German4All – A Dataset and Model for Readability-Controlled Paraphrasing in German

Miriam Anschütz, Thanh Mai Pham,
Eslam Nasrallah, Maximilian Müller, Cristian-George Craciun and **Georg Groh**
Technical University of Munich
miriam.anschuetz@tum.de, grohg@cit.tum.de

Abstract

The ability to paraphrase texts across different complexity levels is essential for creating accessible texts that can be tailored toward diverse reader groups. Thus, we introduce **German4All**, the first large-scale German dataset of aligned readability-controlled, paragraph-level paraphrases. It spans five readability levels and comprises over 25,000 samples. The dataset is automatically synthesized using GPT-4 and rigorously evaluated through both human and LLM-based judgments.

Using German4All, we train an open-source, readability-controlled paraphrasing model that achieves state-of-the-art performance in German text simplification, enabling more nuanced and reader-specific adaptations. We open-source both the dataset and the model to encourage further research on multi-level paraphrasing.

1 Introduction

Text simplification is typically approached as a standardized process, where an input text is simplified to a single, pre-defined complexity level—often determined by the model's training data. However, the audience for simplified language is highly diverse, including people with different reading proficiencies and backgrounds (Stajner, 2021). To better address this diversity, some languages such as German and Spanish differentiate between multiple simplification levels—for example, "plain language" and "easy-to-read" (DE: Leichte Sprache, ES: lectura fácil) (Maaß, 2020; Madina et al., 2023). Despite this, existing simplification systems for German generally treat simplification as a one-size-fits-all task targeting a single output level.

In contrast, English NLP research has explored methods for adapting simplification to multiple complexity levels (Chi et al., 2023; Barayan et al., 2025; Farajidizaji et al., 2024). Yet, such fine-grained control remains underexplored in non-English contexts due to a lack of suitable resources.

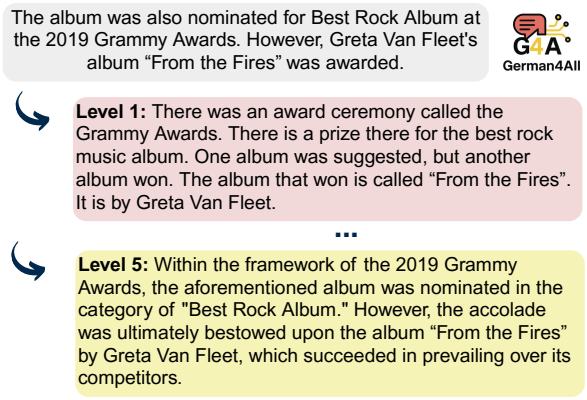


Figure 1: Example from our German4All dataset, translated to English. One input text is paraphrased into five versions at different complexity levels.

While powerful LLMs like GPT-4 can provide high-quality paraphrases in multiple languages, these models are harder to control for specific task settings. Moreover, users have to rely on API endpoints, introducing privacy concerns, or need to deploy large models in their environment (Toshevska and Gievska, 2025). Thus, we identify a need for smaller, task-specific models for readability-controlled paraphrasing.

To address this gap, we introduce German4All, the first large-scale, multi-level paraphrasing corpus for German. The dataset provides paraphrases of source texts across five distinct complexity levels. These levels are defined by their respective target group and range from people with reading difficulties to academic experts. For this paper, we refer to text complexity, readability, and simplification level as similar concepts.

The proposed dataset supports a range of tasks, including simplification, complexification, and readability-controlled paraphrasing. Moreover, the aligned paraphrases can be used in various settings, including iterative simplification or readability assessment. Overall, our contribution can be summarized as follows:

Proceedings of the 18th International Natural Language Generation Conference, pages 390–407
October 29–November 2, 2025.

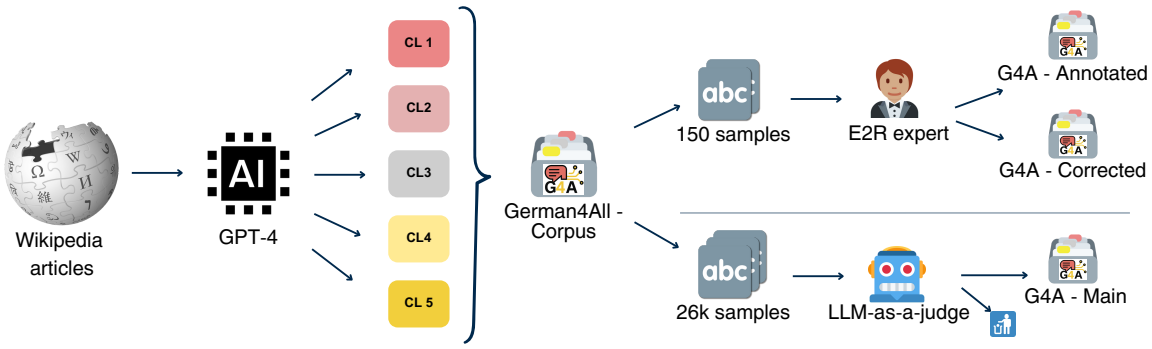


Figure 2: Overview of our dataset creation approach: We take paragraphs from Wikipedia and use GPT-4 to paraphrase them into five different complexity levels (CLs). Then, we manually curate a test set wth 150 samples, while automatically evaluating the main dataset with an LLM judge.

- We release German4All, a large-scale German dataset containing Wikipedia inputs with paraphrases to five different readability levels.
- While the corpus itself is synthesized using GPT4, we conduct a comprehensive evaluation—including an LLM-as-a-judge and feedback from 16 human participants—to validate its quality and usefulness for downstream tasks[1].
- We train a readability-controlled German simplification model on this dataset, which shows state-of-the-art performance compared to existing systems and reflects the styles of different versions of simplified language.

2 Related work

German text simplification German exhibits different styles of simplified language. While simple/-plain language (DE: "einfache Sprache") is a generally simpler version of standard German targeted to a broad audience, such as language learners (DIN-Normenausschuss Ergonomie, 2024), easy language (DE: "Leichte Sprache") is specifically tailored towards people with mental disabilities and reading deficiencies. As such, the simplifications in easy language are way stronger with shorter sentences and more guidance for the reader, e.g., through line breaks after every sentence (DIN-Normenausschuss Ergonomie, 2025). Many NLP resources exist for easy language (Schomacker et al., 2023; Anschütz et al., 2023; Toborek et al., 2023), simple language (Stodden, 2024b; Fruth et al., 2024), and other, more specific target groups like children (Aumiller and Gertz, 2022). In addition, some works provide resources in multiple difficulty levels (Stodden et al., 2023), most often following the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). As such, Spring et al. (2021) were the first to explore readability-controlled simplification for German. Despite these resources, there is a lack of large-scale and publicly available data for multi-level paraphrasing. We close this gap by providing a dataset with aligned paragraphs of different complexity levels. It not only supports simplification but also complexification to higher complexity levels. Complexification can be especially interesting for language learners who want to gradually improve their language proficiency. Moreover, input texts of various complexity can lead to a better generalization in simplification (Chi et al., 2023).

