
CLASSIFYING GERMAN LANGUAGE PROFICIENCY LEVELS USING LARGE LANGUAGE MODELS

Elias-Leander Ahlers

Computer Science Department
University of Münster
Münster, Germany

elias.ahlers@uni-muenster.de

Witold Brunsma

Computer Science Department
University of Münster
Münster, Germany

witold.brunsmann@uni-muenster.de

Malte Schilling

Computer Science Department
University of Münster
Münster, Germany

malte.schilling@uni-muenster.de

ABSTRACT

Assessing language proficiency is essential for education, as it enables instruction tailored to learners' needs. This paper investigates the use of Large Language Models (LLMs) for automatically classifying German texts according to the Common European Framework of Reference for Languages (CEFR) into different proficiency levels. To support robust training and evaluation, we construct a diverse dataset by combining multiple existing CEFR-annotated corpora with synthetic data. We then evaluate prompt-engineering strategies, fine-tuning of a LLaMA-3-8B-Instruct model and a probing-based approach that utilizes the internal neural state of the LLM for classification. Our results show a consistent performance improvement over prior methods, highlighting the potential of LLMs for reliable and scalable CEFR classification.

1 Introduction

Large Language Models (LLMs) such as GPT, LLaMA, and Gemini have demonstrated impressive capabilities in processing and generating human-like text (Minaee et al., 2025). They are now broadly used, for example, as chatbots, for providing customer support (Xu et al., 2024) or as tutoring systems (Kahl et al., 2024).

However, their use as specialized classifiers for complex, fine-grained tasks remains underexplored. One such task is the classification of language proficiency levels as, for example, defined by the Common European Framework of Reference for Languages (CEFR). The CEFR framework defines six proficiency levels (A1–C2), each reflecting distinct linguistic capabilities. Accurate classification into these levels requires models to capture subtle differences in vocabulary, grammar, and discourse. Traditional approaches for CEFR classification rely on manually engineered features and classical machine learning methods such as SVMs or neural networks (Caines and Butterly, 2020; Szűcsyi et al., 2019). While effective to a certain degree, these methods have shown to be labor-intensive and often limited in scalability.

This paper investigates whether LLMs can serve as effective alternatives for CEFR classification. We approach this through systematically evaluating and comparing different strategies that employ LLMs, using prompt engineering, fine-tuning, and a probing based classifier. Prompt engineering explores how the phrasing of model instructions affects performance. Fine-tuning adapts a pre-trained model to the specific classification task. The probing based approach extracts the internal neural state of the Large Language Model of the last layer and utilizes this as an input for a neural network based classifier which is trained in a supervised fashion on a subset of the dataset.

A challenge lies in the lack of a comprehensive, balanced dataset for German CEFR classification. Existing corpora often focus on selected levels (e.g., A2–B2) or lack standardized labeling. To address this, we, as a first contribution, develop a new dataset by combining established corpora with synthetic data generation. This allows us to achieve a balanced distribution across the different CEFR levels (A1–C2) which enables a more reliable evaluation of classification strategies. Our experiments show that using LLMs when supported by a carefully constructed dataset and optimized through prompt engineering, fine-tuning and probing offer a powerful and scalable alternative to traditional approaches for CEFR classification and related linguistic assessment tasks. Considering the internal state of the LLM improves the results to a certain degree. The remainder of the paper is organized as follows. The next section gives an overview on datasets with a focus on German as well as introduces the current state-of-the-art machine learning approaches. The third section explains the construction of a balanced and augmented dataset as well as the different LLM-based classification approaches, we present our results in a comparative evaluation and conclude with a brief discussion.

2 Related Work

In this section, we provide background on key aspects relevant to our study: first, existing datasets for German CEFR classification. Secondly, traditional approaches to modeling language proficiency using classical machine learning methods which we will use as a baseline for comparison.

2.1 Overview of CEFR Datasets in German

Research on CEFR classification for German relies on a small number of publicly available, annotated corpora. In this section, we describe the main datasets (overview, see Table 1) used in our study, focusing on their structure, proficiency level coverage, and suitability for classification tasks.

2.1.1 Falko Corpus

The *Falko corpus* (Reznicek et al., 2010), collected and annotated at Humboldt University Berlin, was designed to facilitate research in corpus linguistics and second language acquisition. It contains a collection of German texts written under controlled conditions by both advanced foreign learners and native speakers, and it includes rich linguistic annotations which include a test score that can be mapped to CEFR levels.

2.1.2 MERLIN Corpus

The *MERLIN corpus* (Boyd et al., 2014) was specifically designed to illustrate CEFR levels using authentic learner texts in German, Italian, and Czech. The corpus covers all six CEFR levels (A1–C2), with a particularly strong representation of the A2, B1, and B2 levels.

The corpus includes a variety of text types, e.g., essays and summaries, collected as part of standardized language tests, ensuring consistency and quality while providing diversity of samples. All texts were annotated with CEFR levels by language experts, providing reliable classifications for our dataset that could be directly included without further processing.

Table 1: Distribution of German CEFR-labeled texts across datasets and levels. Text counts are shown per source and aggregated per CEFR level.

