Revisiting German Complex Word Identification: Contextualized LLMs and Feature Injection

Thorben Schomacker[1,2], **Seid Muhie Yimam**[1], **Chris Biemann**[1], **Marina Tropmann-Frick**[2]
[1]University of Hamburg, [2]Hamburg University of Applied Sciences
thorben.schomacker@haw-hamburg.de

Abstract

Complex word identification (CWI) is essential in text simplification, yet work on German CWI remains comparatively limited. To address this gap, we investigate the capabilities of three state-of-the-art LLMs and compare them to previously proposed baseline systems. We fine-tune the LLMs in three setups: (i) using the target expression only, (ii) using the target expression together with its sentence-level context, and (iii) using the context and injection of classical machine learning features. Our results show that while pretrained-only LLMs fall short, fine-tuned LLMs set new benchmarks for both binary and probabilistic CWI. In addition, embedding the target in its context sentence improves performance, whereas feature injection has no clearly measurable effect. All models in this paper are trained on the probabilistic CWI task and additionally evaluated on the binary task; thus, we publish a single model that supports both evaluation views.

We released all accompanying resources (https://github.com/tschomacker/german-cwi-llm) *and model checkpoints* (https://huggingface.co/collections/tschomacker/german-cwi-llm).

Keywords: Complex Word Identification, Lexical Simplification, Sequence Classification, German Data

1. Introduction

This paper evaluates the capabilities of state-of-the-art LLMs (two LLaMA-style models and one BERT-based model) on German complex word identification (CWI). CWI is a key prerequisite for lexical and text simplification, as it identifies the units that are most likely to block comprehension for a given audience and therefore should be prioritized for simplification or explanation. We compare our results to 14 baseline systems previously reported for the same dataset and task setting.

In addition to the neural models, we consider length- and frequency-based features, as they have been widely used in CWI systems and are reported to be robust cross-linguistic predictors of lexical complexity (Bingel and Bjerva, 2018). More generally, the question of which features best capture word complexity has been investigated in a range of studies (see Gooding (2023, Section 2.1.3) for an overview).

Many CWI approaches operate in a context-independent setting by predicting complexity from the target expression alone, without explicitly modeling the sentence in which it occurs. However, CWI can also be formulated in ways that incorporate context information, for example as a sequence labeling problem (Gooding and Kochmar, 2019). In this work, we explicitly test whether sentence-level context improves German CWI performance when training and evaluating LLMs.

Recent LLM-based approaches have reported strong results on English CWI but more limited performance for German. Although prior work suggests that LLMs may show only modest performance for CWI (Smădu et al., 2024), we evaluate more recent models and training setups to assess current capabilities. We address the following research questions:

RQ1 Does sentence-level context information affect the performance of LLMs on German CWI?

RQ2 Do explicitly injected statistical features provide additional benefit beyond contextualized transformer representations?

RQ3 How sensitive is binary complex word identification performance to the choice of decision threshold when using probabilistic modeling?

2. Related Work

Complex word identification (CWI) is a central component of lexical and automatic text simplification pipelines (Shardlow, 2015, p. 29). While extensively studied for English, German CWI has received comparatively limited attention. Early approaches largely mirrored English work (e.g., Shardlow, 2013) and relied on surface-level indicators such as word length, syllable count, and corpus frequency.

A major step forward was the CWI 2018 Shared Task (Yimam et al., 2018), which released the first publicly available German portion of a multilingual CWI dataset annotated by L2 speakers (Yimam et al., 2017). The dataset provides both binary and probabilistic word-level complexity judgments and

Sentence	*native* **total**	*non-native* **total**	*native* **complex**	*non-native* **complex**	*complexity* **Bin.**	**Prob.**
An der Maschine wurde das <target dwds=2 len=8 syl=2 vow=2>Fahrwerk</target> beschädigt. *The machine's <target dwds=2 len=8 syl=2 vow=2>landing gear</target>* was damaged.	3	9	0	2	1	0.167
An der <target dwds=4 len=8 syl=3 vow=3>Maschine</target> wurde das Fahrwerk beschädigt. *The <target dwds=4 len=8 syl=3 vow=3>machine's</target> landing gear was damaged.*	3	9	0	0	0	0.0
Hauptgrund für die Verschlechterung des Zustandes sei der heiße und trockene Sommer 2003 mit hohen <target dwds=1 len=10 syl=3 vow=4>Ozonwerten</target> . *The main reason for the deterioration in the situation is said to be the hot and dry summer of 2003 with high <target dwds=1 len=10 syl=3 vow=4>ozone levels</target> .*	6	6	5	5	1	0.833
Hauptgrund für die <target dwds=None len=30 syl=8 vow=8>Verschlechterung des Zustandes</target> sei der heiße und trockene Sommer 2003 mit hohen Ozonwerten. *The main reason for the <target dwds=None len=30 syl=8 vow=8>deterioration in the situation</target> is said to be the hot and dry summer of 2003 with high ozone levels .*	6	6	0	1	1	0.083

Table 1: Examples from training data in the CWI 2018 dataset, represented using the preprocessing schema (see Section 3.1.2) we have developed. The first column is the preprocessed input sentence, third and fourth column is the total number of native/non-native annotators and fifth and sixth column refer to the number of annotators, who annotated the target expression as complex. **Bin.** is the binary complexity label, which is one if at least one annotator rated the target as complex and 0 in all other cases. **Prob.** is the probabilistic complexity score, which is the percentage of the annotators marked the word as complex. In our cases $2/12 = 0.167$, $0/9 = 0.0$, $10/12 = 0.833$ and $1/12 = 0.083$

established standard evaluation settings. Baseline systems relied primarily on frequency- and length-based features.

