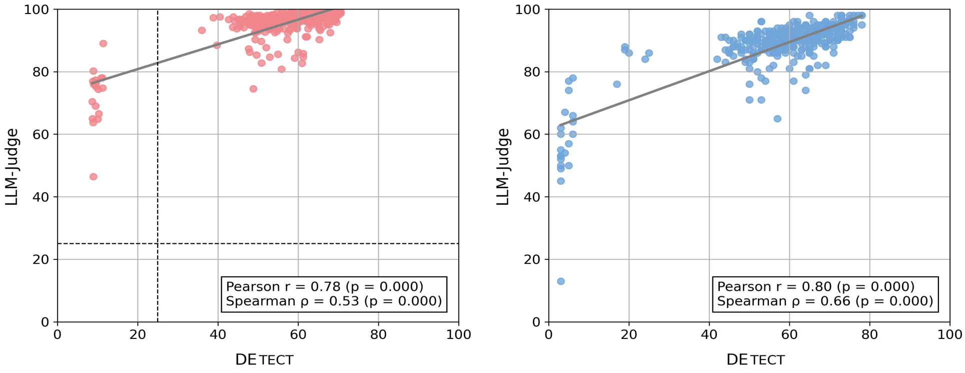
(d) Total Score

ii) *LLM-Judge* vs. DETECT



(a) Simplicity Score

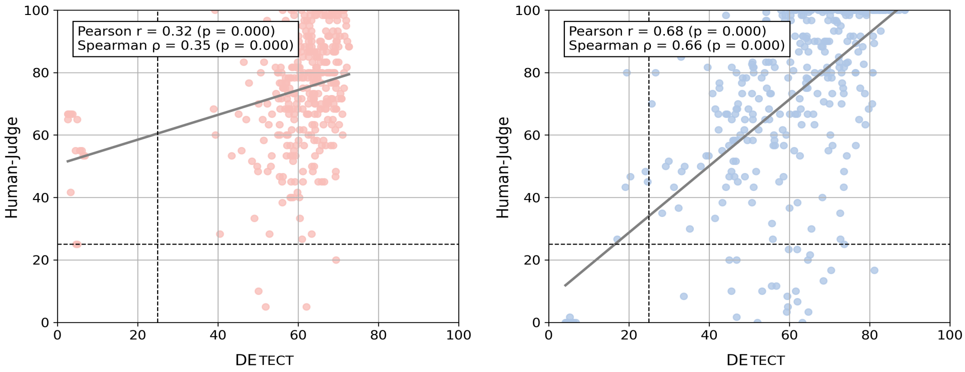
(b) Meaning Preservation Score



(c) Fluency Score

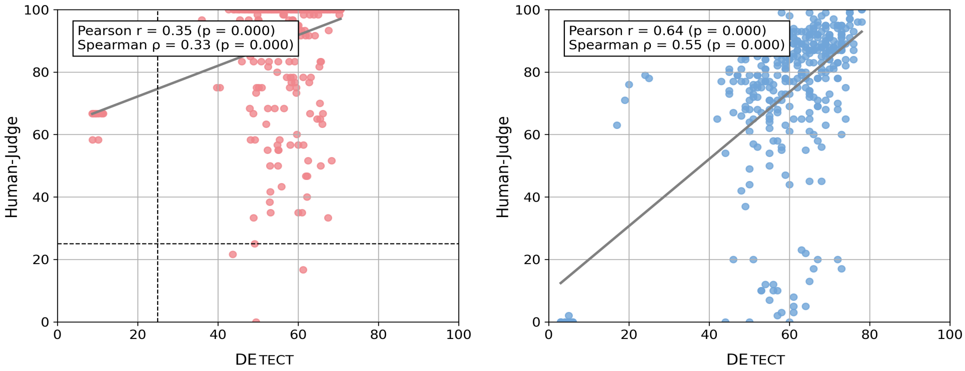
(d) Total Score

iii) *Human-Judge* vs.DETECT



(a) Simplicity Score

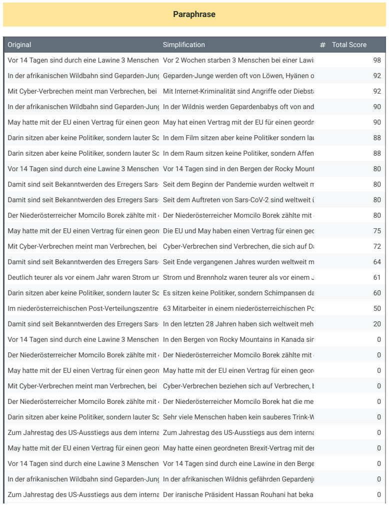
(b) Meaning Preservation Score



(c) Fluency Score

(d) Total Score

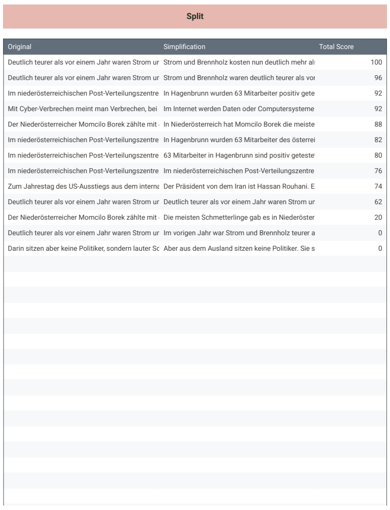
Figure 9: Correlations between *Human-Judge*, *LLM-Judge* and DETECT for each criterion and total scores. Dashed lines indicate "rejection zones" where one model scores below 25.



Paraphrase

Original	Simplification	#	Total Score
Vor 14 Tagen sind durch eine Lawine 3 Menschen	Vor 2 Wochen starben 3 Menschen bei einer Lawi		98
In der afrikanischen Wildbahn sind Geparden-Jun	Geparden-Junge werden oft von Löwen, Hyänen o		92
Mit Cyber-Verbrechen meint man Verbrechen, bei	Mit Internet-Kriminalität sind Angriffe oder Diebst		92
In der afrikanischen Wildbahn sind Geparden-Jun	In der Wildnis werden Gepardenbabys oft von and		90
May hatte mit der EU einen Vertrag für einen geor	May hat einen Vertrag mit der EU für einen geordn		90
Darin sitzen aber keine Politiker, sondern lauter Sc	In dem Film sitzen aber keine Politiker sondern la		88
Darin sitzen aber keine Politiker, sondern lauter Sc	In dem Raum sitzen keine Politiker, sondern Affen		88
Vor 14 Tagen sind durch eine Lawine 3 Menschen	Vor 14 Tagen sind in den Bergen der Rocky Moun		80
Damit sind seit Bekanntwerden des Erregers Sars-	Seit dem Beginn der Pandemie wurden weltweit n		80
Damit sind seit Bekanntwerden des Erregers Sars-	Seit dem Auftreten von Sars-CoV-2 sind weltweit		80
Der Niederösterreicher Momcilo Borek zählte mit	Der Niederösterreicher Momcilo Borek zählte mit		80