Source	A1	A2	B1	B2	C1	C2
Falko EssayL2				83	84	81
Falko SummaryL1						58
Falko SummaryL2					53	53
MERLIN	57	306	331	293	42	4
Synthetic	122					
Our Dataset (1,567)	179	306	331	376	179	196

2.1.3 Additional Resources

While CEFR-labeled resources for German are relatively limited in size and scope, there are significantly larger datasets available for English language proficiency. Most notably, the CEFR-SP dataset (Uchida et al., 2024) contains approximately 17,000 English sentences labeled by language experts. Other significant English datasets include the Cambridge Learner Corpus and the EF-Cambridge Open Language Database (Geertzen et al., 2014).

Overall, these datasets demonstrate the potential scale and variety possible in CEFR-level classification resources, though similar comprehensive collections for German remain a work in progress.

2.2 Traditional Approaches to CEFR Classification

Existing research on automatic CEFR classification in German has mainly focused on feature-based approaches using classical machine learning method. Two main approaches in this area are by Vajjala and Loo (Vajjala and Rama, 2018), extended by Caines et al. (Caines and Butterly, 2020), and, secondly, by Szüügyi et al. (Szüügyi et al., 2019).

Vajjala and Rama (Vajjala and Rama, 2018) proposed to combine handcrafted linguistic features with neural network classifiers to classify German CEFR levels. The authors extracted a variety of features from the texts, including lexical, syntactic, and semantic indicators, which they used to train a neural network classifier. This work, has been extended by Caines et al. (Caines and Butterly, 2020), who incorporated additional linguistic features and applied it to two other languages (Spanish and English) as well. Their system achieved a weighted F1 score of 0.702 for German, demonstrating the effectiveness of feature-based methods in CEFR classification.

Szüügyi et al. (Szüügyi et al., 2019) used a Support Vector Machine (SVM) to classify German CEFR levels on linguistic features from the texts, which again included lexical, syntactic, and semantic features. Furthermore, they tested incorporating a Multi-Layer Perceptron (MLP) classifier. In their work, they simplified the problem by merging the CEFR levels only into the three broader categories: A(A1,A2), B(B1,B2), and C(C1,C2). For these three classes, they report an accuracy of 82%.

These methods rely heavily on features that are hand-crafted by experts and partially operate only on reduced versions of the CEFR scale. This limits their applicability and scalability to real-world applications.

3 Methods

We evaluate several leading Large Language Models for the CEFR classification task, taking into account factors such as model size, inference speed, and performance on German texts (see 4.2). After comprehensive testing and an evaluation of different prompts, we selected the LLaMA-3-8B-Instruct model (Dubey et al., 2024) as our base model for fine-tuning. In the following sections, we will introduce the main parts of our approach: First, the construction of a dedicated and balanced CEFR dataset for German which requires data augmentation. Second, the development and evaluation of various prompt engineering strategies. Third, we present a fine-tuning pipeline based on LLaMA-3-8B-Instruct to assess the effect of supervised adaptation on classification performance. Fourth, we introduce a probing-based analysis in which a neural classifier is trained on the final-layer activations of the LLaMA-3 model to investigate how CEFR-relevant linguistic features are encoded in its internal representations.

3.1 Construction of a Balanced CEFR Dataset

The creation of a comprehensive German CEFR classification dataset presented several challenges due to the limited public availability of CEFR-labeled texts and in particular, as existing datasets tend to be sparse and imbalanced. We addressed this by combining multiple existing corpora along with synthetic generated data to ensure a balanced representation across proficiency levels as required for a robust evaluation.

As a first source, we used the already introduced Falko corpus as it contains high-quality texts for CEFR classification. We selected three subcorpora to ensure a diverse range of text types and proficiency levels:

- **EssayL2**, consists of argumentative essays written by advanced foreign learners of German (CEFR levels between B2 to C2). Topics include feminism, financial policy, and crime. All texts were written under controlled conditions within a 90-minute time limit.
- **SummaryL1**, contains academic summaries from native speakers, which we used as reference samples for the C2 level.
- **SummaryL2**, includes similar summaries written by learners of German as a second language.

Together, these subcorpora offer valuable examples of high-proficient language use and were instrumental in anchoring the upper end of the CEFR scale in our dataset. The EssayL2 subcorpus contained texts from B2 to C2, each annotated with metadata including the author’s C-test score, an integer ranging from 0 to 100 measuring German proficiency. We used the Falko authors’ mapping scheme to assign CEFR levels to each text based on these scores (see Table 2).

Table 2: Mapping of FALKO C-Test Scores to CEFR Levels

Score Range	CEFR Level
60–79	B2
80–89	C1
90–100	C2

We also added the MERLIN corpus which provides a diverse set of texts across all CEFR levels, with a strong focus on A2, B1, and B2 levels. The corpus includes essays and summaries written by learners of German as a second language and is annotated with CEFR levels by language experts.