Among the shared task submissions, tree-based ensemble methods proved competitive. For example, Kajiwara and Komachi (2018) employed random forest classifiers and regressors with features including the number of characters, number of words and the frequency in a corpus written by native speakers and a corpus written by language learner. In particular, they use Lang-8 (Mizumoto et al., 2011), a learner corpus for eight languages (including German). Similarly, Bingel and Bjerva (2018) introduced CoastalCPH, combining an ensemble of feed-forward neural networks and random forests for the binary task, and random forest regressors for probabilistic prediction. We report these baselines and additional systems in Table 3. Taken together, this line of work establishes strong feature-based baselines for German CWI and highlights the usefulness of surface and frequency indicators.

Subsequent work strengthened feature-based baselines. Finnimore et al. (2019) compiled 25 features spanning target-level, subword-level, and sentence-level properties and conducted systematic ablations to derive compact cross-lingual feature sets. Their results show that carefully engineered features can rival or outperform many shared task submissions.

More recently, the BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline (MLSP) reframed complexity assessment as part of a two-stage pipeline, combining lexical complexity prediction (LCP) with lexical simplification (LS) on shared targets and contexts across ten languages, including German (Shardlow et al., 2024). In contrast to earlier CWI setups, LCP targets a continuous complexity score in $[0, 1]$, aligning complexity prediction more directly with downstream simplification. While MLSP provides German instances, the labeled data is not openly released (access is restricted), which limits direct comparability with the CWI 2018 evaluation setting.

Beyond feature-based classification, Gooding and Kochmar (2019) reformulated CWI as a sequence labeling problem. Using a BiLSTM architecture with word- and character-level representations and an auxiliary language modeling objective, they demonstrate that a unified neural model can outperform task-specific systems on English data. This motivates the question of whether modern

	total	**complex**	**simple**	**mean**
Train	6151 ($\sim$ 78%)	3589 ($\sim$ 58%)	2562	0.0783
Dev	795 ($\sim$ 10%)	334 ($\sim$ 42%)	461	0.0798
Test	959 ($\sim$ 12%)	376 ($\sim$ 39%)	583	0.0746
total	7905			

Table 2: The number and ratio of instances in the German portion and the mean value for the probabilistic label in the German CWIG3G2 dataset. Dataset split is the same as for the shared task.

LLMs, as general-purpose models, can improve performance in German CWI as well.

In parallel to these developments in feature engineering and task formulation, recent work has examined whether pretrained neural models can reduce reliance on manual features and improve generalization across domains. However, initial findings suggest that pretrained-only performance remains limited in this setting (Smădu et al., 2024). For German in particular, the CWI 2018 dataset remains the only publicly available resource at the word level, although related corpora addressing subjective sentence-level complexity exist (Naderi et al., 2019; Seiffe et al., 2022) and the MLSP could be used on demand. This resource landscape motivates further investigation into how modern pretrained models perform on German CWI.

3. Methodology

We used the only openly available German CWI corpus (Section 3.1) on three German LLMs (Section 3.2) with different experimental configurations. The results are discussed in Section 4.

3.1. Data

We use the German portion of the CWIG3G2 dataset introduced by Yimam et al. (2017) and used in the CWI 2018 shared task (Yimam et al., 2018).[1] Each instance (total $= 7905$) provides a target expression (which may be a single word or a multiword expression) together with sentence context and both binary and probabilistic complexity annotations.

3.1.1. Labels

In the dataset used for this work, binary complex word labels are defined existentially: a target is annotated as complex if at least one annotator marked it as difficult, and as simple otherwise. The probabilistic labels, in contrast, encode the proportion of annotators who judged a target as complex. As these two labels capture related but distinct notions of lexical complexity, we optimize the model with respect to the probabilistic labels and treat the binary labels as a separate evaluation view rather than as a thresholded variant of the probabilistic score.

For binary evaluation, we follow the shared-task definition and use the dataset-provided existential labels, where a target is considered complex if at least one annotator marked it as difficult. We obtain binary predictions by thresholding the model's predicted score $p = \sigma(z)$ at a decision threshold τ, selected exclusively on the development set (Section 3.4). In the German CWI dataset, there is a total of 23 annotators (12 native and 11 non-native speakers) (Yimam et al., 2017). In the German portion, each target is annotated by 12 annotators, so existential complexity ("at least one annotator") corresponds to $p \geq 1/12 \approx 0.083$. Accordingly, the probabilistic labels are quantized in steps of $1/12$, and the smallest non-zero value is 0.083.