May hatte mit der EU einen Vertrag für einen geor	Die EU und May haben einen Vertrag für einen ge		75
Mit Cyber-Verbrechen meint man Verbrechen, bei	Cyber-Verbrechen sind Verbrechen, die sich auf D		72
Damit sind seit Bekanntwerden des Erregers Sars-	Seit Ende vergangenen Jahres wurden weltweit m		64
Deutlich teurer als vor einem Jahr waren Strom un	Strom und Brennholz waren teurer als vor einem J		61
Darin sitzen aber keine Politiker, sondern lauter Sc	Es sitzen keine Politiker, sondern Schimpansen da		60
Im niederösterreichischen Post-Verteilungszentre	63 Mitarbeiter in einem niederösterreichischen Pc		50
Damit sind seit Bekanntwerden des Erregers Sars-	In den letzten 28 Jahren haben sich weltweit meh		20
Vor 14 Tagen sind durch eine Lawine 3 Menschen	In den Bergen von Rocky Mountains in Kanada sir		0
Der Niederösterreicher Momcilo Borek zählte mit	Der Niederösterreicher Momcilo Borek zählte mit		0
May hatte mit der EU einen Vertrag für einen geor	May hatte mit der EU einen Vertrag für einen geor		0
Mit Cyber-Verbrechen meint man Verbrechen, bei	Cyber-Verbrechen beziehen sich auf Verbrechen,		0
Der Niederösterreicher Momcilo Borek zählte mit	Der Niederösterreicher Momcilo Borek hat die me		0
Darin sitzen aber keine Politiker, sondern lauter Sc	Sehr viele Menschen haben kein sauberes Trink-W		0
Zum Jahrestag des US-Ausstiegs aus dem interna	Zum Jahrestag des US-Ausstiegs aus dem interna		0
May hatte mit der EU einen Vertrag für einen geor	May hatte einen geordneten Brexit-Vertrag mit de		0
Vor 14 Tagen sind durch eine Lawine 3 Menschen	Vor 14 Tagen sind durch eine Lawine in den Berge		0
In der afrikanischen Wildbahn sind Geparden-Jun	In der afrikanischen Wildnis gefährden Gepardenj		0
Zum Jahrestag des US-Ausstiegs aus dem interna	Der iranische Präsident Hassan Rouhani hat beka		0

