Elena Leitner, Georg Rehm

Exploring the Limits of LLMs for German Text Classification: Prompting and Fine-tuning Strategies Across Small and Medium-sized Datasets

Abstract

Large Language Models (LLMs) are highly capable, state-of-the-art technologies and widely used as text classifiers for various NLP tasks, including sentiment analysis, topic classification, legal document analysis, etc. In this paper, we present a systematic analysis of the performance of LLMs as text classifiers using five German datasets from social media across 13 different tasks. We investigate zero- (ZSC) and few-shot classification (FSC) approaches with multiple LLMs and provide a comparative analysis with fine-tuned models based on Llama-3.2, EuroLLM, Teuken and BübleLM. We concentrate on investigating the limits of LLMs and on accurately describing our findings and overall challenges.

1 Introduction

Large Language Models (LLMs) have had a global impact, revolutionising numerous fields and sectors. LLMs leverage very large datasets and advanced architectures, resulting in the ability to process even complex linguistic phenomena. It is generally accepted that they have improved accuracy compared to previous smaller Language Models (LMs), leading to more effective and efficient solutions for various NLP tasks. LLMs have been adopted as text classifiers that demonstrate competitive performance using zero- and few-shot strategies and also fine-tuning, e. g., in English (Pan, García-Díaz, & Valencia-García, 2024; Wang, Pang, Lin, & Zhu, 2024). For German, Münker, Kugler, and Rettinger (2024) test LLMs in an annotation task on Twitter data and report results that are comparable with BERT. Many other studies on social media data demonstrate strong capabilities of LLMs to identify hate speech and offensive language (Bauer, Preisig, & Volk, 2024; He et al., 2024; Zampieri, Rosenthal, Nakov, Dmonte, & Ranasinghe, 2023).

How well do LLMs work when it comes to the German language? To address this question, we investigate the performance of several LLMs on five German datasets using different prompting as well as fine-tuning techniques. Our goal is to find the best solution for binary or multi-class classification with a minimal number of examples with unbalanced class distribution in the data, as well as to investigate why experiments failed and how to improve performance. This includes the following tasks:

- Analysing failures and successes for zero-shot and few-shot prompting approaches as well as for fine-tuning with selected LLMs;
- Analysing the performance of LLMs on selected datasets with regard to the size and distribution of classes;
- Analysing model limitations on a specific linguistic phenomenon presented in selected data (e. g., toxic, offensive or hateful language in social media) as well as in general for the German language.

Our code, results as well as a detailed description of the conducted experiments are available on GitHub.[1]

2 Experiments

Learning Approaches To answer the question which approach is better suited for text classification depending on the size of data and number of classes, we utilised zero- and 8-shot prompting as well as parameter-efficient fine-tuning using QLoRa (Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023). Figure 1 shows a prompt that includes an instruction to classify the text, definitions of classes, and a question. The question for a task was formulated simply, e. g., "Does the text contain any form of offensive language?" Since the LLM does not know which classes we assume to exist, we added definitions. In FSC, for each label, eight random examples were inserted; we used a fixed random seed to ensure reproducibility. To avoid a detailed answer and to get only a category name, we instructed the LLM to answer with one word and to use a category from the list as an answer. Since prompting in the 'native' language enhances LLM comprehension (He et al., 2024), the prompt was formulated in German.

> Please classify the text. The categories are defined as follows: {DEFINITIONS}
>
> Here are a couple of examples of categories assigned by experts: {EXAMPLES}*
>
> {QUESTION} Please answer with one word and use a category from this list as an answer: {CATEGORIES}
>
> Text: {SENTENCE}
>
> Answer: {ANSWER}

Figure 1: Prompting template (translated to English)

For the fine-tuning experiments, we utilised the available train and test sets. We also created validation sets using examples from the train sets for hyperparameter tuning. Due to the different sizes of each dataset, the different numbers and also distributions of classes, we used several hyperparameters to improve performance. However, we observed that when good results were achieved in some tasks, overfitting occurred on

[1] https://github.com/elenanereiss/Limits-of-LLMs-for-German-Text-Classification

other tasks with the same hyperparameters. Thus, to avoid overfitting but to evaluate the models on equal terms, we applied the early stopping technique and set other hyperparameters to default.

Datasets We focus on five small and medium-sized datasets from social media (covering a total of 13 tasks) with different granularities of annotations and unbalanced class distributions, developed for the German language (see Table 1). A detailed overview of the tasks can be found in Appendix A.