Synthetic data generation We rely on synthetic data created with GPT-4 to compile a large-scale, multi-level dataset. Existing resources often rely on or extend other resources (e.g., the Simple German Corpus by Toborek et al., 2023 is contained in the resources by Anschütz et al., 2023). Thus, synthesizing a dataset enables the use of novel data and thus extends the diversity of available datasets. A similar approach was chosen by Klöser et al. (2024), who synthesized complex data from existing human-created simplifications in German to obtain an aligned training corpus. In contrast to this, all our complexity levels are generated by GPT-4, giving us more control over the lexical diversity and the information preservation across all complexity levels. Malik et al. (2024) benchmarked different open- and closed-source models for their ability to adhere to complexity guidance in English and found that only GPT-4 succeeds at this task without further fine-tuning. Similar results were reported by Almasi and Kristensen-McLachlan (2025) who performed similar experiments for Spanish. While

[1] https://github.com/MiriUll/German4All

they find that GPT-4 loses guidance for longer chat histories, the initial answers comply with the complexity guidance in the system prompt.

3 Methodology

An overview of our data creation process is presented in Figure 2. The inputs in our dataset are paragraphs from Wikipedia. For this, we selected the Wikipedia dump from December 2022 in the Cohere/wikipedia-22-12 dataset. This dataset contains Wikipedia paragraphs with at least 100 characters in multiple languages and sorts them by the number of views of their corresponding page. From this data, we randomly selected 26,665 samples from the 3 million most popular German paragraphs as our main data and 150 samples from the 1 million most popular German paragraphs as a test set, while assuring that the two subsets have no overlap. In the full dataset, the upper quartile of words per paragraph is 76 words. Hence, we excluded paragraphs with more than 80 words to ensure a diverse yet consistent paragraph length.

3.1 Complexity levels

Previous works by Chi et al. (2023) or Spring et al. (2021) have defined their complexity levels with existing CEFR levels (Council of Europe, 2001). As such, the language levels of A1, A2, and partly B1 can be considered simple. Yet, these levels were mostly defined for language learners rather than accounting for language barriers of native speakers (Heine, 2017). Hence, the definitions of the different levels focus on the vocabulary and use cases at that level, e.g., introducing yourself in level A1. In contrast, people with learning disabilities have different backgrounds and require info about their everyday lives in an accessible language that goes beyond basic communication as a tourist.

Thus, we do not use CEFR levels but create our own complexity levels that are defined by the respective target groups and aligned with the literacy proficiency levels defined by the OECD (2013, p. 64). We distinguish five complexity levels between easy language for people with reading difficulties (level 1), commonly used language for the general public (level 3), and academic language for experts (level 5). The fine-grained definitions are presented in Appendix A. Most input texts from Wikipedia are between levels 3 and 4, but as we cannot control for their level, they do not count towards the five levels provided.

3.2 Synthetic data generation

Empirical studies by Manning (2024) and Barayan et al. (2025) showed that ChatGPT with GPT-4 as backend can provide competitive and high-quality simplifications. Therefore, we used *gpt-4-turbo-2024-04-09* via the OpenAI Batch API to create our complexity-aware paraphrases. The model's system prompt was "*You are an expert in adapting texts to different complexity levels.*". The more detailed user prompt is shown in subsection F.1. As suggested in the OpenAI engineering guide, we structured the prompt into subsections. In these sections, we define our complexity levels, provide a 1-shot example and further details about the task, guide the model to pay attention to the previously described features of each of the complexity levels, and finally define the JSON output format. Our 1-shot example was randomly selected and manually paraphrased to all five complexity levels. Even though few-shot examples have been shown to improve the output quality (Malik et al., 2024), we tried to find a balance between performance and cost due to the number of input tokens. Therefore, we only provided one example in the prompt.

3.3 Data filtering

Once the paraphrases were synthesized, we employed various automatic filtering steps to ensure a high overall quality of our dataset:

- **Valid JSON format**: The output should be provided in a valid JSON format, and the paraphrases should be ordered by their complexity.
- **Out-of-vocabulary tokens**: We used spaCy (Honnibal et al., 2020) to flag samples with at least three consecutive tokens not in spaCy's vocabulary. These samples were manually reviewed and filtered for false positives.
- **Valid German text**: Using langdetect, we identified all samples containing other languages. Again, these samples were manually reviewed.

All samples that were still flagged as erroneous after the manual review were removed. In addition, we manually inspected random samples and discarded a minor proportion of them. Ultimately, 26,337 out of 26,665 samples remained as the initial version of our German4All-Main dataset.

3.4 Manual test data correction

As outlined before, we sampled a subset of 150 distinct paragraphs from the Wikipedia corpus as

Split	Description	#Samples	#Paraphrases
main	Main dataset, primarily for training	25,459	5 (CL 1-5)
corrected	Manually corrected samples, extended with Leichte Sprache (LS) paraphrases	150	6 (CL 1-5+LS)
annotated	Model outputs together with the manually corrected samples, annotated by the model's shortcomings	132	2

Table 1: Dataset splits and statistics. Each sample contains the original text together with different numbers of paraphrases at different complexity levels (CL).

our test set. This test set was manually revised and corrected by two native speakers, one of them being a Leichte Sprache expert. The 150 manually corrected samples form our German4All-Corrected subset. To enhance traceability and to facilitate further experiments, we annotated the changes that we performed during correction and stored these operations in the German4All-Annotated dataset. This dataset contains triplets of original texts, the original GPT-4 paraphrases to one specific complexity, and our corrected paraphrases. Our correction operations are categorized into six distinct operation types. We distinguish between these operation annotations:

- *removed_info*: Whether we removed (potentially erroneous) information in the correction
- *added_info*: Whether we added information in the correction that was missing from the input
- *corrected_info*: Whether we fixed information in the correction
- *adjusted_complexity*: Whether we corrected the language level to match the target complexity
- *corrected_language*: Whether we corrected language errors
- *hallucination_in_origin*: Whether the original model output contains a hallucination

Furthermore, we manually created Leichte Sprache paraphrases for all samples in the corrected corpus. For this, we generated Leichte Sprache candidates with EasyJon, a free-to-use Leichte Sprache translation tool with an *anthropic/claude-3.5-sonnet* backend (Barbu, 2024). Then, the samples were manually rewritten by a German Easy Language expert.

Overall, this process returned three German4All subsets. An overview is provided in Table 1. While the main corpus is mainly useful for training, the corrected corpus can serve as a gold-standard evaluation dataset or be used for RLHF approaches.

4 Dataset evaluation

Our synthesized data contains more than 25k samples with five complexity level paraphrases each, resulting in a dataset of more than 125k text pairs. Due to this size, a human review of all samples would be infeasible. Therefore, we investigated two different evaluation angles. First, we randomly selected samples from the corpus and performed a human evaluation on these samples. Then, we extended our evaluation to the full corpus by using an LLM-as-a-judge.

4.1 Human evaluation

For the human evaluation, we selected 15 samples with five paraphrases each (=75 text pairs in total) from the Main subset at random. These samples were split into five groups with three samples and 15 original-paraphrased text pairs (3 samples * 5 complexity levels) per group. The paraphrase pairs were presented one by one, grouped by their original text and sorted by their complexity level. For each text pair, we asked the following questions with the answer options in parentheses. The evaluation was originally conducted in German, but we translated it here:

Q1 The paraphrase reflects the content of the original text ... [*incorrectly* | *approximately* | *correctly*]

Q2 How often were pieces of information from the original text omitted in the paraphrase? [*never* | *seldom* | *sometimes* | *often*]

Q3 How often were additional pieces of information added in the paraphrase that were not present in the original text? [*never* | *seldom* | *sometimes* | *often*]

Q4 Skip this question if you selected "never" in the previous question. What types of additional information were added? Multiple options can be selected.

