(Dis)improved?! How Simplified Language Affects Large Language Model Performance across Languages

Anonymous ACL submission

Abstract

Simplified language enhances the accessibility and human understanding of texts. However, whether it also benefits large language models (LLMs) remains underexplored . This paper extensively studies whether LLM performance improves on simplified data compared to its original counterpart. Our experiments span six datasets and eight automatic simplification systems across three languages. We show that English models, including GPT-4o-mini, exhibit a significant performance drop on simplified data. This introduces an intriguing paradox: simplified data is helpful for humans but not for LLMs. At the same time, the performance in non-English languages sometimes improves, depending on the task and quality of the simplifier. Our findings offer a comprehensive view of the potential and limitations of simplified language for LLM performance and uncover severe implications for people depending on simple language.

1 Introduction

Automatic Text Simplification (ATS) is the task of rewriting a text using simpler vocabulary while preserving its original meaning. The goal is to increase readability and make information accessible to a broader audience. The primary target group of simplified language is people with low literacy and mental disabilities, or language learners (Martin et al., 2022). However, previous work has shown that not only people from the target group but even the broad majority of people profit from simplified language (Javourey-Drevet et al., 2022; Murphy Odo, 2022). With this paper, we try to answer if the same holds true for *Large Language Models* (LLMs). Given that LLMs are approaching human-like capabilities (Dubey et al., 2024), it is reasonable to hypothesize that they might also perform better with simplified input.

To answer this question, we select six labeled datasets across three distinct languages (English, German, and Russian) and simplify their texts using eight different pre-trained simplification models and LLMs. Then, we benchmark and compare five large language models, including Llama3.1 (Dubey et al., 2024), Aya Expanse (Dang et al., 2024), and GPT-4o-mini, on both the original and simplified corpora. Our results show a significant change in performance with a strong performance drop for English. An example of such a deteriorated performance can be seen in Figure 1. This drop introduces a severe risk for people who rely on simplified language: If they input prompts or samples in simple language, LLMs may show a worse performance and make more mistakes than with standard English. Especially for tasks with high societal impact, like fake news classification or news summarization, this increases discrimination for already vulnerable target groups. Overall, our contributions can be summarized as follows:

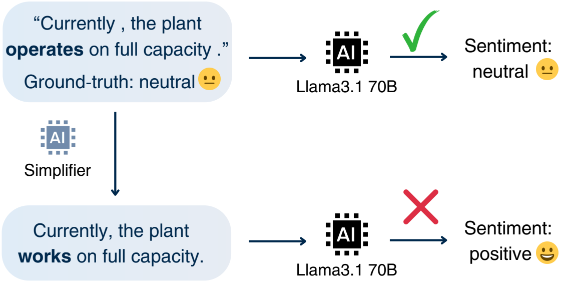

Figure 1: Text sample from the Sentiment Analysis for Financial News dataset (Malo et al., 2014). We test the classification performance of LLMs like Llama3.1 70B on original and automatically simplified data. The sentiment prediction on the original data sample is correct. However, if we use an automatic lexical simplifier that replaced the word "operates" with "works", Llama misclassifies the sample as positive.

- We present a large-scale multilingual benchmark of LLM performance on simplified data, including s.o.t.a. models like Llama3.1, Aya Expanse, and GPT-4o-mini. The simplifications are evaluated on a broad range of metrics, covering readability and meaning preservation.

- Our results indicate a significant performance decline on English simplified data, but with promising improvements in non-English languages.
- All code, simplified data, and model predictions are publicly available for further investigation and experimentation[1].

2 Related work

The impact of ATS on NLP tasks has been studied for many years and for different NLP tasks (Vickrey and Koller, 2008; Schmidek and Barbosa, 2014; Štajner and Popovic, 2016). However, many of the older studies could not use transformers or even large language models and were based on statistical simplification. Among the more recent studies, we identify two research directions: text simplification as data augmentation for pre-training or fine-tuning and text simplification as a pre-processing step to improve inference performance. To investigate the first direction, Van et al. (2021) simplify the training data for LSTM- and BERT-based classification models and evaluate the simplification quality with BLEU only. Results show that different setups of data augmentation with simplification can improve the classifiers. However, they also show that simplifying the data at inference time results in a weaker performance than the original data.

These results are in contrast to other studies that benchmarked simplification as inference preprocessing. Miyata and Tatsumi (2019) tested Google Translator for Japanese to English translations with sentence splitting and further rule-based simplifications. A human evaluation showed that the simplifications yielded strong improvements in the translation outputs. Similarly, Mehta et al. (2020) created an artificial simplification system through back translation and used this system to simplify the machine translation inputs of a low-resource-language translation system. They show improved translation quality across multiple languages. However, the performance changes of the target systems depend on the quality of the ATS systems. As such, Agrawal and Carpuat (2024) investigated how well ATS systems preserve the meaning of the original texts. While human simplifications could improve the performance of a pre-trained question-answering model, automatic simplifications worsened the performance. Our work tries to shed light on the contradicting findings of previous work. For this, we extend the existing research by covering more tasks, languages, and simplifiers. We paint a broader picture of the helpfulness of simplification as pre-processing, especially in times of flexible and powerful LLMs.