We use this label definition and thresholding procedure consistently throughout the paper.

3.1.2. Pre-processing scheme

For our model inputs, we pre-process the dataset by marking the target span in the sentence using `<target>` tags. Depending on the experimental setup, we either provide (i) the target expression without sentence context, (ii) the tagged sentence containing the target, or (iii) the tagged sentence with injected features following. Table 1 shows examples of the resulting input format. Dataset statistics and label distributions are reported in Table 2. The examples in Table 1 also illustrate the quantization of probabilistic labels (e.g., $0.083 = 1/12$ and $0.833 = 10/12$).

3.1.3. Feature Injection

Yimam et al. (2018) use three length features for their models: the number of vowels, syllables and characters and three frequency features: the frequency in Simple Wikipedia, the frequency in the paragraph, and the frequency in the Google Web 1T 5-Grams.

Similarly, we count vowels (vow), syllables (syl) using the spaCy's syllables module[2] and characters (len). For frequency, we used the "Digitales Wörterbuch der deutschen Sprache" ($dwds$, English: digital dictionary of the German language)[3]. The DWDS assigns an integer value to a word, the

[1] www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/complex-word-identification-dataset CC-BY 4.0 License

[2] https://spacy.io/universe/project/spacy_syllables

[3] https://www.dwds.de/d/api

higher the frequency the higher the value. Multiword expressions yield `None`, since they are not supported by the DWDS API (see the 4th example in Table 1).

Concretely, we use the following input templates: (i) **target-only**: `<target> T </target>`; (ii) **context**: the full sentence with `<target>` tags around the marked span; (iii) **context+features**: the tagged sentence, followed by a feature string inside the opening tag, e.g., `<target dwds=2 len=8 syl=2 vow=2>`.

3.2. Models

To select strong German-pretrained models for our CWI experiments, we rely on external German NLU benchmarks as a proxy for general classification capability. Popular benchmark suites such as GLUE (Wang et al., 2018), SUPERGLUE (Wang et al., 2019) and HELM (Liang et al., 2023) do not provide German evaluation data. We therefore use SuperGLEBer (Pfister and Hotho, 2024), a German NLU benchmark suite with 29 tasks across classification, sequence tagging, sentence similarity, and question answering.[4]

SuperGLEBer does not include CWI; we use its *classification* leaderboard solely to guide model selection. We select the top three models by mean classification performance and fine-tune them on CWIG3G2:

1. `LSX-UniWue/LLaMmlein2Vec_7B`[5] (Wunderle et al., 2025)
2. `LSX-UniWue/ModernGBERT_1B`[6] (Wunderle et al., 2025)
3. `LSX-UniWue/LLaMmlein_7B`[7] (Pfister et al., 2025)

All experiments were conducted on a Tesla V100-PCIE-16GB GPU. In the following paragraphs we discuss the (hyper-) parameters used in our experiments. We adopted Pfister and Hotho (2024)'s experimental configuration as closely as possible for our model setups to foster comparability.

Epochs Similar to Pfister and Hotho (2024) we trained our models for 5 epochs. Additionally, we measure the performance on the dev-subset (not on `test`-subset to avoid bleeding) before training to evaluate the models' performance without fine-tuning and to eventually investigate the effects of the compared training paradigms.

Learning rate Similar to Pfister and Hotho (2024) we trained our models with a learning rate of $5\mathrm{e}{-5}$.

Batch size We initially used the same batch size (8) as Pfister and Hotho (2024) but observed that larger per-device batch sizes led to numerical instability (non-finite parameters) during early training steps. We therefore used smaller micro-batches ($= 1$) with gradient accumulation ($= 8$) to improve numerical stability while keeping the effective batch size the same as Pfister and Hotho (2024).

PEFT We apply parameter-efficient fine-tuning (PEFT) (Mangrulkar et al., 2022) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) and quantization (QLoRA). The LoRA configuration is defined by four parameters: the rank r, the scaling factor `lora_alpha`, the task type, and the target modules. We set $r = 8$ and `lora_alpha` $= 32$, following Pfister and Hotho (2024). The task type is sequence classification (`SEQ_CLS`), as we model CWI as a sequence classification task.

For `ModernGBERT_1B`, we apply LoRA adapters to the target modules `Wqkv`, `Wi`, and `Wo`, following the model card.[8] To ensure comparability across architectures, we map these target modules to the structurally corresponding layers in the LLaMA-style models (`LSX-UniWue/LLaMmlein2Vec_7B` and `LSX-UniWue/LLaMmlein_7B`). Concretely, the fused query–key–value projection (`Wqkv`) and the attention output projection (`Wo`) correspond to the attention projections `q_proj`, `k_proj`, `v_proj`, and `o_proj` in LLaMA-style architectures, yielding a comparable attention-focused LoRA setup across all models. For both LLaMA-style models, we perform 4-bit quantization using BitsandBytes.