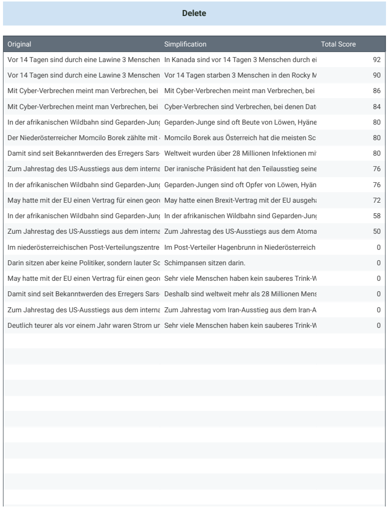
(a) *Paraphrase* samples annotation.



Split

Original	Simplification	Total Score
Deutlich teurer als vor einem Jahr waren Strom ur	Strom und Brennholz kosten nun deutlich mehr al:	100
Deutlich teurer als vor einem Jahr waren Strom ur	Strom und Brennholz waren deutlich teurer als vor	96
Im niederösterreichischen Post-Verteilungszentre	In Hagenbrunn wurden 63 Mitarbeiter positiv gete	92
Mit Cyber-Verbrechen meint man Verbrechen, bei	Im Internet werden Daten oder Computersysteme	92
Der Niederösterreicher Momcilo Borek zählte mit	In Niederösterreich hat Momcilo Borek die meiste	88
Im niederösterreichischen Post-Verteilungszentre	In Hagenbrunn wurden 63 Mitarbeiter des österrei	82
Im niederösterreichischen Post-Verteilungszentre	63 Mitarbeiter in Hagenbrunn sind positiv geteste	80
Im niederösterreichischen Post-Verteilungszentre	Im niederösterreichischen Post-Verteilungszentre	76
Zum Jahrestag des US-Ausstiegs aus dem interna	Der Präsident von dem Iran ist Hassan Rouhani. E	74
Deutlich teurer als vor einem Jahr waren Strom ur	Deutlich teurer als vor einem Jahr waren Strom ur	62
Der Niederösterreicher Momcilo Borek zählte mit	Die meisten Schmetterlinge gab es in Niederöster	20
Deutlich teurer als vor einem Jahr waren Strom ur	Im vorigen Jahr war Strom und Brennholz teurer a	0
Darin sitzen aber keine Politiker, sondern lauter Sc	Aber aus dem Ausland sitzen keine Politiker. Sie s	0