A different research direction was chosen by Anschütz et al. (2024), who used human-supervised simplification corpora to investigate how models change their behavior when working on the original and simplified data. They are the first ones to include LLMs in their investigations and show that models exhibit an incoherent behavior between original and simplified data. However, they only benchmarked GPT3.5-turbo as LLM, and their datasets do not contain ground-truth labels. While they assumed that the human-supervised datasets contain correct simplifications, they cannot measure the actual performance of the classification system without ground-truth labels. We try to overcome this weakness by using labeled datasets and benchmarking the performance of multiple LLMs on these datasets. In addition, we extend the investigation to the task of summarization and not only cover classification tasks.

3 Methodology

Our objective is to compare if the performance of different LLMs changes when the input samples are simplified. For this, we take labeled datasets and simplify the input texts with pre-trained simplification models. Then, we use pre-trained classification models or LLMs to predict the labels on the original and on the simplified inputs. Finally, we evaluate the predictions against the ground-truth labels and examine whether text simplification as pre-processing can improve the models' performance. An overview of our approach is shown in Figure 2. Our investigations range across three distinct languages with six different datasets, eight simplifiers, and six models under test, including LLMs like GPT4o and Llama3.1. All combinations of settings were evaluated independently, and the models did not know if the input text was the original or the simplified version to avoid bias. The different settings will be discussed further in the following subsections.

[1] URL removed for review

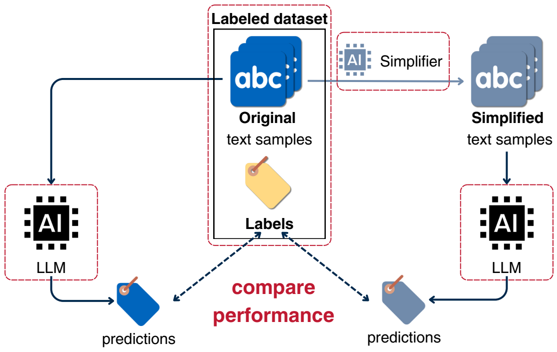

Figure 2: Structure of our investigations. We compare the performance of the same model between the original inputs and their simplified versions. Red boxes indicate that these parts are investigated under different settings.

3.1 Datasets and tasks

Our research covers the tasks of classification and summarization. In classification, the correctness of the label is unambiguous and independent from the chosen metric. In contrast to this, the evaluation of text generation is non-trivial since nuances of text and many language characteristics need to be covered. In addition, ATS systems sometimes struggle to preserve the exact meaning (Säuberli et al., 2024; Agrawal and Carpuat, 2024). Classification tasks like reading comprehension and natural language inference focus on specific text details that can get lost during simplification, even though the simplification is of high overall quality. To avoid depending on small details, we focus our experiments on more content-related tasks like topic and sentiment prediction. We assume that even if the simplifiers remove minor aspects, the overall content should not change significantly, and thus, the ground-truth labels are still correct for the simplified samples.

The selected datasets are shown in Table 1. We experiment with data in English, German, and Russian. All datasets are from the news domain, a general-purpose domain often targeted by ATS literature (Ryan et al., 2023). For each of the datasets, we only worked with the test splits. To reduce the financial efforts of the OpenAI API, we created fixed subsets of the AG News and the sentiment dataset and only used these subsets when prompting this API. In the following, results that are based on these subsets are indicated with †. Each language contains a multi-task dataset that provides data for topic classification and summarization at the same time to enable a multi-task evaluation. The number of classes and granularity of the classes differ among the languages and tasks. The AG News dataset has four very general classes, while the TL;DR dataset focuses more on technical news and its subcategories. For the sentiment task, we purposefully selected a dataset with only three classes (positive, negative, and neutral) to avoid ambiguity due to too fine-grained classes. The summarization task is headline generation, where the models create a headline for the respective news snippet. This task has a strongly abstractive nature and is well-suited to evaluate how well the models can retrieve the most important information from the texts (Scialom et al., 2020).

3.2 Simplifiers

We used eight different pre-trained simplification models for our experiments: two multilingual models for all languages and six language-specific models (four for English, one for German, and one for Russian). Our model selection was limited by the availability and reproducibility of existing approaches. Especially unmaintained or weakly-documented Github repositories make reusing pre-trained models challenging (Stodden, 2024; Kew et al., 2023). Nevertheless, the models that we could run give a good variety of approaches, ranging from lexical to paragraph-level simplification, and are trained for general-purpose or specialized domains. For all models, we used the default configurations provided in their repositories or model cards, and we did not add any further pre-processing. We used these simplification models:

MILES (multiling.) is a lexical simplification pipeline. It uses frequency-based complex word identification and replaces the complex words with a lexical simplifier similar to LSBert (Qiang et al., 2020). It is available in 22 languages, including our investigated languages.

GPT4o mini (multiling.) is one of the state-of-the-art LLMs by OpenAI and offers support for all three languages. We prompted it in a zero-shot manner to simplify the text samples. The simplification prompts per language are presented in Appendix A.

MUSS (EN) stands for “Multilingual Unsupervised Sentence Simplification” and is one of the most popular pre-trained sentence simplification models (Martin et al., 2022). We used the pre-trained `muss_en_mined` checkpoint that utilizes the BART architecture (Lewis et al., 2020). Even

Language	Dataset	Dataset name	Prediction Task	#samples (subset size)	#classes
EN	**AG News**	AG News (Zhang et al., 2015)	topic	7600 (760)	4
	Sentiment	Sentiment Analysis for Financial News (Malo et al., 2014)	sentiment	4846 (970)	3
	TL;DR	tldr_news	topic, summarization	794	5
DE	**Gnad10**	10k German News Articles Datasets (Schabus et al., 2017)	topic	1028	9
	ML SUM	Multilingual summarization (DE) (Scialom et al., 2020)	topic, summarization	579	12
RU	**ML SUM**	Multilingual summarization (RU) (Scialom et al., 2020)	topic, summarization	203	9

Table 1: Overview of all datasets and their classification tasks evaluated in this study.

though MUSS is multilingual, it does not support all the languages we investigate. Due to the long runtime of MUSS, we create simplifications only on the fixed subsets of the data.