[*embellishment* | *explanations or definitions* | *factually incorrect information* | *factually correct information* | *other*]

Q5 The paraphrase for the desired difficulty level is ... [*too easy* | *a bit too easy* | *appropriate* | *a bit too complicated* | *too complicated*]

In total, 16 native German speakers participated in our human evaluation. They participated voluntarily and received no financial compensation. Each participant was assigned to one of the five sample groups, so that we had at least three participants per group. We divided the participants into groups to balance the workload of a single participant while receiving multiple feedback for as many samples as possible, thus increasing the coverage of our evaluation. With this setup, participants needed $\pm$20 minutes to complete their evaluation, and all indeed completed the survey.

	G1	G2	G3	G4	G5	Mean
Q1	0.53	0.29	0.26	0.17	0.28	0.31
Q2	0.73	0.56	-0.05	0.72	0.38	0.47
Q3	0.54	0.18	0.67	0.05	0.56	0.40
Q5	-0.16	0.04	0.13	0.16	0.29	0.09
Q5 Tol.	-0.25	0.2	1.0	1.0	0.31	0.45

Table 2: Inter-annotator agreement measured by Krippendorff's α for questions 1-3 and 5 across annotator groups (G1-G5). For Q5 (complexity level), we also report the agreement with a $\pm$1 tolerance.

Table 2 shows the agreement scores for the different evaluation groups using Krippendorff's α (Krippendorff, 2011). Considering the low number of annotators, the content-related questions show a decent average agreement. Yet, the complexity level question has a mixed result. Since text complexity is a very subjective measure, we also calculated the agreement with a tolerance of one level (e.g., "*too easy*" and "*a bit too easy*" are considered to be an agreement). Still, the agreement for the first evaluation group seems to be very low. However, all three annotators give exactly the same score for a sample in 66% of all pairs, so this is mostly an artifact of the agreement score calculation. Overall, we consider the agreement to be good enough to deduce trends and conclusions from the evaluation.

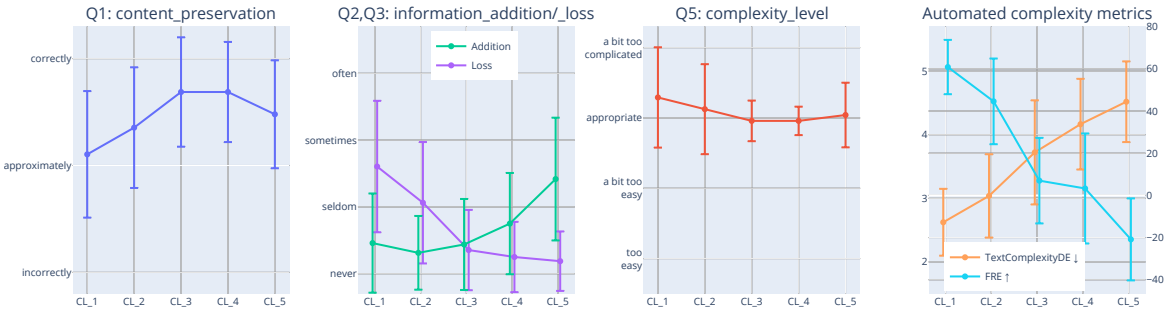
Figure 3a shows the answers for Q1-3 and Q5. The content is best preserved for complexity levels 3 and 4, while levels 1 and 2 show the highest rates of information loss. This is expected for simplifications that leave out (minor) details to improve understanding (Trienes et al., 2024), and thus, a desired feature of the dataset. In contrast, the higher complexity levels add the most information. When looking at the answers from Q4, the most information added in these complexity levels is embellishments and facts. For CL_1, this shifts toward explanations or definitions, which improve the readability of a text. Looking at the complexity levels, we compare the human answers with the Flesh readability ease (FRE) score by Amstad (1978), where a higher FRE score indicates a better readability. In addition, we used the open-source text complexity prediction model for German texts by Anschütz and Groh (2022), trained on the TextComplexityDE dataset (Naderi et al., 2019). Here, the scores range from 1 (easiest) to 7. Most of the texts seem to be appropriate for their respective complexity level, with the strongest human disagreement for the edge levels 1 and 5. Text complexity is a subjective measure, and thus, it can be expected that these levels show a higher answer variation.

Overall, the human evaluation on the randomly selected subset shows that our dataset is of high quality and that the generated paraphrases exhibit the features of their respective complexity levels.

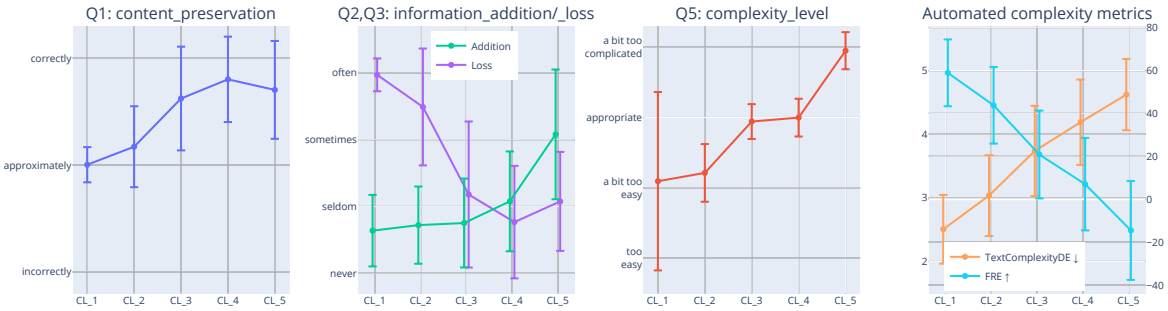
4.2 LLM-as-a-judge

The human evaluation only covers a small proportion of our overall dataset. To acquire a large-scale evaluation of all data samples, we employ an LLM-as-a-judge as automatically evaluating simplification with LLMs has shown good correlation with human judgements in previous work (Liu et al., 2025). We selected *gemma-3-27B-it* as the judge model due to its model size, good multilingual capabilities, and good performance in preliminary experiments (see Appendix B). Our setup reuses the questions and evaluation criteria from the human evaluation. However, every pair of original and paraphrased texts is provided individually without the context of the other paraphrase levels, resulting in $5 * 26,337 = 131,685$ evaluation pairs. The prompt is written in German to avoid code-switching and provides a one-shot example. We utilize structured outputs to force the model to output a parseable JSON output. The full prompt with the task description and further guidance for the model is presented in subsection F.2.

To evaluate and improve our LLM judge, we selected the Annotated subset and tested whether the judge is sensitive to our criteria and correctly annotates the faulty and correct samples. For this, we



(a) Human evaluation results and automated metrics for the respective subset (75 original-paraphrased pairs).



(b) LLM judge evaluation and automated metrics on the full main split (131.365 original-paraphrased pairs).

Figure 3: Evaluation results across questions 1-3 and 5, grouped by the texts' complexity level (CL). The rightmost plot shows two automatic readability measures, TextComplexityDE (Anschütz and Groh, 2022, lower means easier) and Flesh readability scores (Amstad, 1978, higher means easier), to automatically measure the text complexity.

provided the original text and the original model outputs to the LLM judge, and we transferred the textual answers into a numerical scale and performed Spearman correlation tests between our edit operation annotations (compare subsection 3.4) and the judge outputs. For the complexity level, we considered the answers *a bit too easy* and *a bit too complex* as suitable and not requiring an adjustment of the text complexity. We found significant correlations between multiple aspects, including the information_addition (LLM judge) and human hallucination annotations, or complexity_level (LLM judge) and the human complexity adjustments. The full correlation analysis, as well as a comparison between different judge models, is presented in Appendix B. In addition, we provide an example from the dataset and its LLM judge annotations in Appendix E. We concluded that the LLM judge can successfully detect errors in the synthesized outputs, and thus, used it to filter the Main dataset.