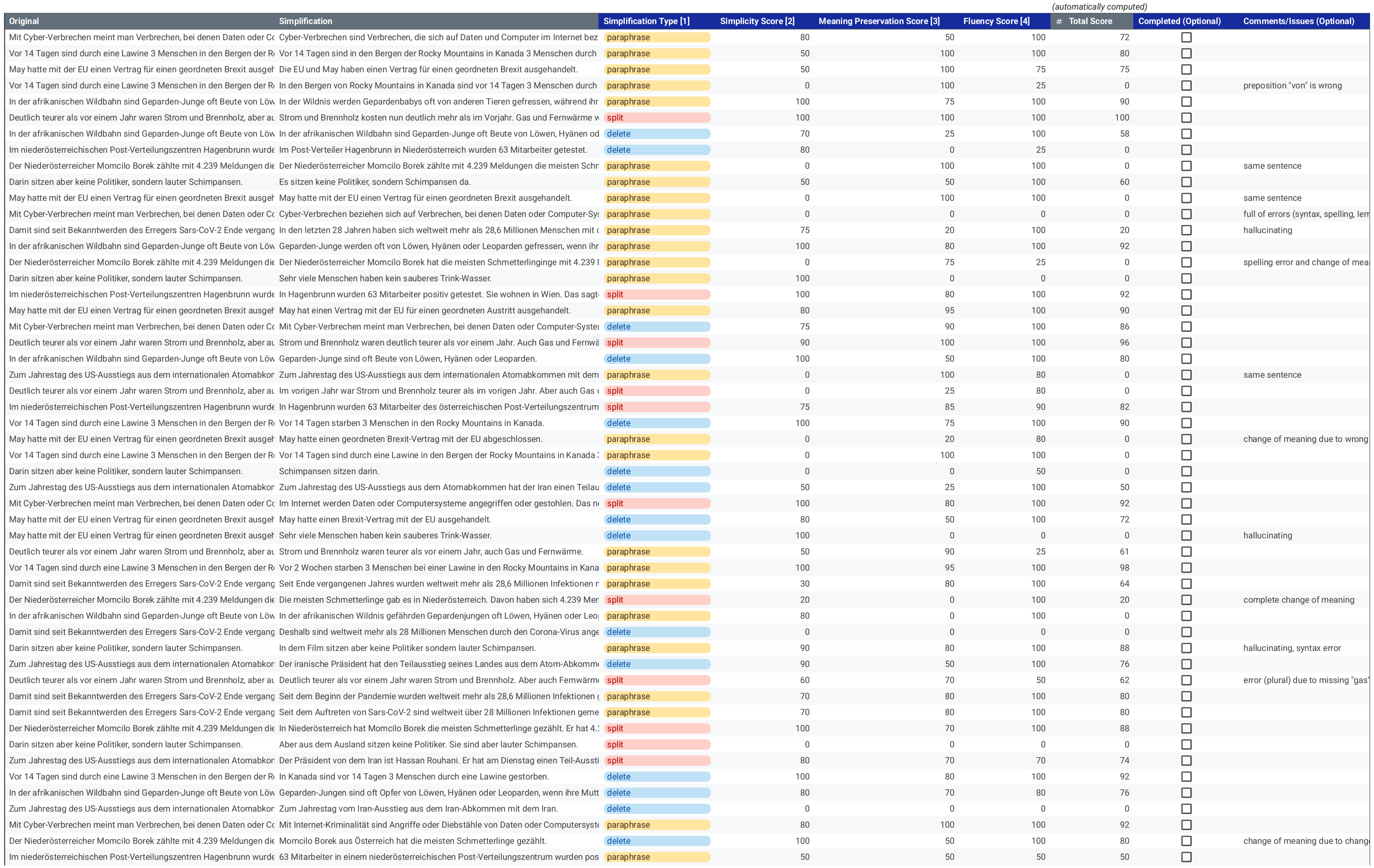
(b) *Split* samples annotation.



Delete

Original	Simplification	Total Score
Vor 14 Tagen sind durch eine Lawine 3 Menschen	In Kanada sind vor 14 Tagen 3 Menschen durch ei	92
Vor 14 Tagen sind durch eine Lawine 3 Menschen	Vor 14 Tagen starben 3 Menschen in den Rocky M	90
Mit Cyber-Verbrechen meint man Verbrechen, bei	Mit Cyber-Verbrechen meint man Verbrechen, bei	86
Mit Cyber-Verbrechen meint man Verbrechen, bei	Cyber-Verbrechen sind Verbrechen, bei denen Dat	84
In der afrikanischen Wildbahn sind Geparden-Jun	Geparden-Junge sind oft Beute von Löwen, Hyäne	80
Der Niederösterreicher Momcilo Borek zählte mit	Momcilo Borek aus Österreich hat die meisten Sc	80
Damit sind seit Bekanntwerden des Erregers Sars-	Weltweit wurden über 28 Millionen Infektionen mi	80
Zum Jahrestag des US-Ausstiegs aus dem interna	Der iranische Präsident hat den Teilausstieg seine	76
In der afrikanischen Wildbahn sind Geparden-Jun	Geparden-Jungen sind oft Opfer von Löwen, Hyän	76
May hatte mit der EU einen Vertrag für einen geor	May hatte einen Brexit-Vertrag mit der EU ausgeh	72
In der afrikanischen Wildbahn sind Geparden-Jun	In der afrikanischen Wildbahn sind Geparden-Jun	58
Zum Jahrestag des US-Ausstiegs aus dem interna	Zum Jahrestag des US-Ausstiegs aus dem Atoma	50
Im niederösterreichischen Post-Verteilungszentre	Im Post-Verteiler Hagenbrunn in Niederösterreich	0
Darin sitzen aber keine Politiker, sondern lauter Sc	Schimpansen sitzen darin.	0
May hatte mit der EU einen Vertrag für einen geor	Sehr viele Menschen haben kein sauberes Trink-W	0
Damit sind seit Bekanntwerden des Erregers Sars-	Deshalb sind weltweit mehr als 28 Millionen Mens	0
Zum Jahrestag des US-Ausstiegs aus dem interna	Zum Jahrestag vom Iran-Ausstieg aus dem Iran-A	0
Deutlich teurer als vor einem Jahr waren Strom ur	Sehr viele Menschen haben kein sauberes Trink-W	0