3.3. Problem Formulation

We model complex word identification as a probabilistic binary classification task and train the model using binary cross-entropy loss. Given an input instance x (sentence and target expression), the model produces a single logit $z \in \mathbb{R}$, which is transformed into a probability via the sigmoid function. The model parameters θ are optimized by minimizing the binary cross-entropy between the predicted probability $p(\text{complex} \mid x)$ and the gold label y:

$$p(\text{complex} \mid x) = \sigma(f_\theta(x)) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

The resulting value $p \in [0, 1]$ is interpreted as a continuous complexity score, corresponding to the es-

[4] https://lsx-uniwue.github.io/SuperGLEBer-site/leaderboard_v1, last access 08.12.25

[5] https://huggingface.co/LSX-UniWue/LLaMmlein2Vec_7B, last access 19.02.26

[6] https://huggingface.co/LSX-UniWue/ModernGBERT_1B, last access 19.02.26

[7] https://huggingface.co/LSX-UniWue/LSX-UniWue/LLaMmlein_7B, last access 19.02.26

[8] https://huggingface.co/LSX-UniWue/ModernGBERT_1B , last access 03.02.2026

timated proportion of annotators who would judge the target expression as complex.

We deliberately adopt this formulation instead of discretizing lexical complexity into multiple ordinal bins and recovering a continuous score via expected values over softmax outputs, as proposed by Smădu et al. (2024). While ordinal binning provides a robust training signal for graded lexical complexity, it introduces an additional discretization step that is not required in our setting, where the target variable is already defined as a probability. Using a sigmoid-based formulation allows us to preserve the original label semantics, avoid arbitrary bin boundaries, and maintain architectural and procedural consistency with standard complex word identification setups.

3.4. Threshold Sensitivity Analysis

Since our model predicts a continuous complexity score $p = \sigma(z)$, a binary decision rule requires selecting a threshold τ. While the annotation scheme defines complexity in existential terms (i.e., at least one annotator marked the word as difficult), the exact numerical threshold depends on the normalization of the probabilistic labels.

To assess the robustness of binary performance with respect to this decision rule, we conduct a threshold sensitivity analysis. On the development set, we evaluate a small set of fixed thresholds $\tau \in \{0.042, 0.083, 0.167\}$, corresponding to values below, equal to, and above the assumed annotator-normalized threshold. We report $F1_{macro}$ for each setting while keeping the probabilistic training procedure unchanged.

Based solely on development performance, we select a single global threshold and keep it fixed for all test evaluations. This ensures that no decision rule is optimized on the test set. Test results are reported under the selected threshold and compared to existing benchmarks.

Figure 1 reports development set performance for different threshold values and training states. The selected threshold is determined exclusively based on development $F1_{macro}$.

Selected threshold We evaluate binary CWI by thresholding the predicted probabilistic score $p = \sigma(z)$ at a fixed decision threshold τ. On the development set, we compare $\tau \in \{0.042, 0.083, 0.167\}$ and select a single global threshold by maximizing $F1_{macro}$. For the fine-tuned models (5 epochs), the best development performance is obtained with $\tau = 1/24 \approx 0.042$, and we therefore fix $\tau = 0.042$ for all subsequent test evaluations.

Note that $\tau = 0.042$ lies below the smallest non-zero gold probability step ($1/12 \approx 0.083$). We nevertheless select τ strictly by development-set macro-F_1 and keep it fixed for all test evaluations.

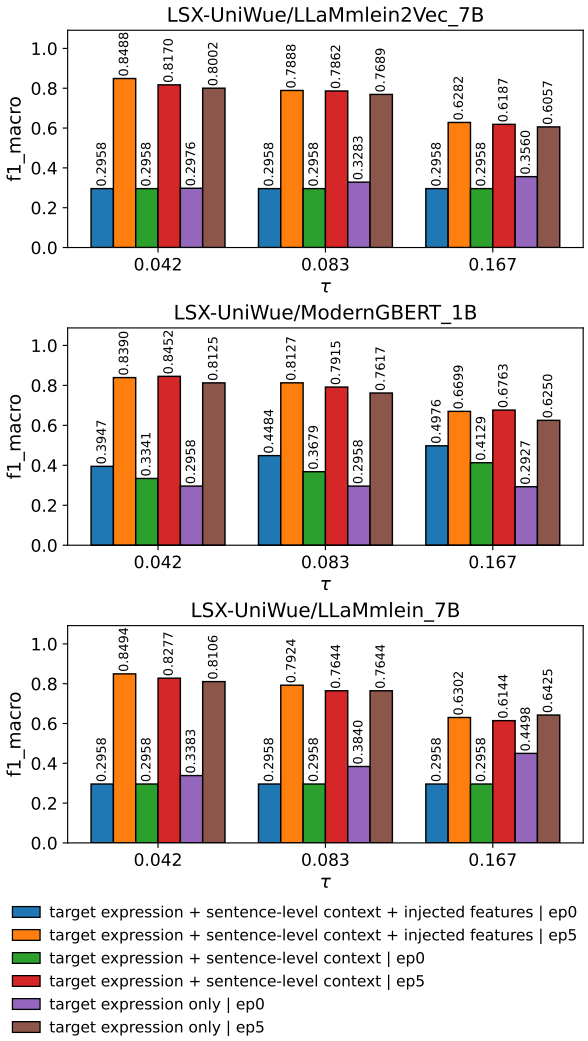

Figure 1: $F1_{macro} \uparrow$ measured on the dev-set with different setups and thresholds τ for three models. Threshold τ is selected on the development set by maximizing $F1_{macro} \uparrow$.