Dataset	Citation	Tasks	Size
German COVID-19 Twitter	[submitted]	informativeness, topic, credibility	643
German Speech Acts	Plakidis and Rehm (2022)	coarse and fine-grained classification	1959
HASOC 2020	Mandl, Modha, Kumar M, and Chakravarthi (2021)	coarse and fine-grained classification	2899
GermEval 2019	Struß, Siegel, Ruppenhofer, Wiegand, and Klenner (2019)	coarse and fine-grained classification	7026
		implicit/explicit offensive language	2888
GermEval 2021	Risch, Stoll, Wilms, and Wiegand (2021)	toxic, engaging, fact claiming	4188

Table 1: Overview of German datasets.

Models We use recent non-European, European and German LLMs such as multilingual Llama 3.2-3B (Grattafiori et al., 2024), European EuroLLM-9B (Martins et al., 2024) and Teuken-7B (Ali et al., 2024) as well as German BübleLM. For prompting we use instruction-tuned models and for fine-tuning the pre-trained base models. Further details can be found in Appendix B.

Statement on Possible Data Contamination We would like to state explicitly that there is a lack of information regarding the training data of the LLMs we experiment with. Their training data may contain training and test sets from the datasets selected in our evaluation study. As reported by Balloccu, Schmidtová, Lango, and Dusek (2024); Samuel, Zhou, and Zou (2025), when data contamination occurs, through memorisation instead of true generalisation, it can lead to inflated evaluation scores.

3 Findings and Challenges

We conducted 78 prompting and 52 fine-tuning experiments in total. All major challenges we faced occurred during prompting. The first challenge was to ensure that an instruction-tuned LLM returns only a class for a given text. During ZSC, in many cases, we received one or more sentences with an explanation of the class. We tried several variants of the prompts; it worked well when we explicitly instructed the LLM to respond with one word and use a class from the list for its answer. Teuken has shown the best performance – on average 99% of answers were one word. With Llama 3.2 we got about 96% and with EuroLLM 79%. During FSC, the rate changed (Teuken: 99.9%, EuroLLM: 91%, Llama 3.2: 85.2%).

We collected the answers from the LLMs and systematised the limitations. In general, we observe the following types of output:

- The answer contained a valid class label: (i) as a word, (ii) in one or more sentences, (iii) translated into German as a word or in one or more sentences (mostly by Teuken for the classes "OTHER", "OFFENSE", "Risk_Reduction" and "Case_Report").
- While the text was classified, no class label was provided (happened often during FSC with LLama 3.2. – "YES", "NO", etc.).
- The text was not classified: (i) due to a lack of context or (ii) due to offensive content (e. g., using Llama 3.2 on all datasets except COVID-19 Twitter).
- Hallucinations that were (i) similar to a predefined class, e. g., "OPFN" instead of "OFFN" or "GGovernm_Decisions" instead of "Governm_Decisions" (this behaviour was mostly observed with EuroLLM), or (ii) random words ("WHO", "Zombies", etc.)

Due to the number of tasks, the second challenge was to filter out class labels from sentences. To get a valid predicted label, we tokenised each output and compared each token with the predefined classes. In cases where we found multiple valid class labels in an answer, we were unable to assign one class automatically, and left these answers unchanged; German translations were mapped to corresponding classes.

Some of our experiments failed technically. Originally, we planned to also test the German LLM LLäMmlein (Pfister, Wunderle, & Hotho, 2024). Unfortunately, we were unable to get an answer from various instruction-tuned models in the form of a class during prompting. Due to this limitation, it was not possible to manually edit each output and filter out a category. Fine-tuning a base model also failed. For LLäMmlein, we got an error message during the initialisation of the tokenizer that we have not been able to fix, which is why we decided to exclude this LLM from our experiments.

Experimenting with n-shot prompting, we found that Teuken and EuroLLM were already working to capacity at 10-shot. EuroLLM began to hallucinate when the maximum input length was exceeded. This is why we reduced the number of examples in FSC to 8, i. e., many-shot classification with 100 examples was not tested. However, we have done some test runs with Llama 3.2, which allows 128,000 input tokens. Already with the first tasks, we noticed that all metrics decreased. On the HASOC 2020 and GermEval 2019 datasets, the F_1-scores were even worse than in ZSC, i. e., around 0.18-0.29 points. We see this as evidence that the use of more examples does not necessarily result in better performance.