(c) *Delete* samples annotation.



(automatically computed)

Original	Simplification	Simplification Type [1]	Simplicity Score [2]	Meaning Preservation Score [3]	Fluency Score [4]	# Total Score	Completed (Optional)	Comments/Issues (Optional)
Mit Cyber-Verbrechen meint man Verbrechen, bei denen Daten oder C	Cyber-Verbrechen sind Verbrechen, die sich auf Daten und Computer im Internet bez	paraphrase	80	50	100	72	☐	
Vor 14 Tagen sind durch eine Lawine 3 Menschen in den Bergen der R	Vor 14 Tagen sind in den Bergen der Rocky Mountains in Kanada 3 Menschen durch	paraphrase	50	100	100	80	☐	
May hatte mit der EU einen Vertrag für einen geordneten Brexit ausge	Die EU und May haben einen Vertrag für einen geordneten Brexit ausgehandelt.	paraphrase	50	100	75	75	☐	
Vor 14 Tagen sind durch eine Lawine 3 Menschen in den Bergen der R	In den Bergen von Rocky Mountains in Kanada sind vor 14 Tagen 3 Menschen durch	paraphrase	0	100	25	0	☐	preposition "von" is wrong
In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löw	In der Wildnis werden Gepardenbabys oft von anderen Tieren gefressen, während ihr	paraphrase	100	75	100	90	☐	
Deutlich teurer als vor einem Jahr waren Strom und Brennholz, aber a	Strom und Brennholz kosten nun deutlich mehr als im Vorjahr. Gas und Fernwärme w	split	100	100	100	100	☐	
In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löw	In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löwen, Hyänen od	delete	70	25	100	58	☐	
Im niederösterreichischen Post-Verteilungszentren Hagenbrunn wurde	Im Post-Verteiler Hagenbrunn in Niederösterreich wurden 63 Mitarbeiter getestet.	delete	80	0	25	0	☐	
Der Niederösterreicher Momcilo Borek zählte mit 4.239 Meldungen die	Der Niederösterreicher Momcilo Borek zählte mit 4.239 Meldungen die meisten Sch	paraphrase	0	100	100	0	☐	same sentence
Darin sitzen aber keine Politiker, sondern lauter Schimpansen.	Es sitzen keine Politiker, sondern Schimpansen da.	paraphrase	50	50	100	60	☐	
May hatte mit der EU einen Vertrag für einen geordneten Brexit ausge	May hatte mit der EU einen Vertrag für einen geordneten Brexit ausgehandelt.	paraphrase	0	100	100	0	☐	same sentence
Mit Cyber-Verbrechen meint man Verbrechen, bei denen Daten oder C	Cyber-Verbrechen beziehen sich auf Verbrechen, bei denen Daten oder Computer-Sy	paraphrase	0	0	0	0	☐	full of errors (syntax, spelling, le
Damit sind seit Bekanntwerden des Erregers Sars-CoV-2 Ende vergang	In den letzten 28 Jahren haben sich weltweit mehr als 28,6 Millionen Menschen mit	paraphrase	75	20	100	20	☐	hallucinating
In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löw	Geparden-Junge werden oft von Löwen, Hyänen oder Leoparden gefressen, wenn ihr	paraphrase	100	80	100	92	☐	
Der Niederösterreicher Momcilo Borek zählte mit 4.239 Meldungen die	Der Niederösterreicher Momcilo Borek hat die meisten Schmetterlinginge mit 4.239	paraphrase	0	75	25	0	☐	spelling error and change of mea
Darin sitzen aber keine Politiker, sondern lauter Schimpansen.	Sehr viele Menschen haben kein sauberes Trink-Wasser.	paraphrase	100	0	0	0	☐	
Im niederösterreichischen Post-Verteilungszentren Hagenbrunn wurde	In Hagenbrunn wurden 63 Mitarbeiter positiv getestet. Sie wohnen in Wien. Das sagt	split	100	80	100	92	☐	