Figure 3b shows the LLM judge's evaluation on the full dataset. In general, the curves look similar to the results of our human evaluation. However, the LLM judge seems to be stricter with the information loss in the lower complexity levels, resulting in only an *approximate* content preservation. As discussed before, text simplification focuses only on the most important information to improve understanding. The strongest differences are evident for the perceived text complexity, as the LLM seems to struggle with the definition of an appropriate complexity level and rather rates the absolute complexity of the text. Therefore, the samples in the lowest complexity level are annotated as *(a bit) too easy* and the samples in the highest complexity level as *a bit too complicated*. Since these are our most extreme levels, we expect them to be as easy/complicated as possible. In addition, the automated metrics show that our samples indeed have the target complexity.

Based on the LLM judge's feedback, we further filtered the main split and automatically removed (a number of) samples where:

- the content_preservation is *incorrect* (340).
- the complexity_level is *too easy* for all complexity levels except CL_1 (6).
- the complexity_level is *too complex* for all complexity levels except CL_5 (0).
- the type of added information is *factually incorrect information* (83) or *other* (5), and *factually correct information* for all except CL_5 (67).
- the information_loss is *often*, but only for samples in CL_5 (403).

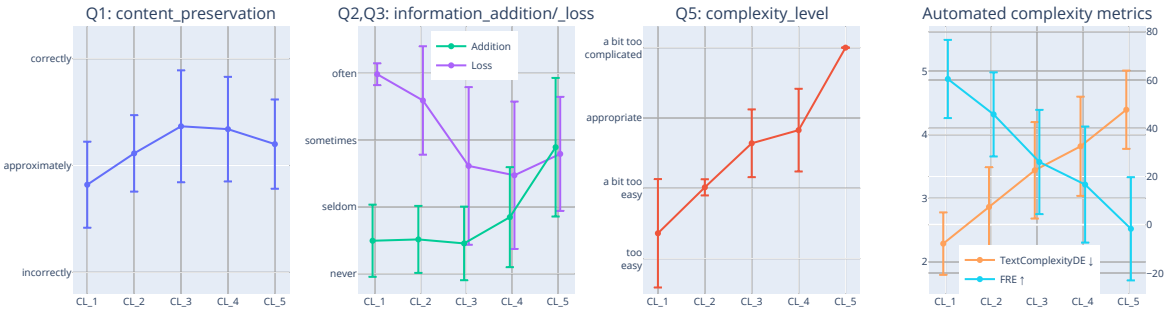


Figure 4: LLM judge evaluation of the distilled model's predictions on the German4All-Corrected test set.

Some samples were flagged by more than one criterion, resulting in 814 removed samples in total, and our final German4All-Main dataset with 25,459 samples. The LLM judge annotations are uploaded in our GitHub repository, so users can create their own filters if needed.

5 Model distillation

The ultimate motivation for our dataset is to create a smaller model that can perform the task of readability-controlled paraphrasing without having to rely on large or expensive models. Therefore, we have inserted and trained LoRA layers (Hu et al., 2022) for a flan-t5-xl model (Chung et al., 2024) that can be inferenced on consumer-sized graphic cards with 12GB of VRAM. We also experimented with different base models such as Llama3.1 8B (Grattafiori et al., 2024) and the German-specific LLäMmlein 7B (Pfister et al., 2025): the former displayed grammatical errors, while the latter struggled with instruction following. Thus, we chose Flan-T5 as base, even though it is already fairly old.

We trained it using a random train:val-80:20 split of the German4All-Main corpus. For every level, we took not only the original input but also the other complexity levels' paraphrases as inputs. This increased our training data by five and leads to a better model generalization as the model sees different styles and complexities as inputs.

For our prompt design, we use the same German complexity level descriptions as for the dataset creation (Appendix A). These descriptions are followed by the task "*Paraphrase the following text to level {level}. {input_text}*" (translated here).

5.1 Evaluating model performance

To evaluate the performance of our distilled model, we benchmarked its performance on the German4All-Corrected test set. Therefore, we created predictions for all complexity levels of all 150 samples, resulting in 750 predictions. Then, we employed the LLM judge and automatic readability metrics from subsection 4.2 to evaluate them. The results are presented in Figure 4: The model successfully learns the characteristics of the different complexity levels and outputs appropriate texts. Yet, from the content perspective, we observe a stronger deviation and a higher content loss compared to the original dataset. To further investigate this, we manually reviewed 75 model predictions (15 per CL). We find that the syntax and style of the samples are of very high quality and match the expected style of the respective level. However, we also found issues with fluency and minor grammatical errors. Moreover, the samples in the higher complexity levels often contain hallucinations, aligning with observations from the LLM judge's *information_addition* criterion. Overall, we find the paraphrases succeed at presenting an input text at different language levels.

5.2 Benchmarking simplification performance

In addition to judging the quality of our model's outputs, we benchmarked and compared its simplification performance compared to other German ATS systems on existing text simplification datasets. For this, we used the EASSE-DE evaluation suite (Stodden, 2024a). First, we compared the performance on our test set with the latest German ATS models. For this, we compare these four models: erlesen-leo-7b (Klöser et al., 2024), mBART-DEplain-APA+web (Stodden et al., 2023), mt5-simple-german-corpus (SGC) (Stodden, 2024b), and Simba (Asghari et al., 2024). The full evaluation results are presented in Appendix D. Our model outperforms the other systems in all evaluation criteria except the compression strength. More interestingly, though, is that our model's target complexity outputs match the characteristics of the

Model	BLEU↑	SARI↑	BS_F1↑	FRE↑	Compression↓	Sent. splits↑	Copies↓
DEplain-web-sent (Stodden et al., 2023)							
References	-	-	-	75.56	0.94	1.87	0.0
mBART-DEplain-APA+web	**17.99**	34.07	**0.43**	68.41	0.85	1.16	0.14
mt5-SGC	3.0	37.02	0.29	74.45	**0.5**	0.95	0.01
German4All-level1 (ours)	5.58	37.85	0.32	**84.81**	0.96	**2.3**	**0.00**
German4All-level2 (ours)	7.67	**38.05**	0.36	77.27	1.01	1.79	**0.00**
Simple German Corpus (Toborek et al., 2023)							
References	-	-	-	64.54	1.26	2.15	0.0
mBART-DEplain-APA+web	**6.69**	28.68	**0.32**	44.23	1.62	1.29	0.17
mt5-SGC	4.67	**43.68**	**0.32**	57.85	**0.66**	1.02	0.03
German4All-level1 (ours)	3.2	41.09	0.27	**79.2**	1.18	**2.11**	**0.00**
German4All-level2 (ours)	4.36	39.41	0.29	65.53	1.37	1.62	0.01
GEOlino (Mallinson et al., 2020)							
References	-	-	-	68.18	0.95	1.32	0.36
mBART-DEplain-APA+web	**55.35**	**44.28**	**0.77**	67.23	0.97	1.09	0.28
mt5-SGC	12.6	29.17	0.49	75.26	**0.75**	0.96	0.04
German4All-level1 (ours)	12.46	29.18	0.44	**83.54**	1.13	**1.89**	**0.00**
German4All-level2 (ours)	19.86	34.05	0.54	75.63	1.2	1.5	0.01
TextComplexityDE (Naderi et al., 2019)							
References	-	-	-	51.64	0.95	2.08	0.0
mBART-DEplain-APA+web	**17.75**	37.37	**0.5**	45.43	0.74	1.31	0.06
mt5-SGC	1.52	33.51	0.27	65.12	**0.34**	0.96	**0.00**
German4All-level1 (ours)	4.28	35.43	0.29	**81.39**	0.77	**3.04**	**0.00**
German4All-level2 (ours)	9.33	**40.56**	0.42	67.11	0.81	2.18	**0.00**