3.5. Evaluation

To ensure comparability, we adopt the same evaluation methodology as Yimam et al. (2018): Mean Absolute Error (MAE) for probabilistic CWI and macro-F_1 ($F1_{macro}$) for binary CWI. We train all models on the probabilistic scores and report MAE against these gold values. For binary evaluation, we follow the shared-task definition using the dataset-provided existential labels and derive binary predictions by thresholding the predicted score $p = \sigma(z)$ at the development-selected threshold $\tau = 0.042$ (Sections 3.1 and 3.4).

4. Results

Table 3 reports performance on probabilistic CWI (MAE) and binary CWI ($F1_{macro}$), where binary predictions are obtained by thresholding model

Model (sorted by publication date)					$F1_{macro}\uparrow$ DEV	$F1_{macro}\uparrow$ TEST	$MAE\downarrow$ DEV	$MAE\downarrow$ TEST
Kajiwara and Komachi (2018): TMU (CWI2018 1st rank)					-	0.7451	-	<u>0.0610</u>
Bingel and Bjerva (2018): CoastalCPH					-	0.6619	-	0.0747
Yimam et al. (2018): baseline					-	<u>0.7546</u>	-	0.0816
Finnimore et al. (2019): monolingual baseline-2-features					**0.795**	0.724	-	-
Finnimore et al. (2019): monolingual full-25-feature set					0.746	0.748	-	-
Finnimore et al. (2019): cross-lingual selected-5-features (EN+ES)					0.783	0.734	-	-
Finnimore et al. (2019): cross-lingual feature selection (ES)					0.774	0.726	-	-
Finnimore et al. (2019): cross-lingual feature selection (EN)					0.760	0.730	-	-
Smădu et al. (2024): Llama-2-7b-ft					-	0.705	-	-
Smădu et al. (2024): Llama-2-13b-ft					-	0.708	-	-
Smădu et al. (2024): Vicuna-v1.5-7b-ft					-	0.675	-	-
Smădu et al. (2024): Vicuna-v1.5-13b-ft					-	0.700	-	-
Smădu et al. (2024): Llama-3-8b-ft					-	0.708	-	-
Smădu et al. (2024): ChatGPT-3.5-turbo-ft					-	0.666	-	-
	epochs	**context**	**features**	**threshold**				
`ModernGBERT_1B`	0	none	none	0.042	0.2958	-	0.6366	-
`ModernGBERT_1B`	0	sentence	none	0.042	0.3341	-	0.5081	-
`ModernGBERT_1B`	0	sentence	all	0.042	0.3947	-	0.2667	-
`ModernGBERT_1B`	5	none	none	0.042	-	0.7626	-	0.0533
`ModernGBERT_1B`	5	sentence	none	0.042	-	0.7921	-	0.0499
`ModernGBERT_1B`	5	sentence	all	0.042	-	0.7829	-	0.0505
`LLaMmlein2Vec_7B`	0	none	none	0.042	0.2976	-	0.4464	-
`LLaMmlein2Vec_7B`	0	sentence	none	0.042	0.2958	-	0.8632	-
`LLaMmlein2Vec_7B`	0	sentence	all	0.042	0.2958	-	0.7799	-
`LLaMmlein2Vec_7B`	5	none	none	0.042	-	0.7633	-	0.0528
`LLaMmlein2Vec_7B`	5	sentence	none	0.042	-	**0.7934**	-	**0.0485**
`LLaMmlein2Vec_7B`	5	sentence	all	0.042	-	0.7921	-	0.0492
`LLaMmlein_7B`	0	none	none	0.042	0.3383	-	0.3332	-
`LLaMmlein_7B`	0	sentence	none	0.042	0.2958	-	0.6449	-
`LLaMmlein_7B`	0	sentence	all	0.042	0.2958	-	0.6952	-
`LLaMmlein_7B`	5	none	none	0.042	-	0.7808	-	0.0523
`LLaMmlein_7B`	5	sentence	none	0.042	-	0.7931	-	0.0490
`LLaMmlein_7B`	5	sentence	all	0.042	-	0.7927	-	0.0488

Table 3: Binary ($F1$) and probabilistic classification (MAE) results. Our results are rounded to the fourth digit after the floating point. Prior best results before are <u>underlined</u> and current best results are in **bold**. All reported binary scores use the selected threshold $\tau = 0.042$ (Section 3.4).

outputs at $\tau = 0.042$ selected on the development set (Section 3.4). We first compare our models to previously reported baselines and then analyze the impact of context and feature injection.

4.1. Comparison to Prior Work

Among previously reported systems, the original shared-task baseline Yimam et al. (2018) achieves the strongest binary performance on the test set ($F1_{macro} = 0.7546$), while TMU (Kajiwara and Komachi, 2018) reports the best probabilistic performance ($MAE = 0.0610$). Later LLM-based approaches evaluated by Smădu et al. (2024) reach test $F1_{macro}$ scores between 0.666 and 0.708, remaining below the strongest feature-based baselines for German.