May hatte mit der EU einen Vertrag für einen geordneten Brexit ausge	May hat einen Vertrag mit der EU für einen geordneten Austritt ausgehandelt.	paraphrase	80	95	100	90	☐	
Mit Cyber-Verbrechen meint man Verbrechen, bei denen Daten oder C	Mit Cyber-Verbrechen meint man Verbrechen, bei denen Daten oder Computer-Syste	delete	75	90	100	86	☐	
Deutlich teurer als vor einem Jahr waren Strom und Brennholz, aber a	Strom und Brennholz waren deutlich teurer als vor einem Jahr. Auch Gas und Fernwä	split	90	100	100	96	☐	
In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löw	Geparden-Junge sind oft Beute von Löwen, Hyänen oder Leoparden.	delete	100	50	100	80	☐	
Zum Jahrestag des US-Ausstiegs aus dem internationalen Atomabko	Zum Jahrestag des US-Ausstiegs aus dem internationalen Atomabkommen mit dem	paraphrase	0	100	80	0	☐	same sentence
Deutlich teurer als vor einem Jahr waren Strom und Brennholz, aber a	Im vorigen Jahr war Strom und Brennholz teurer als im vorigen Jahr. Aber auch Gas	split	0	25	80	0	☐	
Im niederösterreichischen Post-Verteilungszentren Hagenbrunn wurde	In Hagenbrunn wurden 63 Mitarbeiter des österreichischen Post-Verteilungszentrum	split	75	85	90	82	☐	
Vor 14 Tagen sind durch eine Lawine 3 Menschen in den Bergen der R	Vor 14 Tagen starben 3 Menschen in den Rocky Mountains in Kanada.	delete	100	75	100	90	☐	
May hatte mit der EU einen Vertrag für einen geordneten Brexit ausge	May hatte einen geordneten Brexit-Vertrag mit der EU abgeschlossen.	paraphrase	0	20	80	0	☐	change of meaning due to wrong
Vor 14 Tagen sind durch eine Lawine 3 Menschen in den Bergen der R	Vor 14 Tagen sind durch eine Lawine in den Bergen der Rocky Mountains in Kanada	paraphrase	0	100	100	0	☐	
Darin sitzen aber keine Politiker, sondern lauter Schimpansen.	Schimpansen sitzen darin.	delete	0	0	50	0	☐	
Zum Jahrestag des US-Ausstiegs aus dem internationalen Atomabko	Zum Jahrestag des US-Ausstiegs aus dem Atomabkommen hat der Iran einen Teilau	delete	50	25	100	50	☐	
Mit Cyber-Verbrechen meint man Verbrechen, bei denen Daten oder C	Im Internet werden Daten oder Computersysteme angegriffen oder gestohlen. Das n	split	100	80	100	92	☐	
May hatte mit der EU einen Vertrag für einen geordneten Brexit ausge	May hatte einen Brexit-Vertrag mit der EU ausgehandelt.	delete	80	50	100	72	☐	
May hatte mit der EU einen Vertrag für einen geordneten Brexit ausge	Sehr viele Menschen haben kein sauberes Trink-Wasser.	delete	100	0	0	0	☐	hallucinating
Deutlich teurer als vor einem Jahr waren Strom und Brennholz, aber a	Strom und Brennholz waren teurer als vor einem Jahr, auch Gas und Fernwärme.	paraphrase	50	90	25	61	☐	
Vor 14 Tagen sind durch eine Lawine 3 Menschen in den Bergen der R	Vor 2 Wochen starben 3 Menschen bei einer Lawine in den Rocky Mountains in Kana	paraphrase	100	95	100	98	☐	
Damit sind seit Bekanntwerden des Erregers Sars-CoV-2 Ende vergang	Seit Ende vergangenen Jahres wurden weltweit mehr als 28,6 Millionen Infektionen	paraphrase	30	80	100	64	☐	
Der Niederösterreicher Momcilo Borek zählte mit 4.239 Meldungen die	Die meisten Schmetterlinge gab es in Niederösterreich. Davon haben sich 4.239 Me	split	20	0	100	20	☐	complete change of meaning