Table 3: Performance comparison on four German simplification datasets. We follow the approach by Stodden (2024b). Thus, we evaluated the outputs in terms of the reference-based metrics BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), and BERTscore (Zhang* et al., 2020), as well as with the linguistic annotations FRE, compression rate, number of sentence splits, and number of exact copies. The best value per dataset is bolded.

respective references, i.e., to-level-1 wins for the level 1 references, but to-level-2 wins for the level 2. This shows that our model indeed distinguishes between different complexity levels.

Finally, we evaluated our model on publicly available sentence- or paragraph-level datasets, following the approach by Stodden (2024b). Table 3 shows the performance when prompting our model to output texts in complexity levels 1 or 2. We compare them with two models that performed best in terms of SARI and FRE. The full EASSE-DE reports can be found on git[2]. Our models achieve SOTA SARI scores and the highest number of sentence splits, showing a consistent performance on all datasets. Yet, they fall short in terms of meaning overlap metrics like BLEU and BERTscore. EASSE-DE outputs the samples with the worst relative score compared to the other models. We manually reviewed these samples and found that often, the input and reference data are noisy, and our model outputs are quite good, matching the expected styles of levels 1 and 2. An example for this is the following sample from the DEplain-web-sent corpus (Stodden et al., 2023):

Input: "Ein beliebter Badeort mit vielen Stränden und Wassersport ist Deutschlands größter See , der Bodensee , ganz im Süden ." (*EN: A popular seaside resort with many beaches and water sports is Germany's largest lake, Lake Constance, in the far south.*)

German4All-level1: "In Deutschland gibt es einen großen See . Der Bodensee ist sehr groß . Viele Menschen gehen dort zum Baden ." (*EN: There is a large lake in Germany . Lake Constance is very big . Many people go there to swim .*)

German4All-level2: "Der Bodensee ist ein großer See in Deutschland . Er liegt im Süden . Dort gibt es viele Strände und viele Möglichkeiten , Wassersport zu machen ." (*EN: Lake Constance is a large lake in Germany . It is located in the south . There are many beaches and many opportunities to do water sports .*)

Both models provide a high-level simplification of the input and preserve most of the content. The level 1 simplification is easier than level 2, with stronger content compression and shorter sentences. Nevertheless, both simplifications receive low automatic scores due to the poor reference. This prob-

[2] https://github.com/MiriUll/German4All/tree/main/easse-de_reports

lem with automatic meaning preservation metrics was also reported by Barayan et al. (2025, Section 5) and will probably remain unsolved until a multilingual paraphrase-aware metric is developed.

6 Conclusion

In this paper, we present German4All, the first aligned multi-level paraphrasing dataset in German. With more than 25k samples and a Flan-T5 model trained on this data, our work enables large-scale research on text simplification, paraphrasing, and readability assessment—going beyond what was previously possible for German.

Limitations

GPT4 has shown impressive performance and high agreement with human annotators in readability ranking tasks (Engelmann et al., 2024) or when generating simplifications (Klöser et al., 2024). While we tried to give an exhaustive evaluation of our dataset with 16 native speakers and an LLM-as-a-judge, the dataset was synthesized using an LLM. Thus, it could still contain errors or biases, and users of the dataset must be aware of these limitations. Moreover, the human annotators have a considerably high degree of education and can't be considered people from the easy or plain language target groups. This restriction was necessary so that the annotators could also evaluate the samples in the higher complexity levels 4 and 5. Nevertheless, it results in a lack of target group integration that is recommended for readability evaluations (Säuberli et al., 2024) and has been done for other works in simplification (Gao et al., 2025; Anschütz et al., 2024).

In terms of content, we focused on data from Wikipedia due to its permissive license. This results in samples that are very descriptive and have an explanatory nature. Yet, we not only took paragraphs from the abstracts, but from the full articles. Thus, our data also contains paragraphs that describe happenings and the life lines of people. These samples can be considered similar to news articles, and thus, our dataset can be used for paraphrases in different domains and use cases. The Wikipedia samples could have been part of GPT-4's training data and thus contaminated. However, we don't benchmark GPT-4, but use it to create our dataset. Thus, we don't see any problems with this exposure.

Finally, our simplifications show a very good structural understanding of simplified language and introduce sentence splits and rewritings where necessary. However, while very complex terms are often removed, the level of factual term explanations and guiding the user to interpret the content could be increased (Hewett et al., 2024). However, to ensure factual correctness in these definitions, we did not include this angle of elaborative simplification in our dataset.

Ethical considerations

Readability-controlled paraphrasing is an effort to tailor texts to the needs of specific readers. Thus, it tries to overcome the limitations of text simplification that assumes a homogeneous target group and only provides standardized simplifications. As such, these single simplifications may still be too complex for some readers, while others might find the level of simplification discriminating (Maaß, 2020). Therefore, our resources aim to help reduce discrimination while increasing true accessibility.

However, synthesizing a dataset with LLMs like GPT-4 might introduce biases into our dataset that we could not control for. As such, the LLM outputs reflect the biases and limitations of GPT-4's training data. Moreover, our input data relies on Wikipedia, a user-generated information platform with potentially erroneous content. Thus, users of our dataset and model should be aware of potentially wrong or outdated information and always evaluate the outputs against other sources.

Data availability statement

All experimental results that were discussed in this work are publicly available on our Git repository or on Huggingface. This includes the different versions of our dataset, results from the human, LLM judge, and EASSE-DE evaluation, and all model predictions.

Acknowledgements

Our paper uses closed-source models from OpenAI. To create and evaluate our corpus, we spent approximately 500$.
This paper is based on joint work in the context of Mai Pham's master's thesis (Thanh Mai Pham, 2024). In addition, we thank Robert Schauer, Maximilian Eckert, and Ricardo Ebene for their preliminary model fine-tuning results.
This research has been funded by the German Federal Ministry of Research, Technology and Space

(BMFTR) through grant 01IS23069 Software Campus 3.0 (Technical University of Munich) as part of the Software Campus project "LIANA".
The authors used an AI writing assistant in the form of ChatGPT to improve formulations and phrasing in the paper. Yet, no novel text was generated, and all corrections were thoroughly revised by the authors.

References

Mina Almasi and Ross Kristensen-McLachlan. 2025. Alignment drift in CEFR-prompted LLMs for interactive Spanish tutoring. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 70–88, Vienna, Austria. Association for Computational Linguistics.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich.