To address the remaining gap at the lower end of the proficiency spectrum, in particular the lack of A1-level texts, we employed a synthetic data generation approach using a large language model. We chose the Claude 3.5 Sonnet model (Anthropic, 2024), an advanced LLM developed by Anthropic. While we used LLaMA-3-8B-Instruct as our classification model, we opted for Claude for data generation for several reasons. First, Claude has demonstrated superior ability in following complex instructions and maintaining consistent quality (Huang et al., 2024). Additionally it shows good performance on German language understanding (Park et al., 2024). Second, as a larger model with a broader training data coverage, it allows for more nuanced and authentic generation of German learner texts (Kim et al., 2024). The prompt engineering process for data generation underwent several iterations until a desired output quality was reached. Initial attempts produced texts that were too simplistic or thematically biased toward primary school topics. To overcome this, we refined the approach and adjusted the prompts to avoid stereotypical beginner subjects, incorporated explicit CEFR A1 criteria from the Goethe Institute, and used prompts written in German, which produced noticeably improved results. The final prompt is included in the appendix (appendix section A). Using this approach, we generated 122 synthetic texts. Each sample was manually reviewed to ensure appropriateness and quality before inclusion in the dataset.

The final dataset consists of 1,567 texts spanning across all six CEFR levels (A1–C2). Table 1 shows the distribution of texts by source and level.

3.2 Prompt Engineering for CEFR Classification

We developed and refined a prompting strategy through several iterations, each aimed at improving the model’s classification accuracy. Prompt engineering plays a crucial role in aligning the model’s behavior with the task objective by structuring inputs and instructions in a way that leverages the model’s internal representations (White et al., 2023). A well-designed prompt should guide the model to focus on relevant textual features in order to produce consistent label predictions, even without access to explicit training signals. In particular, prompts should clearly define the task as well as the expected output format to ensure reliable and interpretable model responses.

We started with an **English base prompt** (see Appendix D for full prompts, Table 8) that formulates the classification task in explicit terms: it instructs the model to assess a given text and assign a proficiency level according to the CEFR scale. The prompt specifies both the decision criterion (CEFR classification) and the required response format (a single level label, such as A1–C2), ensuring that the model focuses on the intended task and avoids generating additional explanations.

Next, we switched to German language prompts as Pires et al. found (Pires et al., 2019) that LLMs often show a better performance when the prompt language matches the task language. In initial tests, the model frequently included justifications for its classification when using the German prompt. This might skew the results and, therefore, we appended an instruction that the model should suppress any additional explanation or reasoning for the classification. The final German prompt (**German zero-shot prompt**, given in Appendix D, Table 8) emphasizes both the classification standard and the required output format, helping to reduce noise in the model’s predictions.

Building on this, we implemented a few-shot prompting strategy (**German few-shot prompt**, see Table 8 in Appendix D). The prompt included the base instruction and incorporated an example for each CEFR level, demonstrating typical text characteristics at each level. Examples were carefully selected from previously misclassified texts to highlight common challenges and improve the model’s ability to distinguish between adjacent levels. By explicitly showing

Table 3: Hyperparameters used for fine-tuning the LLaMA-3-8B-Instruct model.

Hyperparameter	Value
Training Settings	
Learning rate	2e-4
Number of epochs	5 (cutoff after 3)
Batch size	1
Gradient accumulation steps	1
Warmup steps	400
Weight decay	0.001
Learning rate scheduler	Linear
LoRA Configuration	
LoRA rank (r)	32
LoRA alpha	32
LoRA dropout	0.03
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Bias	none
Gradient checkpointing	True
Use DORA / RSLoRA	False

example texts at each level, the few-shot format provided additional context for the LLM and served as an implicit decision guide.

3.3 Fine-Tuning LLaMA-3 for CEFR Classification

To complement our prompt-based experiments, we fine-tuned a Large Language Model specifically on the CEFR classification task. Fine-tuning allows a model to adapt more closely to domain-specific data and structure (Devlin et al., 2019). It can yield significant improvements in performance over prompt-based strategies.

We selected the LLaMA-3-8B-Instruct model (Dubey et al., 2024) as the foundation for fine-tuning. This choice was driven by several factors: the model has a balanced size of 8 billion parameters, it is openly available, and has demonstrated strong capabilities in instruction following which is essential for adapting to our structured classification task. Moreover, the model has shown promising performance in German-language understanding.

Fine-tuning follows a supervised learning approach, for which we used the dataset introduced in Section 3.1. The dataset was split into a training (154 samples per CEFR level) and a test set (25 samples per level), ensuring balanced representation across proficiency levels. Training samples were formatted following the LLaMA-3 input structure, with system prompts, user messages, and assistant responses clearly separated by special tokens (using the LLaMA-3 tokenizer and paired with their corresponding CEFR labels). For fine-tuning, we used LoRA (Hu et al., 2021). Training was performed on a NVIDIA RTX A6000 GPU, with the process completing in approximately 20 minutes. Fine-tuning was conducted using the hyperparameters given in Table 3.

3.4 Probing-based Neural Classifier

Large language models (LLMs) contain internal states which refers to representations that typically remain hidden during inference. Probing of internal states allows to trace the internal processing of the LLM and has shown to contain additional information (Ameisen et al., 2025; Lindsey et al., 2025). As one example, internal states have previously been applied in hallucination detection, where classifiers were trained using the internal representations as an input and were able to identify hallucinated outputs with higher accuracy than response-based methods (Azaria and Mitchell, 2023; Ridder and Schilling, 2024b). This demonstrates that internal states contain rich information about the model’s internal reasoning, beyond what is expressed in its generated text.