Cochrane and Medeasi (EN) are both based on the HuggingFace space simplification-model-app. Both are built upon a BART model fine-tuned for simplification in the medical domain. The Medeasi checkpoint uses the sentence-level MED-EASi dataset (Basu et al., 2023), while Cochrane is fine-tuned on the paragraph-level Cochrane dataset (Devaraj et al., 2021).

SimplifyText (EN) uses the Keep it Simple (KiS) approach by Laban et al. (2021) and is a GPT2-based simplification model.

DEplain (DE) is a German text simplification model based on mT5 (Stodden, 2024) and fine-tuned on the DEplain-APA corpus (Stodden et al., 2023).

Russian simplification (RU) is a Russian sentence simplification model. It is based on ruT5 and was fine-tuned on the RuSimpleSentEval (Sakhovskiy et al., 2021) and the RuAdapt (Dmitrieva and Tiedemann, 2021) datasets.

3.3 Classifiers and LLMs

We evaluate the behavior of different classification and summarization systems when exposed to original and simplified versions of text. Our models under test span from DeBERTa-based classification systems to the latest open- and closed-source large language models. Table 2 gives an overview of the models and settings that we investigated.

For each English classification dataset, we fine-tuned two DeBERTaV3-base classifiers (He et al., 2023). The first classifier was trained on the original data, while the other classifier was fine-tuned on the data simplified with the SimplifyText model. We selected this model for simplification because it received the best scores among the open-source models in our unsupervised simplification evaluation (see subsection 3.4). Every training was conducted for one epoch with a learning rate of $2 \cdot 10^{-5}$. We trained the models on the datasets' training splits, so the test splits used for our investigation were still unseen for the models. With this training setup, we can test how much the models adapt to the specific style of simplification and if text simplification as pre-processing or data augmentation during training is beneficial for performance.

The second part of our study investigated the performance of large language models. For this, we selected four LLMs, three open-source models from Meta's Llama3.1 family (Dubey et al., 2024) and Aya Expanse 8B from Cohere for AI (Dang et al., 2024), and the closed-source GPT4o-mini from OpenAI. Llama3.1 is a multilingual LLM with a context length of 128k tokens. For our experiments, we use the instruction-tuned versions with 8B and 70B parameters to account for performance differences due to model size. Llama3.1 70B is loaded with bitsandbytes' 8-bit quantization. Unfortunately, Llama is not available in Russian. In contrast, Aya Expanse 8B exhibits powerful multilingual capacities and supports 23 languages, including the three in our study. For GPT, we limited our experiments to fixed subsets to reduce the financial efforts.

For the predictions themselves, we used the same zero-shot prompt for all four models. The prompts per dataset are presented in Appendix B. A native German or Russian speaker created each of the non-English prompts. Even if we told the models to only predict the topic and not provide any reasoning, some of the outputs still contained more content than the topic. We tried to account for the most common phrases among them during post-processing. Therefore, we lower-cased all model outputs and removed phrases like “The topic of this snippet is”. In addition, some labels were a combination of multiple terms, e.g., `sci/tech` in AG News. If only one part, e.g., only `sci`, was predicted, we considered this prediction correct and replaced it with the proper topic name.

Model	Setting	Language(s)
DeBERTaV3	FT Orig	EN
DeBERTaV3	FT Simple	EN
Llama3.1 8B Instruct	Zero-shot	EN, DE
Llama3.1 70B Instruct	Zero-shot	EN, DE
Aya Expanse 8B	Zero-shot	EN, DE, RU
GPT-4o-mini	Zero-shot	EN, DE, RU

Table 2: Overview of all models under test. Traditional models are fine-tuned on either the original training data or a simplified version of it. The LLMs are prompted in a zero-shot manner.

3.4 Unsupervised simplification evaluation

Previous work has investigated the impact of human-supervised simplifications (Anschütz et al., 2024), but for our datasets, human supervision is not feasible. In contrast, we investigate the impact of automatic text simplification, and thus, we need to evaluate the quality of the automatic simplifications. Our datasets are not targeted to simplification, and hence, no reference simplification exists. Therefore, we based our evaluation on unsupervised metrics that evaluate the simplification against the source instead of comparing it against a reference. While a human evaluation would be the best solution, this is infeasible for our large-scale study setup with multiple languages, datasets, and simplifiers. To still provide an insightful evaluation of the simplifications, we not only evaluate the overall simplification quality but also the readability of the texts and the meaning preservation independently. To measure the readability of the texts and the simplicity-gain through simplification, we used the Flesh-Reading-Ease (FRE) (Flesch, 1948). It is a statistical measure based on the number of words per sentence and the average word length. It can be adapted for many languages, including German and Russian. The score ranges from 0 to 100, with a higher score indicating a higher readability. We used the Python implementation of the textstat package and the German adaption by Amstad (1978).