In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löw	In der afrikanischen Wildnis gefährden Gepardenjungen oft Löwen, Hyänen oder Leo	paraphrase	80	0	100	0	☐	
Damit sind seit Bekanntwerden des Erregers Sars-CoV-2 Ende vergang	Deshalb sind weltweit mehr als 28 Millionen Menschen durch den Corona-Virus ange	delete	0	0	0	0	☐	
Darin sitzen aber keine Politiker, sondern lauter Schimpansen.	In dem Film sitzen aber keine Politiker sondern lauter Schimpansen.	paraphrase	90	80	100	88	☐	hallucinating, syntax error
Zum Jahrestag des US-Ausstiegs aus dem internationalen Atomabko	Der iranische Präsident hat den Teilausstieg seines Landes aus dem Atom-Abkomm	delete	90	50	100	76	☐	
Deutlich teurer als vor einem Jahr waren Strom und Brennholz, aber a	Deutlich teurer als vor einem Jahr waren Strom und Brennholz. Aber auch Fernwärm	split	60	70	50	62	☐	error (plural) due to missing "gas"
Damit sind seit Bekanntwerden des Erregers Sars-CoV-2 Ende vergang	Seit dem Beginn der Pandemie wurden weltweit mehr als 28,6 Millionen Infektionen	paraphrase	70	80	100	80	☐	
Damit sind seit Bekanntwerden des Erregers Sars-CoV-2 Ende vergang	Seit dem Auftreten von Sars-CoV-2 sind weltweit über 28 Millionen Infektionen geme	paraphrase	70	80	100	80	☐	
Der Niederösterreicher Momcilo Borek zählte mit 4.239 Meldungen die	In Niederösterreich hat Momcilo Borek die meisten Schmetterlinge gezählt. Er hat 4.	split	100	70	100	88	☐	
Darin sitzen aber keine Politiker, sondern lauter Schimpansen.	Aber aus dem Ausland sitzen keine Politiker. Sie sind aber lauter Schimpansen.	split	0	0	0	0	☐	
Zum Jahrestag des US-Ausstiegs aus dem internationalen Atomabko	Der Präsident von dem Iran ist Hassan Rouhani. Er hat am Dienstag einen Teil-Aussti	split	80	70	70	74	☐	
Vor 14 Tagen sind durch eine Lawine 3 Menschen in den Bergen der R	In Kanada sind vor 14 Tagen 3 Menschen durch eine Lawine gestorben.	delete	100	80	100	92	☐	
In der afrikanischen Wildbahn sind Geparden-Junge oft Beute von Löw	Geparden-Jungen sind oft Opfer von Löwen, Hyänen oder Leoparden, wenn ihre Mutt	delete	80	70	80	76	☐	
Zum Jahrestag des US-Ausstiegs aus dem internationalen Atomabko	Zum Jahrestag vom Iran-Ausstieg aus dem Iran-Abkommen mit dem Iran.	delete	0	0	0	0	☐	
Mit Cyber-Verbrechen meint man Verbrechen, bei denen Daten oder C	Mit Internet-Kriminalität sind Angriffe oder Diebstähle von Daten oder Computersyst	paraphrase	80	100	100	92	☐	
Der Niederösterreicher Momcilo Borek zählte mit 4.239 Meldungen die	Momcilo Borek aus Österreich hat die meisten Schmetterlinge gezählt.	delete	100	50	100	80	☐	change of meaning due to chang
Im niederösterreichischen Post-Verteilungszentren Hagenbrunn wurde	63 Mitarbeiter in einem niederösterreichischen Post-Verteilungszentrum wurden pos	paraphrase	50	50	50	50	☐	

(d) Overall grading tool interface.

Figure 10: Ranking of *Human-Judge* quality scores. Following the RANK & RATE protocol by Maddela et al. (2023), we used Google Sheet to collect *Human-Judge* scores. Before kicking-off the grading, we hosted instruction sessions with the human experts to show them how to work with the tool.