For all experiments, we used LLaMA-3-8B. We chose the base version rather than the Instruct model, since our setup does not require text generation but relies solely on extracting contextualized embedding vectors (CEVs) for the given text samples. The instruction-tuned variant is optimized for producing conversational and helpful continuations, which could introduce unwanted biases in the internal representations. To avoid such effects, we used the base model for

all CEV extractions. Input to the LLM were the tokenized texts from the dataset. These sequences of texts were passed to the model for processing. Internal states of the last layer for the last token of a text were extracted using the Transformers library (Hugging Face, 2025), which provides direct access to the final-layer embeddings for all tokens in a sequence.

We trained a multi-layer perceptron (MLP) classifier on the CEFR-annotated dataset described in a supervised fashion. The final MLP architecture consists of four fully connected layers (`input_size=1024-512-256-6`) with ReLU activations and softmax in the output layer, following previous probing architectures (Ridder and Schilling, 2024a; Su et al., 2024). Training used a learning rate of 1×10^{-3} and an L2 regularization of 0.001 (hyperparameters were tuned using a grid search over architecture, learning rate, and regularization). The network receives the contextualized embedding vectors (CEVs) from the last hidden layer of the LLaMA-3-8B model as input, with the assigned CEFR proficiency label as the target output. These embeddings served as input features for the classifier. As the data sets are small, we used five fold cross validation to evaluate the classification performance.

4 Results

In this section, we present the results of our CEFR classification experiments, structured into four main parts: prompt engineering, comparison of different LLMs, probing approach, and fine-tuning. We evaluated the model’s performance using standard classification metrics including accuracy, precision, recall, and F1 score. Additionally, we introduced a *group accuracy* metric to account for the continuous nature of CEFR levels, considering classifications of adjacent levels as correct. As another metric, we defined a *mean classification distance*, which quantifies the average gap between predicted and true language proficiency levels. Distances are assigned as integer penalties, with adjacent levels counting as 1 and more distant errors receiving proportionally larger penalties. In contrast to binary accuracy, this approach provides more nuanced evaluation by weighting the severity of classification errors. This approach better reflects the practical application of CEFR classification, as boundaries between levels are at least to a certain degree subjective and fluid. For selected setups, we additionally report confusion matrices to highlight common misclassifications between adjacent CEFR levels. Together, these results offer insights into the strengths and limitations of LLM-based CEFR classification and help identify promising directions for further improvement.

4.1 Prompt Engineering Results

To evaluate the impact of prompt design on classification accuracy, we tested the LLaMA-3-8B-Instruct model using several prompt variants applied to our CEFR-labeled dataset (see Section 3.1), which includes a balanced distribution of 1,567 learner texts across all six CEFR levels. These included an English base prompt, a German zero-shot prompt, and a German few-shot prompt. Table 4 summarizes the performance of each prompt type.

The **English Base Prompt** served as our initial method for instructing the model to classify texts according to CEFR levels. While straightforward in its instruction to classify texts according to CEFR levels, performance was limited (overall performance, see Table 4, for detailed metrics on each level see appendix C, Table 7). This held also true for specific levels, for instance, the model achieved a precision of 0.471 and recall of 0.640 for intermediate levels like B1. The corresponding confusion matrix revealed a tendency to default to middle-range levels (B1–B2), suggesting that the model struggled to differentiate between extreme proficiency levels (see Fig. 1 (a)). Particularly notable is the model’s tendency to misclassify A1 and A2 texts as a B1 level, with 21 out of 25 A1 texts and all 25 A2 texts being incorrectly classified as B1. This showed as well in the *mean classification distance* of 1.12. The systematic error

(a) English Base Prompt			(b) German Zero-Shot Prompt			(c) German Few-Shot Prompt										
Actual	Predicted		Actual	Predicted		Actual	Predicted		Actual	Predicted		Actual	Predicted			
	A1	A2	B1	B2	C1	C2	A1	A2	B1	B2	C1	C2	A1	A2	B1	
A1	3	1	21	0	0	0	5	7	6	0	0	0	15	5	5	0
A2	0	0	25	0	0	0	0	4	14	0	0	0	3	9	13	0
B1	0	0	24	1	0	0	0	1	23	1	0	0	0	1	16	7
B2	0	0	17	8	0	0	0	0	21	4	0	0	0	0	0	21
C1	0	0	7	18	0	0	0	0	6	19	0	0	0	0	8	6
C2	0	0	0	25	0	0	0	0	6	19	0	0	0	0	1	22

Figure 1: Confusion matrices for different prompt engineering approaches using the LLaMA-3-8B-Instruct model: (a) English Base Prompt, (b) German Zero-Shot Prompt, (c) German Few-Shot Prompt. Each matrix visualizes predicted CEFR levels (columns) against true labels (rows), with cell shading indicating prediction density. The mean classification distances are: English Base Prompt = 1.120, German Zero-Shot Prompt = 1.051, German Few-Shot Prompt = 0.467.

Table 4: Performance comparison across different prompt types, the neural network-based probing classifier, and the fine-tuning approach (group accuracy includes adjacent levels).