Miriam Anschütz and Georg Groh. 2022. TUM social computing at GermEval 2022: Towards the significance of text statistics and neural embeddings in text complexity prediction. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 21–26, Potsdam, Germany. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Miriam Anschütz, Tringa Sylaj, and Georg Groh. 2024. Images speak volumes: User-centric assessment of image generation for accessible communication. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 27–40, Miami, Florida, USA. Association for Computational Linguistics.

Hadi Asghari, Christopher Richter, Freya Hewett, Larissa Wunderlich, and Theresa Züger. 2024. Simba: Ai-powered text simplification.

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, Marseille, France. European Language Resources Association.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul-Gerhard Barbu. 2024. Entwicklung einer anwendung zum Übersetzten von texten in leichter/einfacher sprache mithilfe von large language models (llms). Master's thesis, Rosenheim Technical University of Applied Sciences. Advised and supervised by Prof. Dr. Gerd Beneken and Prof. Dr. Marcel Tilly.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. Learning to paraphrase sentences to different complexity levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Council for Cultural Co-operation. Education Committee. Modern Languages Division Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

DIN-Normenausschuss Ergonomie. 2024. Einfache Sprache - Anwendung für das Deutsche - Teil 1: Sprachspezifische Festlegungen (plain language - application for the german language - part 1: Language-specific provisions, DIN 8581-1).

DIN-Normenausschuss Ergonomie. 2025. Empfehlungen für Deutsche Leichte Sprache (guidance for German Easy Language, DIN SPEC 33429).

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Leon Fruth, Robin Jegan, and Andreas Henrich. 2024. An approach towards unsupervised text simplification on paragraph-level for German texts. In *Proceedings*

of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 77–89, Torino, Italia. ELRA and ICCL.

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. *Preprint*, arXiv:2507.01479.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, and Ava Spataru et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Antje Heine. 2017. Deutsch als Fremd- und Zweitsprache – eine besondere Form Leichter Sprache? Überlegungen aus der Perspektive des Faches DaF/-DaZ". *„Leichte Sprache "im Spiegel theoretischer und angewandter Forschung*, pages 401–414.

Freya Hewett, Hadi Asghari, and Manfred Stede. 2024. Elaborative simplification for German-language texts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 29–39, Kyoto, Japan. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lars Klöser, Mika Beele, Jan-Niklas Schagen, and Bodo Kraft. 2024. German text simplification: Finetuning large language models with semi-synthetic data. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 63–72, St. Julian's, Malta. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Joseph Liu, Yoonsoo Nam, Xinyue Cui, and Swabha Swayamdipta. 2025. Evaluation under imperfect benchmarks and ratings: A case study in text simplification. *Preprint*, arXiv:2504.09394.

Christiane Maaß. 2020. *Easy language–plain language–easy language plus: Balancing comprehensibility and acceptability*. Frank & Timme, Berlin.

Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2023. Easy-to-read language resources and tools for three european languages. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, page 693–699, New York, NY, USA. Association for Computing Machinery.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Sabine Manning. 2024. Ki-tools für einfache sprache: Leistungen von 10 tools auf einen blick. Published 22.05.2024, last accessed 30.04.2025.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language. *arXiv preprint*.

OECD. 2013. Oecd skills outlook 2013: First results from the survey of adult skills. OECD Publishing, http://dx.doi.org/10.1787/9789264204256-en.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Björn Plüster. 2023. Leolm: Igniting german-language llm research.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Thorben Schomacker, Michael Gille, Marina Tropmann-Frick, and Jörg von der Hülls. 2023. Data and approaches for German text simplification – towards an accessibility-enhanced communication. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 63–68, Ingolstadt, Germany. Association for Computational Lingustics.

Nicolas Spring, Annette Rios, 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, Held Online. INCOMA Ltd.

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. 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, Miami, Florida, USA. Association for Computational Linguistics.

Regina Stodden. 2024b. Reproduction & benchmarking 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, Torino, Italia. ELRA and ICCL.

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, Toronto, Canada. Association for Computational Linguistics.

Thanh Mai Pham. 2024. German4all: Paraphrasing german texts to different complexity levels with gpt-generated synthetic data. Master's thesis, Technical University of Munich. Advised and supervised by Miriam Anschütz and Georg Groh.

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, Toronto, Canada. Association for Computational Linguistics.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron C. Wallace, and Junyi Jessy Li. 2024. InfoLossQA: Characterizing and recovering information loss in text simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 4263–4294.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Human operation	LLM Judge criterion	Phi-4	Llama-3.3-70B-Instruct	Qwen2.5-72B-Instruct	Gemma-3-27B-it
corrected info	**content preservation**	**-0.23 (0.007)**	-0.05 (0.566)	**-0.24 (0.006)**	**-0.27 (0.001)**
removed info	**information addition**	**0.38 (0.000)**	0.13 (0.126)	**0.30 (0.001)**	**0.51 (0.000)**
hallucination in origin	**information addition**	**0.23 (0.008)**	0.10 (0.231)	0.20 (0.022)	**0.35 (0.000)**
added info	**information loss**	0.14 (0.101)	0.13 (0.130)	0.16 (0.059)	0.13 (0.153)
adjusted complexity	**complexity level**	-0.07 (0.401)	0.19 (0.031)	**0.32 (0.000)**	**0.34 (0.000)**

Table 4: Spearman correlation strength and p-values (in brackets) of human operations in the *German4All-annotated* dataset and the LLM judge criteria. Statistically significant correlations (p-values ≤ 0.01) are bolded.

A Complexity level specifications

We distinguish between these five complexity levels. The following descriptions were originally provided in German, but translated to English for the paper.

1. Leichte Sprache Plus (literal translation: Easy Language Plus)
   - *Target group*: People with reading difficulties, including people with learning disabilities and those who have only recently started to learn German.
   - *Characteristics*: Very short sentences, only short and frequently used words, direct speech, avoidance of abbreviations, metaphors, or irony.
   - *Examples areas*: simple instructions, accessible websites.
2. Simple German for beginners
   - *Target group*: Non-native speakers with basic knowledge of German.
   - *Characteristics*: Simple sentence structures, basic vocabulary, strong focus on important information, avoidance of culture-specific expressions.
   - *Example areas*: Language learning materials, introductory web texts.
3. Commonly used language
   - *Target group*: General public with different levels of education.
   - *Characteristics*: Clear, structured sentences, focus on comprehensibility, avoidance of technical terms.
   - *Example areas*: Wide-ranging news portals, blogs.
4. Elevated everyday language
   - *Target group*: Regular readers with a good understanding of the language.
   - *Characteristics*: More varied vocabulary, occasional technical terminology with explanations, complex sentence structures.
   - *Example areas*: Specialist blogs, quality newspapers.
5. Academic language
   - *Target group*: Academics and experts.
   - *Characteristics*: Complex sentence structures, specialized terminology, use of technical terms.
   - *Example areas*: Specialist journals, scientific publications.

B LLM judge ablation

We compared multiple LLMs to see if they properly capture our evaluation criteria. For this, we selected four recent multilingual open-source models. The annotations of human operations in the *German4All-annotated* dataset give an indication about the errors in the GPT-4 outputs. For example, if the human manually removed information from the output, that means that GPT-4 has added information that should not be there. Similarly, if the human has adjusted the text complexity, that means

the previous version of the text was improper for the target complexity level. We selected five human operations in the dataset and paired them with criteria in our evaluation setup. An ideal judge model should have high correlations for all these pairs. The results are presented in Table 4.