The second aspect of our evaluation is the overall simplification quality. For this, we use two different scores, which are LENS_SALSA (Heineman et al., 2023) and REFeREE (Huang and Kochmar, 2024). Both metrics are learned metrics that were fine-tuned to mimic human annotation scores. LENS_SALSA is working on the word- and sentence-level and predicts and scores edit annotations that are performed during simplification. In contrast to this, REFeREE employs a multi-step fine-tuning process that aligns the metric scores with traditional metrics like BLEU (Papineni et al., 2002) and performs a multi-aspect evaluation of the fluency and simplicity of the generated text. While LENS_SALSA ranges from 0 to 100, REFeREE only ranges from -1 to 1. Therefore, we rescale the REFeREE values to make them comparable with the other metrics.

Finally, the third evaluation criterion is testing if the simplification preserves the original text’s meaning. This is especially important for content classification tasks, as in our study. Again, we select two metrics to evaluate the factuality of the simplifications. First, we use FactCC (Kryscinski et al., 2020), which has shown the best human correlation on factuality evaluations like the FRANK dataset (Pagnoni et al., 2021). It was originally designed for the evaluation of abstractive summarization, but since some of our simplification systems perform complex operations close to summarization, we consider this metric suitable. FactCC employs a binary classification to predict whether the summary is factually consistent with its source. For our evaluation, we calculate the percentage of samples that are deemed correct to end up with a value between 0 and 100 again. The last metric is MeaningBERT (Beauchemin et al., 2023), which is specifically targeted toward meaning preservation in text simplification.

We provide a detailed evaluation and correlation analysis only for English, as FRE is the only unsupervised metric that we could find for German and Russian simplification.

4 Results and Discussion

4.1 Simplification evaluation

We evaluate the simplifications in English based on three criteria: the readability of the texts, the over-

Metric	Original	MILES	Cochrane	Medeasi	SimplifyText	MUSS	GPT4o mini
				AG News			
FRE	48.78	54.13	**70.22**	58.92	65.93	53.64 †	59.11 †
REFeREE	-	36.08	72.48	67.19	71.0	65.35 †	**87.84** †
LENS_SALSA	-	53.0	66.56	62.41	64.66	60.74 †	**70.65** †
FactCC	-	**91.63**	52.37	85.04	60.39	84.87 †	85.53 †
Meaning_BERT	-	**91.56**	67.41	85.62	83.29	90.06 †	82.72 †
				Sentiment			
FRE	55.43	61.76	**73.34**	65.73	65.52	58.97 †	61.76 †
REFeREE	-	51.6	56.74	55.49	67.59	65.61 †	**75.46** †
LENS_SALSA	-	60.34	65.88	56.42	**69.85**	64.29 †	69.34 †
FactCC	-	96.22	54.5	91.48	73.85	95.26 †	**96.29** †
Meaning_BERT	-	84.84	50.19	**85.12**	76.74	83.27 †	78.68 †
				TL;DR			
FRE	57.27	63.85	**76.2**	67.74	62.08	60.73	62.32
REFeREE	-	39.88	75.25	76.0	79.93	79.48	**84.64**
LENS_SALSA	-	60.54	72.05	72.9	73.95	72.84	**75.74**
FactCC	-	**90.93**	49.75	87.03	66.37	86.23	88.92
Meaning_BERT	-	**89.11**	67.89	70.18	84.22	88.76	87.77

Table 3: Unsupervised simplification evaluation of the English simplifiers. For all metrics, higher scores indicate better simplification quality. The best scores per metric are bolded. †evaluated only on subset

all simplification quality, and the faithfulness of the simplifications. For this, we automatically score the simplifications with five different metrics (see subsection 3.4 for details). Table 3 shows the metrics scores for the English simplifications. In terms of readability, the Cochrane simplifier achieves the highest scores, indicating the biggest simplicity gain. Interestingly, the FRE scores of GPT4o-mini are rather low compared to the other simplifiers, indicating that it performs rather conservative simplification. Nevertheless, it achieves the best overall simplification quality across all datasets. This is probably due to its great fluency and overall capacities. In terms of faithfulness, MILES has the best scores across almost all datasets. This is expected since it is a lexical simplification system that does not rewrite the sentences but only replaces some complex words within. Overall, all simplification systems show a good performance and can be used for further experiments.

4.2 Model performances

To investigate if the model performances change when we simplify the input texts, we compare the accuracies of all classification tasks and the rougeL scores (Lin, 2004) for the summarization tasks as implemented in Huggingface evaluate. For each dataset, we report the results of the two fine-tuned DeBERTa classifiers and the four LLMs in a zero-shot setting. In addition, we tested whether the changes in accuracy were statistically significant. For this, we performed a related t-test with the hypothesis that the average of the two distributions was the same. If the p-value is smaller than 0.05, we reject this hypothesis and can conclude that the accuracy change is significant. The results for the English classification task are presented in Table 4. Overall, the fine-tuned classifiers (DeBERTa Orig and DeBERTa Simple) show the best accuracies, with GPT-4o-mini coming the closest. However, nearly all models show a decreased classification performance if the inputs are simplified. No performance improvement is statistically significant. However, the majority of the simplifications introduce a severe performance drop of up to 20 percentage points. The sentiment dataset is the dataset with the most significant performance changes, even though it has the fewest and most distinct classes. The performance decreases are especially remarkable for the DeBERTa classifier, which was fine-tuned on simplified data. This model exhibits a drop in performance even when the same simplifier is used for training and testing. A similar problem can be observed with GPT4o-mini, which exhibits a performance drop even when it is working on its own simplification outputs. However, statistically significant performance changes of the GPT4o-mini simplifications are scarce.