Prompt Name	Accuracy	Group Accuracy
English Base Prompt	23.3%	64.6%
German Zero-Shot Prompt	33.3%	75.3%
German Few-Shot Prompt	59.3%	94.0%
Probing NN Classifier	65.83%	99.2%
Fine-tuned Instruct	LLaMA-3-8B-	76.7%
		100.0%

pattern suggested that the English-language prompt might be limiting the model’s ability to accurately assess German language proficiency.

The transition to a German prompt marked a significant improvement in classification performance. The **German Zero-Shot Prompt** demonstrated more nuanced classification abilities, particularly for intermediate proficiency levels. This approach showed improved differentiation between adjacent levels, with the confusion matrix revealing more balanced distribution of classifications across CEFR levels (see Fig. 1 (b)). Overall, these improvement suggests that language alignment between the prompt and the classification task plays a crucial role in model performance (*mean classification distance* improved as well to 1.051).

Next, we observed further substantial improvements in classification accuracy across all proficiency levels for the **German Few-Shot Prompt**. The confusion matrix revealed a more diagonal pattern, indicating better discrimination between adjacent proficiency levels (see Fig. 1 (c); detailed metrics are provided in appendix C, Table 7)). Especially interesting was the model’s improved handling of extreme proficiency levels. A1 classification showed a precision of 0.833 and recall of 0.600, while C2 classification achieved a precision of 0.629 and recall of 0.880 (*mean classification distance* of 0.467 highlights this improvement as well). This balanced performance across the proficiency spectrum suggests that the few-shot examples helped the model develop a more nuanced understanding of the characteristics associated with each CEFR level.

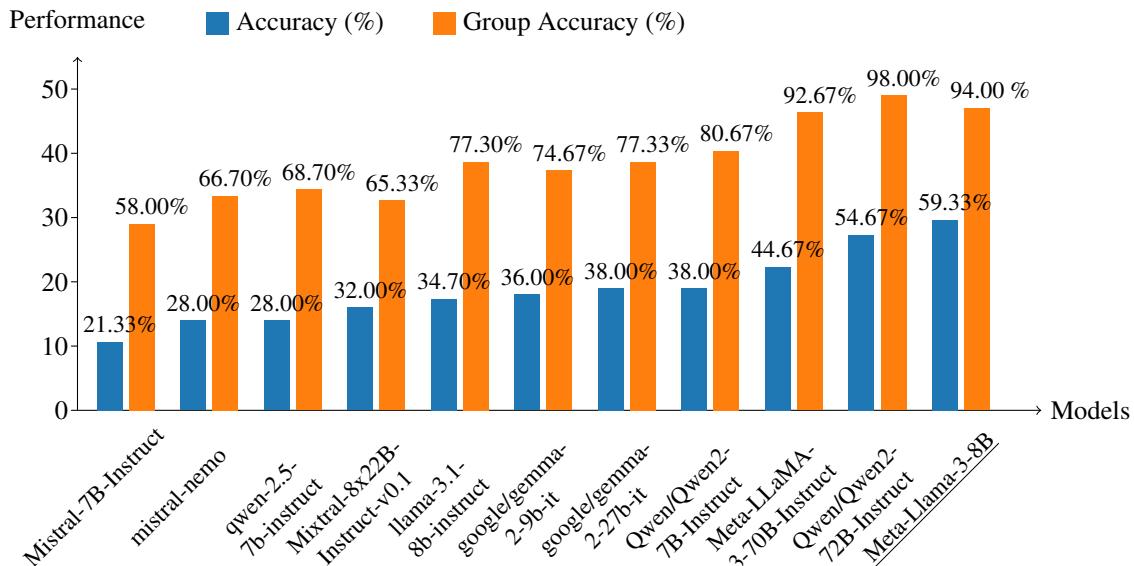


Figure 2: Performance comparison of different language models on CEFR classification, showing both exact accuracy and group accuracy (includes adjacent levels), sorted by Accuracy (names are model names as found on huggingface).

4.2 Performance Comparison Across LLMs

To evaluate the baseline capabilities of various pre-trained language models on the CEFR classification task, we tested twelve widely-used instruction-tuned LLMs on our balanced dataset using the German few-shot prompt. Models were selected to represent different architectures, parameter sizes, and training sources. Each model was prompted in German using the same CEFR classification instruction to ensure consistency across evaluations. We recorded both exact classification accuracy and a form of a more relaxed group-level accuracy (that includes adjacent levels). Figure 2 illustrates the performance comparison across all evaluated models.

4.3 Performance of the Probing Neural Network Classifier

A neural network based classifier was trained that uses the full internal state of the last layer as an input and maps this to a class label. Test results for the classifier are shown in Table 5 when using five-fold cross validation. The accuracy achieved was 65.83% with a precision of 0.665, recall of 0.658, , and an F1 score of 0.659. Additionally, the group accuracy was 99.2%, indicating that almost all misclassifications occurred between adjacent levels only. The corresponding confusion matrix is given in Figure 3 which further shows that there are only small deviations. Overall, the neural network classifier shows an improvement across all metrics when compared to querying the LLM using only language (see Table 4 for comparison of the results). This indicates that the internal states contain additional supporting information.