Across all three model families evaluated in this work, fine-tuned configurations substantially outperform both prior LLM-based systems and earlier feature-based baselines. The best binary performance on the test set is achieved by the contextualized `LLaMmlein2Vec_7B` model with $F1_{macro} = 0.7934$, followed closely by `LLaMmlein_7B` with $F1_{macro} = 0.7931$. For probabilistic CWI, the lowest MAE is 0.0485 (LLaMmlein2Vec_7B, contextualized), improving considerably over the previously best reported MAE of 0.0610. To the best of our knowledge, these results constitute the strongest reported performance on the German CWIG3G2 benchmark under directly comparable evaluation settings.

4.2. Pretrained-only vs. Fine-Tuning

We use pretrained-only to denote evaluation of the pretrained sequence-classification models without any task-specific fine-tuning (0 training epochs);

this is not a prompting-based setup. We report these pretrained-only results on the development set to characterize the starting point before fine-tuning but do not report on the test set to avoid bias the experimental decision process based on test results. For instance, LLaMmlein_7B achieves $F1_{macro} = 0.3383$ (no context) and 0.2958 (with context) on the development set, with corresponding MAE values between 0.3332 and 0.6952 depending on the configuration. Similar patterns are observed for LLaMmlein2Vec_7B and ModernGBERT_1B. Fine-tuning dramatically reduces MAE and improves macro-F_1 across all model families, moving from weak pretrained-only development performance to strong test performance after five epochs.

4.3. Effect of Sentence-Level Context (RQ1)

To address RQ1, we compare target-only configurations with contextualized setups. For both LLaMA-style models, incorporating sentence context consistently improves performance.

For `LLaMmlein2Vec_7B`, adding context increases test $F1_{macro}$ from 0.7633 (no context) to 0.7934 (sentence context), while reducing MAE from 0.0528 to 0.0485. A similar trend holds for `LLaMmlein_7B`, where contextualization improves $F1_{macro}$ from 0.7808 to 0.7931 and reduces MAE from 0.0523 to 0.0490.

`ModernGBERT_1B` shows the same qualitative behavior: contextualization increases test $F1_{macro}$ from 0.7626 to 0.7921 and reduces MAE from 0.0533 to 0.0499.

Overall, sentence-level context yields consistent gains across architectures and evaluation metrics, supporting RQ1: contextual information positively affects German CWI performance in fine-tuned LLMs.

4.4. Effect of Feature Injection (RQ2)

To address RQ2, we compare contextualized setups with and without injected statistical features. The results do not indicate consistent improvements from feature injection.

For `LLaMmlein2Vec_7B`, adding features slightly decreases binary performance ($0.7934 \rightarrow 0.7921$) and slightly increases MAE ($0.0485 \rightarrow 0.0492$). For `LLaMmlein_7B`, feature injection leads to a small reduction in $F1_{macro}$ ($0.7931 \rightarrow 0.7926$) with only marginal changes in MAE ($0.0490 \rightarrow 0.0488$).

For `ModernGBERT_1B`, the differences between contextualized configurations with and without features are small and do not reveal a systematic pattern. Depending on the threshold and metric, feature injection may result in slight improvements or slight degradations, but effect sizes remain minor.

On the test set, feature injection does not yield consistent gains over the contextualized setup without features; differences are small and sometimes negative. On the development set, feature injection can improve macro-F_1 in some configurations, suggesting that its benefit may be unstable and sensitive to split/model interactions.

4.5. Threshold Sensitivity (RQ3)

RQ3 investigates how sensitive binary CWI performance is to the choice of decision threshold when models are trained on probabilistic labels. Since our models output continuous scores $p \in [0, 1]$, binary predictions require selecting a threshold τ.

For the fine-tuned models (5 epochs), development performance is highest at $\tau = 0.042$ across all model families and setups, and it decreases for larger thresholds ($\tau = 0.083$, $\tau = 0.167$). For example, for `LLaMmlein2Vec_7B` (5 epochs, contextualized), dev $F1_{macro}$ decreases from 0.8170 at $\tau = 0.042$ to 0.7862 at $\tau = 0.083$ and 0.6187 at $\tau = 0.167$. In contrast, pretrained-only (0 epochs) results do not follow a consistent monotonic pattern with respect to τ, which is expected because the unfitted classifier head yields poorly calibrated scores.

Importantly, although the gold probabilistic labels are quantized in steps of approximately $1/12$, the decision threshold still meaningfully affects predictions derived from model outputs. Selecting τ based on development performance therefore remains necessary. In our experiments, $\tau = 0.042$ consistently yields the highest development performance and is fixed for all reported test results.

Overall, binary performance is sensitive to the threshold choice, but the ranking between model architectures and experimental setups remains stable across the tested values. This indicates that while absolute $F1_{macro}$ scores vary with τ, the qualitative conclusions of this study are robust to reasonable threshold variations.