Our prompt features a sanity check on whether the LLM understood the prompt. As such, the answer to question 4 (type of addition) should be *'NaN'* if the answer to question 3 was *never*. All models succeeded with this prompt understanding test. Concerning the human annotations, no LLM judge can properly detect information that was missing from the GPT-4 outputs. However, Phi4 and Gemma3 can detect all other content-related errors, like hallucinations and incorrect information. For the complexity evaluation, only Qwen2.5 and Gemma3 have a statistically significant correlation with the human annotations. Therefore, Gemma3 is our preferred judge model and was used for testing the full dataset.

C Distillation training parameters

For model fine-tuning, we used transformers 4.52.4 and peft 0.15.2 with the following configuration:

```
lora_config = LoraConfig(
    r=32,
    lora_alpha=16,
    target_modules=["q", "v", "k", "o"],
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.SEQ_2_SEQ_LM,
)

tr_args = Seq2SeqTrainingArguments(
    output_dir=output_dir,
    eval_strategy="steps",
    save_strategy="steps",
    save_steps=20000,
    eval_steps=100000,
    learning_rate=3e-4,
    per_device_train_batch_size=64,
    per_device_eval_batch_size=64,
    auto_find_batch_size=True,
    weight_decay=0.01,
    num_train_epochs=2,
    predict_with_generate=True,
    logging_dir="./logs",
    logging_steps=1000,
    bf16=True,
    fp16=False,
    report_to="none",
    lr_scheduler_type="linear",
    warmup_steps=2750
)
```

D Benchmarking German ATS models on our dataset

We compare our model against the best-performing models according to Stodden (2024b) in terms of simplicity and SARI scores, as well as two more recent German simplification models. All models were fine-tuned on different datasets, and the base versions range from mBart to Llama3 (Grattafiori et al., 2024) and the German-specific Leo-LM (Plüster, 2023).

The highest readability score is obtained by our German4All model, prompted to output texts at level 1, while the erlesen model shows the lowest. Mt5-SGC has the lowest compression rate, mostly because it just outputs one sentence, no matter the input length. Our model learns the nuances between simplifications to levels 1 and 2 and achieves higher BLEU, SARI, and BERTscore, i.e., is closer to the reference, when creating outputs of the respective level.

E Further example

Here is an example from the main corpus and the respective LLM judge annotations:

Orig text: "Die vier zum Teil beschaädigten Nischenfiguren der Südfassade unterhalb des Giebels stellen Drama, Gesang, Musik und Tanz dar. Sie wurden von dem Stuttgarter Bildhauer Theodor Bausch geschaffen." (*EN: The four partially damaged niche figures on the south facade below the gable represent drama, singing, music, and dance. They were created by the Stuttgart sculptor Theodor Bausch.*)

Level 1: "An einem großen Haus gibt es Figuren aus Stein. Sie sind ein bisschen kaputt. Die Figuren zeigen vier Dinge: ein Theaterstück, Singen, Musik und Tanzen. Ein Mann namens Theodor Bausch hat sie gemacht." (*EN: There are stone figures on a large house. They are a little damaged. The figures depict four things: a play, singing, music, and dancing. A man named Theodor Bausch made them.*)
→ LLM judge annotations: content_preservation = approximately, information_loss = often, information_addition = seldom, complexity_level = a bit too complicated

Level 2: An der Südseite eines Gebaudes gibt es vier beschadigte Statuen. Diese Statuen zeigen vier Künste: Schauspiel, Gesang, Musik und Tanz. Ein Künstler aus Stuttgart, Theodor Bausch, hat sie erschaffen. (*There are four damaged statues on the south side of a building. These statues depict four arts: acting,*

Model	BLEU↑	SARI↑	BS_F1↑	FRE↑	Compression↓	Sent. splits↑	Copies↓
German4All - Target complexity level 1							
References	-	-	-	75.7	0.73	1.96	0.0
erlesen-leo-7b	4.55	38.15	0.22	49.25	1.08	1.1	0.01
mBART-DEplain-APA+web	2.74	35.88	0.26	52.1	0.49	0.72	0.0
mt5-SGC	0.2	38.65	0.2	70.5	**0.22**	0.51	0.0
Simba	4.85	41.29	0.28	65.09	0.71	1.48	0.0
German4All-level1 (ours)	**14.47**	**53.9**	**0.45**	**77.2**	0.69	**1.99**	0.0
German4All-level2 (ours)	10.85	47.43	0.41	67.18	0.8	1.49	0.0
German4All - Target complexity level 2							
References	-	-	-	64.02	0.82	1.45	0.0
erlesen-leo-7b	8.44	42.42	0.28	49.25	1.08	1.1	0.01
mBART-DEplain-APA+web	4.22	37.98	0.35	52.1	0.49	0.72	0.0
mt5-SGC	0.19	36.86	0.2	70.5	**0.22**	0.51	0.0
Simba	5.59	41.93	0.33	65.09	0.71	1.48	0.0
German4All-level1 (ours)	10.85	50.41	0.41	**77.2**	0.69	**1.99**	0.0
German4All-level2 (ours)	**15.66**	**51.7**	**0.48**	67.18	0.8	1.49	0.0

Table 5: Performance comparison of different German ATS models on our German4All-Corrected test set. Best values are bolded. Our models show the highest scores and best fit the characteristics of the references in the respective complexity level.

singing, music, and dance. They were created by an artist from Stuttgart named Theodor Bausch.)
→ LLM judge annotations: content_preservation = approximately, information_loss = sometimes, information_addition = seldom, complexity_level = a bit too easy

Level 3: An der Südwand eines Gebaudes befinden sich vier teilweise beschadigte Statuen. Sie symbolisieren Dramatik, Gesang, Musik und Tanz. Diese Kunstwerke wurden vom Stuttgarter Bildhauer Theodor Bausch geschaffen. (*EN: On the south wall of a building there are four partially damaged statues. They symbolize drama, singing, music, and dance. These works of art were created by the Stuttgart sculptor Theodor Bausch.*)
→ LLM judge annotations: content_preservation = correct, information_loss = seldom, information_addition = seldom, complexity_level = appropriate

Level 4: Die Südfassade des Gebaudes ziert eine Reihe von Nischenfiguren, die trotz teils sichtbarer Schaden die Künste des Dramas, Gesanges, der Musik und des Tanzes darstellen. Sie sind Werke des renommierten Stuttgarter Bildhauers Theodor Bausch. (*EN: The south facade of the building is adorned with a series of niche figures which, despite some visible damage, represent the arts of drama, singing, music, and dance. They are the work of the renowned Stuttgart sculptor Theodor Bausch.*)
→ LLM judge annotations: content_preservation = correct, information_loss = never, information_addition = seldom, complexity_level = appropriate

Level 5: Die Südfassade des architektonischen Werkes weist vier beschadigte Nischenfiguren auf, welche die Disziplinen Drama, Gesang, Musik und Tanz reprasentieren. Diese künstlerischen Darstellungen wurden von Theodor Bausch, einem Bildhauer aus Stuttgart, gefertigt und reflektieren kulturelle Ausdrucksformen der bildenden Künste. (*EN: The south facade of the architectural work features four damaged niche figures representing the disciplines of drama, singing, music, and dance. These artistic representations were created by Theodor Bausch, a sculptor from Stuttgart, and reflect cultural forms of expression in the fine arts.*)
→ LLM judge annotations: content_preservation = approximately, information_loss = sometimes, information_addition = often, complexity_level = a bit too complicated

The samples are all of high quality. However, levels 3 and 4 achieve the best content preservation, while level 1 removes information ("Südfassade unterhalb des Giebels" (*south facade below the gable*), "Stuttgarter") and level 5 adds information ("architektonischen Werkes" (*architectural work*), "reflektieren kulturelle Ausdrucksformen der bildenden Künste" (*reflect cultural forms of expression in the fine arts*)).