Our results show that all classifiers, even powerful LLMs like GPT-4o-mini, exhibit a performance decrease when working with simplified inputs. An

Model	Original	Original (subset)	MILES	Cochrane	Medeasi	Simplify Text	MUSS	GPT4o mini
AG News - Classification (*accuracy*)								
DeBERTa Orig	94.5	94.34 †	-6.58*	-1.07*	-2.79*	-3.71*	-1.58 †	-3.16*†
DeBERTa Simple	90.26	90.26 †	-3.0* †	-0.61*	-0.83*	-1.7*	-1.05 †	-1.32†
AyaExpanse8B	83.13	80.53	-0.14	-2.91*	-1.43*	-1.56*	+0.39†	-0.66 †
Llama3.1 8B	80.12	78.68 †	-1.3*	-1.96*	-1.48*	-1.58* †	+0.27 †	-5.26*†
Llama3.1 70B	79.97	80.26 †	-0.55*	-0.21	+0.08	-0.36	-0.79 †	+1.45†
GPT4o-mini	-	84.08 †	-0.66 †	+1.18 †	-0.79 †	± 0.0 †	± 0.0 †	-0.53†
Sentiment - Classification (*accuracy*)								
DeBERTa Orig	88.16	86.08 †	-6.0*	-13.91*	-1.98*	-5.65*	-0.82 †	+0.41†
DeBERTa Simple	87.49	87.53 †	-6.46*	-12.57*	-1.73*	-3.8*	-1.13 †	-1.24†
AyaExpanse8B	68.4	68.76 †	-5.52*	-17.33*	+0.02	-6.47*	-2.37†	-4.12†
Llama3.1 8B	68.17	68.56 †	-8.95*	-20.57*	-1.1	-14.39*	-7.01*†	-6.5*†
Llama3.1 70B	78.23	78.76 †	-3.96*	-10.1*	-1.98*	-5.97*	-4.74*†	-1.96†
GPT4o-mini	80.84	80.72 †	-4.09*	-14.76*	-1.01*	-9.8*	-3.19†	-0.72†
TL;DR - Classification (*accuracy*)								
DeBERTa Orig	76.32	-	-4.91*	-1.39	-15.37*	-0.25	-2.27*	-1.01
DeBERTa Simple	74.56	-	-3.53*	-0.13	-9.07*	+0.25	-0.38	+0.13
AyaExpanse8B	58.19	-	+0.63	-0.76	-0.13	+0.75	+1.51	+0.63
Llama3.1 8B	44.84	-	-3.4*	-1.26	-3.15	+0.75	± 0.0	-3.91*
Llama3.1 70B	56.55	-	-5.79*	-4.91*	-6.68*	-2.27	-1.01	-1.13
GPT4o-mini	65.74	-	± 0.0	± 0.0	-2.39	-2.01	-0.75	-0.75
TL;DR - Summarization (*rougeL*)								
AyaExpanse8B	23.09	-	-2.04*	-5.95*	-4.59*	-2.17*	-0.88*	-0.79*
Llama3.1 8B	23.89	-	-3.17*	-6.4*	-6.08*	-2.34*	-1.37*	-0.98*
Llama3.1 70B	27.04	-	-2.81*	-7.43*	-7.04*	-2.9*	-1.62*	-0.76
GPT4o-mini	25.86	-	-2.67*	-7.72*	-6.3*	-1.99*	-2.01*	-0.02

Table 4: Changes in performance across all English datasets. For most of the models and simplifiers, the scores decrease (red boxes). Only a few combinations show improved performance (blue boxes). * statistically significant change ($p < 0.05$), significant changes have a darker color, †evaluated and compared only on the fixed subset

obvious explanation for this behavior would be that the simplification systems alter the meaning of the input samples. However, while the MILES simplifier has the highest meaning preservation according to the automated metrics (compare Table 3), it is among the simplifiers with the strongest performance drops for the classifiers. Therefore, we reject the faithfulness alone as a simple explanation of this behavior. Moreover, MILES is only a lexical simplification system that performs minimal changes, indicating that the choice of words in simplified language is more relevant to the classifiers than the sheer amount of edit operations. This aligns with previous research by Anschütz et al. (2024), who find that the Levenshtein distance between original and simplified samples only has a weak correlation with label changes in LLMs.

Table 5 and Table 6 show the results for German and Russian respectively. First of all, we can see that the FRE scores increase for all ATS systems, indicating that the simplifiers successfully improved the readability of the samples. Again, the GPT4o-mini simplifications achieve a comparatively small readability improvement. For Russian, we observe hardly any statistically significant changes, except for some strong improvements of Aya Expanse on the classification task. In general, both Russian models show an extremely weak summarization performance in terms of rougeL score, even for the original data. Therefore, the changes on simplified data have only minor importance as the models don't seem to fulfill the task at all. For German, we observe many improvements, especially for the Gnad10 classification task. In addition, simplifications by GPT4o show the most significant improvements and only one significant performance drop. This is even the case in the summarization task. Our results allow for two interpretations: Most of the models are primarily trained on English data, and they seem to overfit more to the standard language style in their pre-training. Therefore, their performance on English simplified language drops significantly. However, for languages with weaker LLM support, we assume less overfitting. Thus, these models can benefit from simplifications, especially if they are of high, human-like quality, as