5. Conclusion

In this work, we evaluated state-of-the-art German-pretrained LLMs on complex word identification and compared them to established feature-based baselines and previously reported LLM systems. Across all three model families, fine-tuned configurations substantially outperform pretrained-only setups and set new benchmarks for both binary ($F1_{macro}$) and probabilistic (MAE) German CWI.

Our experiments show that sentence-level context consistently improves performance, confirming that lexical complexity is not purely a property

of isolated target expressions but benefits from contextual modeling. In contrast, injecting explicit length- and frequency-based features does not yield consistent gains once models are fine-tuned, suggesting that such information may already be implicitly encoded in pretrained representations.

These findings indicate that German CWI can be effectively addressed using compact, parameter-efficient fine-tuning of modern LLMs. At the same time, the large performance gap between pretrained-only and fine-tuned settings highlights that lexical complexity prediction remains a task requiring targeted supervision. Finally, we emphasize that complexity judgments are inherently audience-dependent (Gooding et al., 2021), and future work should further investigate personalized and domain-specific modeling approaches for German CWI.

6. Outlook

Beyond complex word identification as a standalone task, our models can be interpreted as a proxy for lexical complexity, thereby serving both as a modeling signal and as an evaluation bridge for LCP. In this sense, CWI approximates a decision boundary in the underlying complexity space, complementing continuous LCP approaches that aim to model graded difficulty.

A further use case is supporting the evaluation of lexical complexity models. In particular, token- or span-level complexity estimates can be aggregated to text-level signals when combined with complementary indicators (e.g., sentence-level properties Schomacker et al., 2024). Such composite assessments may provide a more informative view of text complexity than relying solely on classical readability formulas (Tanprasert and Kauchak, 2021). Initial findings (e.g., Deilen et al., 2023; Anschütz et al., 2023; Schomacker et al., 2026) showcase LLMs used end-to-end without necessarily relying on CWI as "classic" pipelines did. Having CWI-specific models still offers benefits in terms interpretability and modularity.

Although newer models will likely be released in the future, our results suggest that the qualitative trends observed in this study may extend beyond the specific models evaluated here. In particular, incorporating sentence-level context was associated with improved performance in our experiments, whereas explicit feature injection did not consistently lead to additional gains under the configurations we tested.

Finally, the BERT-based model (`ModernGBERT_1B`) may represent a practical option for deployment scenarios. Despite having substantially fewer parameters than the 7B LLaMA-style models, it achieves comparable performance in our experiments. This indicates that smaller models can, under certain conditions, provide a reasonable balance between computational efficiency and predictive quality.

7. Limitations

First, our models are trained to predict probabilistic complexity scores and we derive binary predictions by thresholding the predicted score at a fixed decision threshold ($\tau = 0.042$), selected on the development set. While this avoids tuning on the test set, binary $F1_{macro}$ values are nevertheless dependent on the chosen decision rule and should be interpreted together with the threshold sensitivity results (RQ3). In addition, the probabilistic gold labels in CWIG3G2 are quantized in steps of approximately $1/12$ (smallest non-zero value 0.083), which constrains how finely label semantics can be reflected in binary decision boundaries.

Second, the dataset is relatively small (7905 instances across train/dev/test in our split) and restricted to the CWIG3G2 domain (news) and annotation population. It nevertheless remains the only publicly available German word-level CWI dataset. We therefore prioritize controlled comparison over large-scale pretraining variation. As a result, the reported improvements may not fully generalize to other German genres, domains (e.g., administrative texts, health communication), or reader groups with different proficiency profiles.

Third, our feature injection approach is intentionally simple and may not fully exploit structured scalar features. Therefore, our findings should not be interpreted as evidence that lexical features are generally unhelpful, but rather that this particular integration strategy does not provide additional benefit over contextualized transformer representations. Our approach may underutilize the features and is sensitive to tokenization and formatting choices. More structured integration methods, e.g., dedicated feature embeddings, could yield different outcomes and remain for future work.

Finally, we evaluate only three pretrained model families and a single compute setting (one GPU type) with a fixed hyperparameter configuration. Although we align our setup with prior work for comparability, alternative choices (e.g., different LoRA target modules, ranks, quantization settings, or longer training) might affect absolute performance and potentially interact with the impact of context and feature injection. In all experiments a single random seed was used, following prior benchmark setups. While this facilitates comparability, we acknowledge that reporting variance across multiple seeds would further strengthen robustness claims.

Ethics statement

This work explores large language models for complex word identification (CWI) in German, with the goal of supporting accessibility and text simplification. While the task itself poses low direct risk, several ethical considerations apply.

Pretrained language models may encode societal biases present in their training data, which could affect how lexical complexity is assessed. In addition, the notion of "complexity" is inherently subjective and depends on reader characteristics such as language proficiency and background knowledge. Our models are trained on annotated data that reflect specific annotator perspectives and may not generalize to all user groups.

CWI systems should therefore be used as supportive tools rather than replacements for human judgment, particularly in sensitive contexts. Careful evaluation is required before deployment to ensure fair and appropriate use.