4.4 Fine-tuned Model Performance Analysis

To evaluate the benefits of model adaptation, we fine-tuned the LLaMA-3-8B-Instruct model on our balanced CEFR dataset which showed the highest accuracy (Fig. 2). This aims for the model to learn domain-specific patterns in the selected language and better distinguish between subtle linguistic features associated with each CEFR level. The fine-tuning process revealed consistent and stable improvement in model performance. The training loss curve showed a steady decrease across epochs, with significant improvements in the early stages followed by gradual stabilization. The stabilization of training loss around epoch 3 suggested optimal convergence without overfitting, validating our choice of training parameters. This pattern was particularly encouraging given the relatively small size of our training dataset, indicating very fast but still effective learning from limited examples.

Table 5: Performance metrics for the neural network based classifier across CEFR levels. The classifier was trained on the internal CEV states of the LLM as an input and produces the corresponding class labels. Results are mean results on the test set for 5-fold cross validation.

Class	Precision	Recall	F1 Score
A1	0.889	0.800	0.842
A2	0.600	0.750	0.667
B1	0.524	0.550	0.537
B2	0.632	0.600	0.615
C1	0.647	0.550	0.595
C2	0.700	0.700	0.700
Weighted Avg	0.665	0.658	0.659

Actual	Predicted					
	A1	A2	B1	B2	C1	C2
A1	16	4	0	0	0	0
A2	2	15	3	0	0	0
B1	0	6	11	3	0	0
B2	0	0	7	12	0	1
C1	0	0	0	4	11	5
C2	0	0	0	0	6	14

Figure 3: Confusion matrix for the neural network based classifier, highlighting a reduced confusion between adjacent CEFR levels.

Actual	Predicted					
	A1	A2	B1	B2	C1	C2
A1	21	4	0	0	0	0
A2	3	18	4	0	0	0
B1	0	2	21	2	0	0
B2	0	0	4	16	5	0
C1	0	0	0	4	21	0
C2	0	0	0	0	7	18

Figure 4: Confusion matrix for the fine-tuned LLaMA-3-8B model, showing improved accuracy and reduced confusion between neighboring CEFR levels.

The fine-tuned LLaMA-3-8B-Instruct model demonstrated substantial improvements across all six proficiency levels, achieving a weighted F1 score of 0.769 across the complete CEFR spectrum. The resulting confusion matrix, shown in Figure 4, demonstrates substantially improved classification accuracy and a more balanced performance across all CEFR levels. The distribution of misclassifications shows a clear pattern of only adjacent-level errors. This pattern is particularly clear in the B1-C1 range, where all misclassifications occur between neighboring proficiency levels. For instance, B2 texts were only misclassified to either B1 or C1, suggesting that the model maintains a reasonable understanding of the proficiency spectrum even when making errors. This pattern of adjacent-level errors is particularly relevant when considering the continuous nature of language proficiency assessment. When accounting for this through our group accuracy metric, which considers classifications of adjacent levels as partially correct, the model achieves a perfect group accuracy of 100% on the test set. This result indicates that even when the model makes errors, it maintains a fundamentally good understanding of the proficiency spectrum. An improved *mean classification distance* of as low as 0.233 for the fine-tuned model further demonstrates that adaptation worked quite well.

The fine-tuned model showed distinct performance patterns across different proficiency levels, revealing both strengths and areas for potential improvement. At the extremes of the proficiency spectrum, the model showed particularly strong performance, with A1 and C2 levels achieving the highest F1 scores of 0.857 and 0.837 respectively (Table 6).

The A1 level performance is particularly noteworthy, combining high precision (0.875) with strong recall (0.840). This balanced performance suggests robust capabilities in identifying beginner-level texts, a crucial ability for practical applications in language assessment. The perfect precision (1.000) achieved at the C2 level, although with lower recall (0.720), indicates high reliability when the model identifies advanced-level texts. In the intermediate range, we observed more nuanced performance patterns. B1 level texts were classified with strong consistency, achieving an F1 score of 0.778 with notably high recall (0.840). This suggests effective identification of intermediate proficiency markers. However, the B2 level presented the most significant challenge, with an F1 score of 0.681, reflecting the inherent difficulty of distinguishing this transitional proficiency level.

5 Discussion and Conclusion

The presented approach demonstrates the effectiveness of Large Language Models (LLMs) in classifying German texts according to CEFR proficiency levels. Through careful prompt engineering and fine-tuning of the LLaMA-3-8B-Instruct model, we achieved a weighted F1 score of 0.769 for the fine-tuned model which represents a substantial improvement over previous state-of-the-art results, such as the 0.702 reported by Caines & Buttery (Caines and But-

Table 6: Performance metrics of the fine-tuned model across CEFR levels.

Class	Precision	Recall	F1 Score
A1	0.875	0.840	0.857
A2	0.750	0.720	0.735
B1	0.724	0.840	0.778
B2	0.727	0.640	0.681
C1	0.636	0.840	0.724
C2	1.000	0.720	0.837
Weighted Avg	0.785	0.767	0.769

ter, 2020). Notably, our model achieved perfect *group accuracy* (100%), demonstrating excellent performance in assigning texts to the correct general CEFR region—that is, either the exact level or a directly adjacent one. This represents a significant improvement over prior work using broader level groupings. For instance, Szüügyi et al. (Szüügyi et al., 2019) reported 82% accuracy using a three-level classification (A, B, C) based on a Linear SVM and manual feature engineering. When applying the evaluation scheme to our results (see Table 4), our fine-tuned model achieves 76.7% accuracy and 100% group accuracy, indicating a notable improvement. Such performance appears particularly valuable in practical applications, where the boundaries between CEFR levels often have to be considered fluid and subjective. The fine-tuned model’s ability to consistently place texts within the correct region of the proficiency spectrum, even if not always at the exact level, demonstrates its reliability as a tool for language assessment support.