Model	Orig.	DEplain	MILES	GPT4o mini
Gnad10 - Classification (*accuracy*)				
FRE	46.41	61.34	59.96	52.55
AyaExpanse8B	26.75	+7.1*	+2.34*	+4.28*
Llama3.1 8B	50.78	-5.64*	-3.7*	+0.19
Llama3.1 70B	33.85	+7.4*	-1.85	+7.88*
GPT4o-mini	58.95	-4.77*	+3.21*	+1.17
ML SUM DE - Classification (*accuracy*)				
FRE	48.84	61.06	62.32	53.25
AyaExpanse8B	49.74	+3.46	-1.73	+3.11
Llama3.1 8B	62.0	-1.9	-0.51	+2.42
Llama3.1 70B	61.14	± 0.0	-6.74*	+5.18*
GPT4o-mini	77.72	-7.77*	-2.07*	-1.55
ML SUM DE - Summarization (*rougeL*)				
AyaExpanse8B	17.46	-10.97*	-3.05*	-1.7*
Llama3.1 8B	14.78	-9.19*	-1.99*	-0.71
Llama3.1 70B	15.63	-9.08*	-1.43*	+0.65
GPT4o-mini	16.1	-9.98*	-1.4*	+0.24

Table 5: Accuracy changes on German data, * statistically significant change ($p < 0.05$)

Model	Orig.	Russian simpl.	MILES	GPT4o mini
ML SUM RU - Classification (*accuracy*)				
FRE	48.33	51.66	70.74	49.01
AyaExpanse8B	32.02	+4.93	+8.37*	+14.29*
GPT4o-mini	67.98	+1.97	-1.97	-0.49
ML SUM RU - Summarization (*rougeL*)				
AyaExpanse8B	2.79	+0.16	-0.82	-0.82
GPT4o-mini	0.99	-0.49	± 0.0	± 0.0

Table 6: Accuracy changes on Russian data, * statistically significant change ($p < 0.05$)

with GPT4o-mini.

5 Limitations

We provide an extensive evaluation of the employed simplification models, evaluating them for their simplicity gain, simplification quality, and meaning preservation. However, we fully rely on automatic metric scores and don't extend the evaluation with a human review. Alva-Manchego et al. (2021) have shown that traditional metrics don't cover all aspects of simplification. Therefore, a deeper analysis of the models' strengths and shortcomings should be done in the future.

In addition to this, our investigation only covers a limited set of NLP tasks. We selected the sentiment and classification tasks to avoid biases due to automatic evaluation metrics and insufficient meaning preservation of the simplification models. In addition, we tested the performance on summarization as a generation task. Nevertheless, it would be valuable to add further NLP tasks to draw a broader picture of LLM behavior. Moreover, since the results indicate that simplifications can improve the performance of non-English languages, this research should be extended to further languages.

Finally, we used the same prompts for all models and tested them in a zero-shot setting. This could mean that the models could not unfold their full potential and that the performances could be improved further. However, we don't evaluate the models on an absolute scale; rather, we compare the performance of simplified and original texts. All experiments are conducted under the same setting, and thus, the limitations of the zero-shot setting should not affect our overall results.

6 Conclusion

Experiments across six datasets, eight ATS systems, and three languages show that English LLMs exhibit a severe performance drop when switching from original to simplified language. However, simplified texts can enhance performance at inference time for non-English languages. We thus encourage content creators to prioritize using simple language online as a way to improve LLMs' downstream performance and comprehension and to open their models to a broader audience.

7 Ethical considerations

Our work uncovers novel insights into how LLMs perform on simplified language. We don't create any new datasets or models, and thus, there is no harm coming from our investigations. However, we find some alarming behavior in most of the LLMs as our results show that they decrease their performance when using simplified language in English. This can have severe implications for people with low literacy or mental disabilities when using platforms like ChatGPT: When a user asks the chatbot for a summarization of a news snippet in plain language, the models are more likely to make mistakes in these interactions. These people are already a vulnerable target group that struggles to verify information on the internet due to information barriers of overly complicated texts. When easy-to-use and trust-evoking platforms like chatbots show a worse performance when interacting with those people, this implies severe discrimination of users of simplified language that we uncovered with this work.

References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

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.

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

Miriam Anschütz, Edoardo Mosca, and Georg Groh. 2024. Simpler becomes harder: Do LLMs exhibit a coherent behavior on simplified corpora? In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 185–195, Torino, Italia. ELRA and ICCL.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. *Preprint*, arXiv:2302.09155.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

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.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.

Yichen Huang and Ekaterina Kochmar. 2024. REFeREE: A REference-FREE model-based metric for text simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13740–13753, Torino, Italia. ELRA and ICCL.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C. Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of french. *Applied Psycholinguistics*, 43(2):485–512.

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, Singapore. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Research article. *Journal of the Association for Information Science and Technology*, 65(4):782 – 796.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-then-translate: Automatic preprocessing for black-box translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8488–8495.

Rei Miyata and Midori Tatsumi. 2019. Evaluating the suitability of human-oriented text simplification for machine translation. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 147–155. Waseda University.

Dennis Murphy Odo. 2022. The Effect of Automatic Text Simplification on L2 Readers' Text Comprehension. *Applied Linguistics*, 44(6):1030–1046.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

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.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lsbert: A simple framework for lexical simplification. *Preprint*, arXiv:2006.14939.