Lay Summary

Some words are harder to understand than others. This can make texts difficult to read, especially for people with lower reading skills or those learning a language. Our research looks at how artificial intelligence (AI) can automatically find these difficult words. This task is called complex word identification (CWI).

We focus on German and test modern AI language models to see how well they can recognize difficult words. These models are already trained on large amounts of text, but we also adapt them further using a small amount of task-specific data. This process is called fine-tuning.

Our results show that these fine-tuned models can identify difficult words very accurately. They perform better than earlier methods that rely heavily on manually designed features, such as word length or how often a word appears in texts.

We also find that context matters: a word may be easy or difficult depending on the sentence it appears in. Models that consider the full sentence perform better than those that look at words in isolation.

Overall, our work shows that modern AI models can be an effective tool for detecting difficult words in German sentences. This could support applications such as simplifying texts, improving accessibility, or helping people better understand written information.

8. Bibliographical References

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bart\lomiej 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. URL: https://aclanthology.org/2023.findings-acl.74, doi:10.18653/v1/2023.findings-acl.74.

Joachim Bingel and Johannes Bjerva. 2018. Cross-lingual complex word identification with multitask learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 166–174, New Orleans, Louisiana, USA. Association for Computational Linguistics. URL: https://aclanthology.org/W18-0518/, doi:10.18653/v1/W18-0518.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in Easy Language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria. URL: https://aclanthology.org/2023.tsar-1.1/.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong Baselines for Complex Word Identification across Multiple Languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 970–977, Minneapolis, Minnesota, USA. Association for Computational Linguistics. URL: https://aclanthology.org/N19-1102, doi:10.18653/v1/N19-1102.

Sian Gooding. 2023. *A Personalised Approach to Lexical Complexity*. Doctoral Thesis, University of Cambridge, Cambridge, England. URL: https://doi.org/10.17863/CAM.116567.

Sian Gooding and Ekaterina Kochmar. 2019. Complex Word Identification as a Sequence Labelling Task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics. URL: https://aclanthology.org/P19-1109/, doi:10.18653/v1/P19-1109.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word Complexity is in the Eye of the Beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics. URL: https://aclanthology.org/2021.naacl-main.351, doi:10.18653/v1/2021.naacl-main.351.

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 *Proceedings of the International Conference on Learning Representations*, Online. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex Word Identification Based on Frequency in a Learner Corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana, USA. Association for Computational Linguistics. URL: https://aclanthology.org/W18-0521/, doi:10.18653/v1/W18-0521.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. URL: https://openreview.net/forum?id=iO4LZibEqW.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. URL: https://github.com/huggingface/peft.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective Assessment of Text Complexity: A Dataset for German Language. Version Number: 1. URL: https://arxiv.org/abs/1904.07733, doi:10.48550/ARXIV.1904.07733.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German Language Understanding Evaluation Benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics. URL: https://aclanthology.org/2024.naacl-long.438/, doi:10.18653/v1/2024.naacl-long.438.

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. URL: https://aclanthology.org/2025.acl-long.111/, doi:10.18653/v1/2025.acl-long.111.

Thorben Schomacker, Miriam Anschütz, Regina Stodden, Georg Groh, and Marina Tropmann-Frick. 2024. Overview of the GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE). In *Proceedings of GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)*, pages 1–14, Vienna, Austria. Association for Computational Lingustics. URL: https://aclanthology.org/2024.germeval-1.1.

Thorben Schomacker, Burak Tinman, Chris Biemann, and Marina Tropmann-Frick. 2026. LLMs for Easy Language Translation: A Case Study on German Public Authorities Web Pages. In *KI 2025: Advances in Artificial Intelligence*, pages 252–261, Cham. Springer Nature Switzerland. doi:10.1007/978-3-032-02813-6_20.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective Text Complexity Assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources

Association. URL: https://aclanthology.org/2022.lrec-1.74.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics. URL: https://aclanthology.org/P13-3015/.

Matthew Shardlow. 2015. *Lexical Simplification: Optimising the Pipeline*. Dissertation, University of Manchester, Manchester, United Kingdom. URL: https://pure.manchester.ac.uk/ws/portalfiles/portal/54575176/FULL_TEXT.PDF.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics. URL: https://aclanthology.org/2024.emnlp-main.933/, doi:10.18653/v1/2024.emnlp-main.933.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics. URL: https://aclanthology.org/2021.gem-1.1, doi:10.18653/v1/2021.gem-1.1.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. URL: http://aclweb.org/anthology/W18-5446, doi:10.18653/v1/W18-5446.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New Encoders for German Trained from Scratch: Comparing ModernGBERT with Converted LLM2Vec Models. ArXiv:2505.13136 [cs]. URL: http://arxiv.org/abs/2505.13136, doi:10.48550/arXiv.2505.13136.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, USA. Association for Computational Linguistics. URL: https://aclanthology.org/W18-0507/, doi:10.18653/v1/W18-0507.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822, Varna, Bulgaria. INCOMA Ltd. URL: https://aclanthology.org/R17-1104/, doi:10.26615/978-954-452-049-6_104.