A neural network based classifier that was trained on the internal states of the LLaMA3 model showed an improvement compared to pure prompting approaches when applied to the same LLM. The internal state of the model appears to contain further valuable information that support language level classification. Importantly, the fine-tuning approach outperformed the neural network probing classifier which demonstrates that without fine-tuning the LLaMA model does not contain sufficient information for a better performance.

Overall, our results demonstrate several key advances: strong performance in distinguishing extreme proficiency levels (A1 and C2), consistent handling of intermediate levels, and reliable classification of adjacent CEFR levels. These results suggest significant potential for LLMs to support language assessment processes, particularly when combined with human expertise.

As language assessment tools continue to evolve, the integration of LLM-based approaches shows promise in enhancing the efficiency and reliability of the CEFR classification process. Finally, a promising direction lies in leveraging LLMs not only for classification but also for generation of texts on a specific CEFR level or rewriting existing ones to match a target proficiency level. One challenge for this is the limited availability of high-quality, labeled training data, as generating or annotating such texts typically requires the involvement of language assessment experts. The presented classification model could support such approaches as it can serve as a reliable tool for automatically labeling or validating large volumes of generated data. This can facilitate the creation of synthetic datasets that are both diverse and consistently aligned with CEFR standards.

References

- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Calum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>.
- Anthropic. Claude 3.5 sonnet model card and announcement. <https://www.anthropic.com/news/clause-3-5-sonnet>, June 2024. Model card with performance benchmarks and capabilities.
- Amos Azaria and Tom Mitchell. The internal state of an ILM knows when it’s lying. *arXiv preprint arXiv:2304.13734*, 2023.
- Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. The MERLIN corpus: Learner Language and the CEFR. In *LREC*, pages 1281–1288. Reykjavik, Iceland, 2014.
- Andrew Caines and Paula Buttery. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.689>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, ..., and the Llama Team, AI@Meta. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783, July 2024. URL <https://arxiv.org/abs/2407.21783>.

- Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database(EFCamDat). In *Proceedings from 31st Second Language Research Forum*, 2014. URL <https://api.semanticscholar.org/CorpusID:37833484>.
- Goethe Institut. Deutschkurse und Prüfungen. <https://www.goethe.de/de/spr/kur/stu.html>, 2020. Accessed on 2024-09-09.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, June 2021. doi: 10.48550/ARXIV.2106.09685.
- Zhen Huang, Zengzhi Wang, Shijie Xia, and Pengfei Liu. Olympic Arena Medal Ranks: Who Is the Most Intelligent AI So Far? *CoRR, abs/2406.16772*, June 2024. doi: 10.48550/arXiv.2406.16772. URL <https://arxiv.org/abs/2406.16772>.
- Hugging Face. Transformers, 2025. URL <https://github.com/huggingface/transformers>. Accessed on 11.09.2025.
- Sebastian Kahl, Felix Löffler, Martin Maciol, Fabian Ridder, Marius Schmitz, Jennifer Spanagel, Jens Wienkamp, Christopher Burgahn, and Malte Schilling. Evaluating the impact of advanced LLM techniques on AI-lecture tutors for a robotics course. *arXiv preprint arXiv:2408.04645*, 2024.
- Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. Evaluating language models as synthetic data generators, 2024. URL <https://arxiv.org/abs/2412.03679>.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. URL <https://arxiv.org/abs/2402.06196>.
- Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. Multiprageval: Multilingual pragmatic evaluation of large language models, 2024. URL <https://arxiv.org/abs/2406.07736>.
- Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL <https://aclanthology.org/P19-1493/>.
- Marc Reznicek, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas. Das Falko-Handbuch: Korpusaufbau und Annotationen. *Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin*, 2010.
- Fabian Ridder and Malte Schilling. The HalluRAG Dataset: Detecting Closed-Domain Hallucinations in RAG Applications Using an LLM’s Internal States, 2024a.
- Fabian Ridder and Malte Schilling. The hallurag dataset: Detecting closed-domain hallucinations in rag applications using an llm’s internal states. *arXiv preprint arXiv:2412.17056*, 2024b.
- Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models, 2024.
- Edit Szüügyi, Soören Etler, Andrew Beaton, and Manfred Stede. Automated assessment of Language Proficiency on German Data. *KONVENS 2019*, 2019.
- Satoru Uchida, Yuki Arase, and Tomoyuki Kajiwara. Profiling English sentences based on CEFR levels. *ITL-International Journal of Applied Linguistics (Belgium)*, pages 103–126, March 2024. doi: 10.1075/itl.22018.uch.
- Sowmya Vajjala and Taraka Rama. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2018)*, pages 147–153, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0515. URL <https://aclanthology.org/W18-0515/>.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023. URL <https://arxiv.org/abs/2302.11382>.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, pages 2905–2909, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3661370. URL <https://doi.org/10.1145/3626772.3661370>.