Michael 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, Toronto, Canada. Association for Computational Linguistics.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplesenteval-2021 shared task: evaluating sentence simplification for russian. In *Proceedings of the International Conference "Dialogue*, pages 607–617.

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 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Jordan Schmidek and Denilson Barbosa. 2014. Improving open relation extraction via sentence restructuring. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3720–3723, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. 2023. Why larger language models do in-context learning differently? In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Regina Stodden. 2024. 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, 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.

Hoang Van, Zheng Tang, and Mihai Surdeanu. 2021. How may I help you? using neural text simplification to improve downstream NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4074–4080, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352, Columbus, Ohio. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Simplifications to augment context

Previous work by Van et al. (2021) experimented with simplification as data augmentation. One of their experiments was to concatenate the original and simplified texts together. Our benchmarking of LLMs is done in a zero-shot setting, so data augmentation at training time is out-of-scope for our work. However, we tested how the models perform when they see the concatenated versions during inference time. For this, we created two additional versions of the data samples. First we concatenated the simplified text to the original one with just a whitespace in between. This version is called *orig+simp*. To ablate whether the accuracy changes are based on the input length or the additional context, we also created a version where the original samples were concatenated to themselves (identified as *orig+orig*). Due to the shorter context window of the DeBERTa models and their fine-tuning to a specific input style and length, we only ran these ablations on the LLMs. In addition, we only tested these settings for the classification tasks. Figure 3 shows the accuracy curves of the original, simple, and the two combined versions. For the larger models like GPT4o-mini and LLama 70B (bottom two), the augmentations seem to make no difference. Moreover, for Aya Expanse, the concatenations on the TL;DR dataset seem to worsen the performance even further. In contrast to this, for the smaller Llama 8B model, the *orig+simple* versions can improve the performance by over 10 percentage points, matching the performance of the larger Llama model. In contrast, we only see minor improvements or even decreased performances of the *orig+orig* concatenations. This implies that the simplifications give additional context or explanations to the original content that can improve the zero-shot performance of some of the smaller language models. This aligns with previous findings that different LLM sizes perform in-context-learning differently and that smaller models orient themselves more on the task description, while larger models rely on the knowledge they obtained during pre-training (Shi et al., 2023). For our experiments, this means that adding the simplifications to the original samples has a higher impact on the model performance of the smaller models.

A LLM simplification prompts

We used GPT4o-mini to create high-quality simplifications. We used the following prompt where `sample` is replaced by the text to be predicted. For German and Russian, the prompt is translated, respectively.

Simplify (EN): {“role”: {“system”, “content”: “You are a helpful assistant. You will be provided with sentences from news articles. Your task is to simplify the texts to enhance readability. You must not alter the meaning and don’t provide reasoning.” },
{“role”: “user”, “content”: “{`sample`} - Simplification: ”}

Simplify DE: {“role”: {“system”, “content”: “Du bist ein hilfreicher Assistent. Du bekommst Sätze aus Nachrichtenartikeln. Deine Aufgabe ist es, die Texte zu vereinfachen, um die Verständlichkeit zu erhöhen. Du darfst den Inhalt nicht verändern und brauchst keine Begründungen angeben.” },
{“role”: “user”, “content”: “{`sample`} - Vereinfachung: ”}

Simplify RU: {“role”: {“system”, “content”: “Ты - полезный помощник. Тебе будут предоставлены предложения из новостных статей. Твоя задача - упростить текст, чтобы повысить его читабельность. Ты не должен изменять смысл и приводить аргументы.” },
{“role”: “user”, “content”: “{`sample`} - Упрощение: ”}

B LLM Prediction prompts

We used the same system prompts for all four large language models and prompted them in a zero-shot manner. The prompts differ per dataset and