F System prompts

The following figures show the prompts that we used to generate the dataset and evaluate it using an LLM judge. All prompts were provided in German to avoid code switching, but were translated to

English for the paper.

F.1 Synthetic data generation

```
**Context**
There are five Complexity levels: <definitions from App. A>
```

```
**Example**
Text: <input text>
Paraphrases in json Format:
{
    "1": "Paraphrase at complexity level 1",
    "2": "Paraphrase at complexity level 2",
    "3": "Paraphrase at complexity level 3",
    "4": "Paraphrase at complexity level 4",
    "5": "Paraphrase at complexity level 5"
}
```

```
**Task**
Paraphrase the given text to five different complexity levels. When creating the
paraphrases, you should proceed step by step,considering the target group and the
specific characteristics and areas of application of each complexity level. Do
this task in the form of an inner monologue. Do not explain your thought process,
but present the final paraphrased texts directly.
```

```
Text: <input text>
```

```
**Response in json format**
{
    "1": "Level 1 text",
    "2": "Level 2 text",
    "3": "Level 3 text",
    "4": "Level 4 text",
    "5": "Level 5 text"
}
```

Figure 5: Prompt structure for generating paraphrases at five complexity levels. Colored boxes separate each section. First, we describe each complexity level. Then, we show a 1-shot example. Afterward, we provide the task description, the input text, and the response format.

```
**Beispiel**
Text: Die Ortschaft Danbury geht auf die Gründung einer vorchristlichen Wallburg
zurück. Funde lassen auf eine erste Siedlung in der Eisenzeit schließen. Nach
Römern und Angelsachsen, wurde das Gebiet um Danbury im 11. Jahrhundert von
dänischen Stämmen erobert. Nach der Eroberung Englands durch Wilhelm den Eroberer
1066 wurde das Land um Danbury von den Normannen besiedelt. Das älteste, heute
noch erhaltene Gebäude, ist die Kirche St. John the Baptist, die im 13.
Jahrhundert errichtet wurde.
Paraphrasierungen im json Format:
{
    "1": "Danbury ist ein Ort. In Danbury gab es früher eine große alte Burg.
    Viele verschiedene Leute haben dort gelebt. Die ersten Menschen kamen vor sehr
     langer Zeit. Später kamen Menschen aus Dänemark und dann aus Frankreich. In
    Danbury steht eine sehr alte Kirche. Sie ist ungefähr 800 Jahre alt.",
    "2": "Danbury ist ein Ort mit einer alten Burg, die noch vor der christlichen
    Zeit gebaut wurde. Zuerst lebten dort Menschen in der Eisenzeit. Dann kamen
    Römer, Angelsachsen und später Dänen. Nach einer großen Schlacht kamen
    Menschen aus Frankreich, die Normannen. Die älteste Kirche dort heißt St. John
     the Baptist und wurde im Mittelalter gebaut.",
    "3": "Danbury ist bekannt für seine historische Burg, die vor der christlichen
     Ära errichtet wurde. Die ersten Bewohner kamen während der Eisenzeit. über
    die Jahrhunderte hinweg wurde das Gebiet von Römern, Angelsachsen und später
    von Dänen bewohnt. Nach der normannischen Eroberung Englands im Jahr 1066
    wurde Danbury von den Normannen übernommen. Die Kirche St. John the Baptist,
    die älteste in Danbury, stammt aus dem 13. Jahrhundert.",
    "4": "Danbury besitzt eine lange Geschichte, die mit einer prähistorischen
    Festung beginnt. Archäologische Funde deuten darauf hin, dass die Region
    bereits in der Eisenzeit bewohnt war. Nach den Römern und Angelsachsen kamen
    im 11. Jahrhundert dänische Eroberer. Die normannische Eroberung Englands im
    Jahr 1066 brachte weitere Veränderungen mit sich, und Danbury fiel in die
    Hände der Normannen. Die Kirche St. John the Baptist, erbaut im 13.
    Jahrhundert, ist das älteste noch bestehende Gebäude.",
    "5": "Die historische Entwicklung von Danbury lässt sich bis zu einer prä
    christlichen Festungsanlage zurückverfolgen. Archäologische Befunde belegen
    eine frühe Besiedlung während der Eisenzeit. Die Abfolge der
    Herrschaftswechsel von Römern zu Angelsachsen und später zu dänischen Stämmen
    im 11. Jahrhundert illustriert die komplexe Sozialstruktur dieser Ära. Mit der
     normannischen Eroberung Englands im Jahre 1066 wurde Danbury Teil eines
    erweiterten Herrschaftsgebietes. Die Kirche St. John the Baptist aus dem 13.
    Jahrhundert dient als architektonisches Zeugnis dieser vielschichtigen
    Geschichte."
}
```

Figure 6: German 1-shot example provided in the original prompt.

F.2 LLM judge

```
You will receive an original text and a paraphrased version.

The paraphrases can be available in 5 different levels of difficulty. We define
these levels as follows: <definitions from App. A>
```

```
Your task is to evaluate the paraphrasing under the following aspects:

- content_preservation: How well does the content of the paraphrase match the
original text? Pay particular attention to whether the meaning has been changed or
 simplified so that nuances are lost or new interpretations are created.
- information_loss: Is information missing from the paraphrase that appears in the
 original?
- information_addition: Does the paraphrase contain additional information that
does not appear in the original? This includes explanatory paraphrases or examples
 that introduce new elements of meaning.
- type_of_addition: If you answered 'never' to the previous question, answer 'NaN'
 here. Otherwise, if additional information is included, indicate what type it is:
 e.g., explanations, embellishments, or correct or incorrect facts.
- complexity_level: How well does the paraphrase match the given level of
complexity? Pay attention to the linguistic features of the respective level, not
just the content.
```

```
**IMPORTANT:**
- If abstract terms (e.g. 'universalist') are replaced by simple paraphrases (e.g.
 'for all people'), this is considered an additional explanation.
- If the paraphrase creates a shift in content (e.g. from ideological concept to
simple statement), this may constitute *factually incorrect information*.
- The assessment must be made exclusively in the following JSON format and must
comply with this schema:
{json_schema}

Return **only** the JSON object -- without any further text.
```

Figure 7: LLM judge system prompt with the task description and additional guidance.

```
Evaluate this text:
Original text: <input text original text>
Paraphrased text: <input text paraphrased text>
Complexity level of the paraphrase: <input text complexity level>

Pay particular attention to whether the meaning is lost or changed by
simplifications.
```

```
**Response in json format**
{
    'content_preservation': 'incorrectly'|'approximately'|'correctly'
    'information_loss': 'never'|'seldom'|'sometimes'|'often',
    'information_addition': 'never'|'seldom'|'sometimes'|'often',
    'type_of_addition': list[
        'embellishment'|'explanations/definitions'|
        'factually_incorrect_information'|'factually_correct_information'|
        'other'|'NaN'
    ],
  'complexity_level': 'too_easy'|'a_bit_too_easy'|'appropriate'|
  'a_bit_too_complicated'|'too_complicated',
}
```

Figure 8: LLM Judge user prompt and JSON output format.