A Prompt for Generation of Synthetic Data

German prompt used for synthetic data generation, following definition by (Goethe Institut, 2020):

Bitte generiere Texte mit dem CEFR Niveau A1. Diese sollten länger (Circa 600 Wörter) sein. Versuche Themen zu finden, welche nicht mit Schule/Kindheit in Verbindung stehen. Es sollte sich um Texte für Deutsch Sprachler mit dem Level A1 handeln.

Hier ist eine Definition des A1 Levels:

Kann vertraute, alltägliche Ausdrücke und ganz einfache Sätze verstehen und verwenden, die auf die Befriedigung konkreter Bedürfnisse zielen. Kann sich und andere vorstellen und anderen Leuten Fragen zu ihrer Person stellen - z. B. wo sie wohnen, welche Leute sie kennen oder welche Dinge sie haben - und kann auf Fragen dieser Art Antwort geben. Kann sich auf einfache Art verständigen, wenn die Gesprächspartner langsam und deutlich sprechen und bereit sind zu helfen.

B Learning curves for Fine-Tuning the LLM

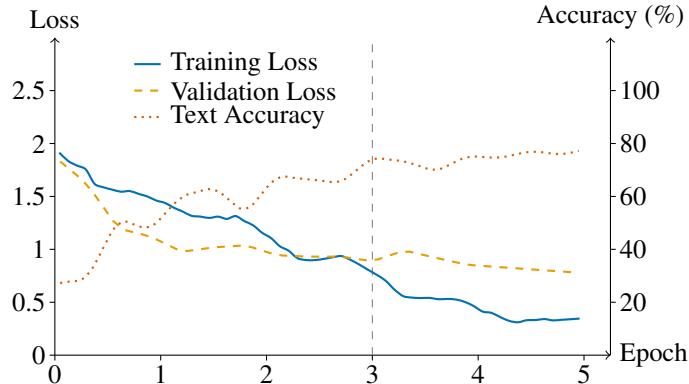


Figure 5: Training loss (blue), validation loss (orange), and text accuracy (red) during training of the model. The vertical line marks the training cutoff used.

C Detailed Performance Metrics for Prompt Engineering Approaches

Table 7: Performance metrics for different prompts (LLaMA-3-8B-Instruct).

English Base Prompt				German Zero-Shot Prompt				German Few-Shot Prompt			
Class	P	R	F1	Class	P	R	F1	Class	P	R	F1
A1	1.000	0.120	0.214	A1	1.000	0.480	0.649	A1	0.833	0.600	0.698
A2	0.000	0.000	0.000	A2	0.579	0.440	0.500	A2	0.600	0.360	0.450
B1	0.255	0.960	0.403	B1	0.303	0.920	0.455	B1	0.471	0.640	0.542
B2	0.154	0.320	0.208	B2	0.093	0.160	0.118	B2	0.568	0.840	0.677
C1	0.000	0.000	0.000	C1	0.000	0.000	0.000	C1	0.546	0.240	0.333
C2	0.000	0.000	0.000	C2	0.000	0.000	0.000	C2	0.629	0.880	0.733
Avg	0.247	0.467	0.286	Avg	0.327	0.500	0.406	Avg	0.609	0.600	0.586

D Prompt Variations for CEFR Classification

Table 8: Overview of prompt variants used for CEFR Classification. Each prompt was tested in isolation to evaluate its impact on model performance and behavior on our data set.

Prompt Name	Original Prompt	English Translation	Accuracy	Group Accuracy
English Base Prompt	Classify the language level of a given text according to the Common European Framework of Reference for Languages (CEFR). Respond with only the corresponding CEFR level (A1, A2, B1, B2, C1 or C2).	Classify the language level of a given text according to the Common European Framework of Reference for Languages (CEFR). Respond with only the corresponding CEFR level (A1, A2, B1, B2, C1 or C2).	23.3%	64.6%
German Zero-Shot Prompt	Bewerte die Sprachkenntnisse des bereitgestellten deutschen Textes gemäß dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER/CEFR). Antworte NUR mit der entsprechenden Stufe: A1, A2, B1, B2, C1 oder C2, *keiner* Begründung.	Assess the language proficiency of the given German text according to the CEFR. Respond ONLY with the corresponding level: A1-C2, no justification.	33.3%	75.3%
German Few-Shot Prompt	Klassifiziere die Sprachkenntnisse des bereitgestellten deutschen Textes gemäß dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER/CEFR). Antworte NUR mit der entsprechenden Stufe: A1, A2, B1, B2, C1 oder C2, NICHT MEHR. Gebe auch *keine* Begründung! Hier sind jeweils Beispiele: A1: [...] A2: [...] B1: [...] B2: [...] C1: [...] C2: [...]	Classify the language proficiency of the following German text according to the CEFR. Respond ONLY with one of the following levels: A1-C2, NOTHING MORE. Do not give any justification! Here are examples for each level: A1: [...] A2: [...] B1: [...] B2: [...] C1: [...] C2: [...]	59.3%	94.0%