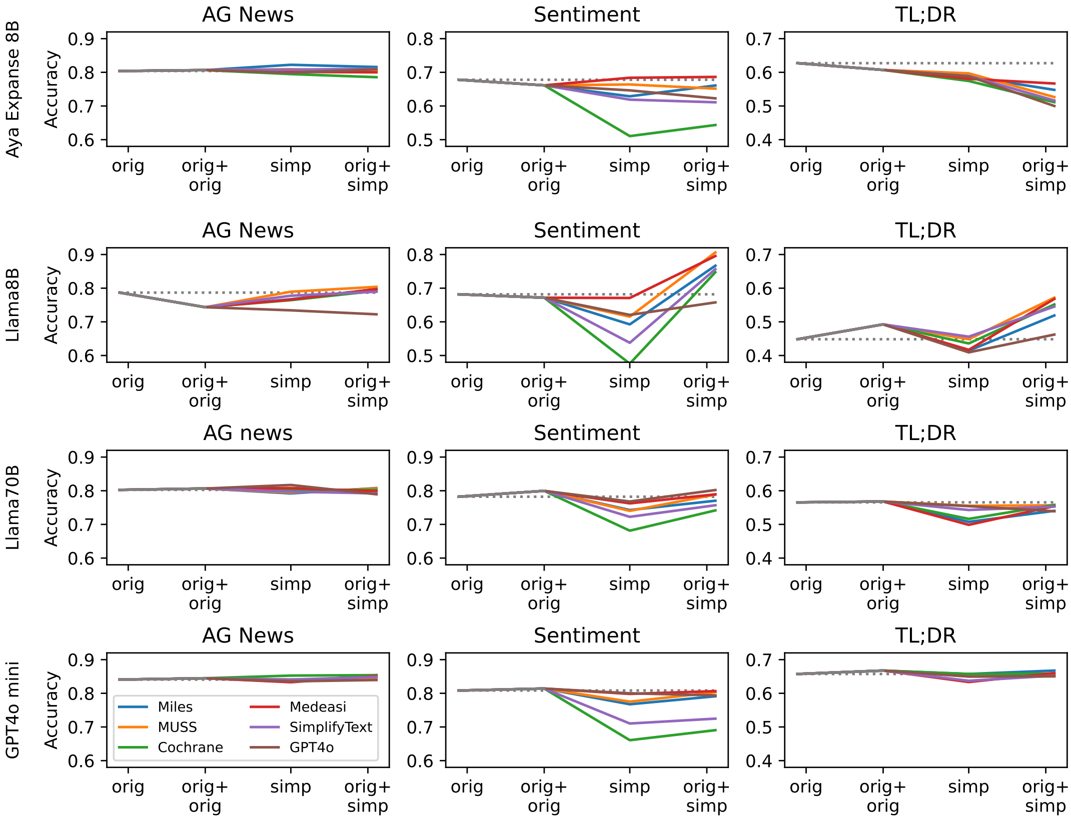

Figure 3: Combining original and simplified texts as inputs. Accuracy difference by Aya Expanse 8B, Llama8B, Llama70B, and GPT4o mini. We additionally plot the accuracy of the original data as a dashed grey line to enhance comparability.

language. Below are example prompts for classification and summarization tasks where `sample` is replaced by the text to be predicted. All remaining prompts can be found in our Github repository.

AG News (EN): {“role”: {“system”, “content”: “You are a helpful assistant. You will be provided with sentences from news articles. Classify each query into a news topic. There are four possible topics: world, sports, business or sci/tech. You must not choose another topic. Answer only with one single word and do not provide reasoning.” }, {“role”: “user”, “content”: “{`sample`} - The topic is”}

Sentiment (EN): {“role”: {“system”, “content”: “You are a helpful assistant. You will be provided with sentences from articles. Classify the sentiment of each query. There are three possible sentiments: positive, neutral or negative. You must not choose another sentiment. Answer only with one single word and do not provide reasoning.”}, {“role”: “user”, “content”: “{`sample`} - The sentiment is”}

TL;DR (EN): {“role”: {“system”, “content”: “You are a helpful assistant. You will be provided with sentences from news articles. Classify each query into a news topic. There are five possible topics: ’Sponsor’, ’Big Tech & Startups’, ’Science & Futuristic Technology’, ’Programming & Design & Data Science’ and ’Miscellaneous’. You must not choose another topic. Answer only with one single word and do not provide reasoning.” }, {“role”: “user”, “content”: “{`sample`} - The topic is”}

Gnad10 (DE): {“role”: {“system”, “content”: “Du bist ein hilfreicher Assistent. Du bekommst Sätze aus Nachrichtenartikeln. Ordne jede Anfrage einem Nachrichtenthema zu. Es gibt neun mögliche Themen: Web, Panorama, International, Wirtschaft, Sport, Inland, Etat, Wissenschaft und Kultur. Du darfst kein anderes Thema wählen. Antworte nur mit einem einzigen Wort und gib keine Begründung an.” }, “role”: “user”, “content”: “{`sample`} - Das Thema ist”}

ML SUM (DE): {“role”: {“system”, “content”: “Du bist ein hilfreicher Assistent. Du bekommst Sätze aus Nachrichtenartikeln. Ordne jede Anfrage einem Nachrichtenthema zu. Es gibt zwölf mögliche Themen: politik, wirtschaft, geld, panorama, sport, muenchen, digital, karriere, bildung, reise, auto und stil. Du darfst kein anderes Thema wählen. Antworte nur mit einem einzigen Wort und gib keine Begründung an.” },
{“role”: “user”, “content”: “{sample} - Das Thema ist”}

ML SUM (RU): {“role”: {“system”, “content”: “Ты - полезный ассистент. Тебе будут предоставлены предложения из новостных статей. Классифицируй каждый запрос в соответствии с темой новости. Темы даны на английском языке, и есть девять возможных тем: science, politics, mosobl, culture, social, incident, economics, sport, moscow. Ты не должен выбирать какую-либо другую тему. Отвечай только одним словом и не объясняй.” },
{“role”: “user”, “content”: “{sample} - Тема”}

Summarize (EN): {“role”: {“system”, “content”: “You are a helpful assistant. You will be provided with sentences from news articles. Your task is to create a headline that summarizes the content. Answer only with one sentence and don’t provide reasoning.” },
{“role”: “user”, “content”: “{sample} - The headline is”}

Summarize DE: {“role”: {“system”, “content”: “Du bist ein hilfreicher Assistent. Du bekommst Sätze aus Nachrichtenartikeln. Deine Aufgabe ist es, einen Titel zu verfassen, der den Inhalt zusammenfasst. Antworte nur mit einem Satz und gib keine Begründung an.” },
{“role”: “user”, “content”: “{sample} - Der Titel ist”}

Summarize RU: {“role”: {“system”, “content”: “Ты - полезный помощник. Тебе будут предоставлены предложения из новостных статей. Твоя задача - придумать заголовок, который обобщает содержание статьи. Отвечай только одним предложением и не приводи аргументы.” },
{“role”: “user”, “content”: “{sample} - Заголовок:”}