4 Results and Discussion

The results of the prompting and fine-tuning experiments are shown in Table 2. The fine-tuned LLMs achieved the best results in all tasks. A big difference between prompting and fine-tuning can be found in the tasks with fine-grained classes. F_1-scores doubled for almost all LLMs. The F_1-scores on the Speech Acts Dataset (coarse) using Llama

	Dataset and Task	A	Llama 3.2			EuroLLM			Teuken			BübleLM		
			p	r	f1	p	r	f1	p	r	f1	p	r	f1
COVID-19 Twitter		zs	.84	.61	.59	.53	.40	.39	.67	.53	.49	–	–	–
	informativeness	fs	.50	.57	.52	.74	.60	.64	.68	.63	.65	–	–	–
	3 classes	t	.70	.73	.70	.77	.79	**.78**	.65	.68	.65	.73	.72	.72
		zs	.23	.23	.18	.31	.28	.26	.46	.37	.29	–	–	–
	topic	fs	.51	.44	.40	.24	.26	.18	.62	.38	.31	–	–	–
	6 classes	t	.71	.66	**.67**	.52	.53	.50	.54	.61	.56	.61	.57	.58
		zs	.41	.45	.18	.34	.18	.23	.41	.39	.38	–	–	–
	credibility	fs	.54	.36	.41	.45	.63	.39	.56	.51	.44	–	–	–
	3 classes	t	.50	.52	.50	.54	.56	**.54**	.47	.48	.47	.51	.52	.51
Speech Acts		zs	.31	.22	.15	.20	.24	.18	.25	.19	.12	–	–	–
	coarse	fs	.42	.32	.29	.21	.22	.18	.19	.24	.19	–	–	–
	6 classes	t	.64	.67	**.65**	.69	.56	.59	.58	.60	.56	.48	.54	.49
		zs	.13	.11	.10	.15	.13	.12	.03	.11	.03	–	–	–
	fine	fs	.28	.20	.17	.10	.14	.08	.11	.17	.10	–	–	–
	17 classes	t	.33	.38	.34	.34	.32	.31	.40	.43	**.39**	.27	.31	.28
HASOC 2020		zs	.66	.62	.59	.60	.62	.53	.63	.51	.22	–	–	–
	coarse	fs	.70	.49	.55	.63	.67	.63	.63	.52	.24	–	–	–
	2 classes	t	.76	.81	.78	.75	.79	.76	.78	.81	**.79**	.78	.78	.78
		zs	.38	.38	.29	.33	.33	.21	.26	.27	.07	–	–	–
	fine	fs	.39	.30	.29	.36	.32	.24	.23	.28	.25	–	–	–
	4 classes	t	.49	.58	.50	.48	.56	.51	.46	.59	.48	.49	.54	**.51**
GermEval 2019		zs	.66	.62	.56	.65	.65	.57	.16	.50	.24	–	–	–
	coarse	fs	.65	.50	.34	.68	.70	.66	.16	.50	.24	–	–	–
	2 classes	t	.76	.77	.76	.76	.78	**.76**	.73	.76	.74	.73	.76	.74
		zs	.36	.37	.36	.33	.34	.23	.29	.32	.21	–	–	–
	fine	fs	.41	.26	.28	.33	.31	.23	.40	.29	.11	–	–	–
	4 classes	t	.42	.45	.42	.40	.45	.40	.44	.48	**.44**	.40	.46	.41
		zs	.54	.53	.26	.51	.48	.28	.43	.50	.46	–	–	–
	offensive	fs	.57	.51	.14	.54	.52	.37	.60	.50	.47	–	–	–
	2 classes	t	.68	.76	**.70**	.68	.71	.69	.67	.74	.69	.65	.73	.67
GermEval 2021		zs	.61	.60	.60	.57	.48	.40	.62	.53	.37	–	–	–
	toxic	fs	.59	.60	.59	.55	.54	.54	.62	.56	.44	–	–	–
	2 classes	t	.67	.68	.68	.70	.69	.69	.70	.71	**.70**	.68	.66	.67
		zs	.55	.54	.53	.56	.49	.46	.56	.52	.33	–	–	–
	engaging	fs	.51	.41	.43	.53	.52	.52	.59	.55	.35	–	–	–
	2 classes	t	.66	.68	.67	.66	.68	.67	.64	.66	.64	.67	.68	**.67**
		zs	.63	.61	.53	.57	.48	.34	.67	.50	.26	–	–	–
	factClaiming	fs	.61	.62	.59	.58	.57	.58	.60	.55	.39	–	–	–
	2 classes	t	.77	.77	**.77**	.75	.76	.76	.72	.72	.72	.72	.71	.71

Table 2: Precision, recall and macro F_1-score for zero-shot (zs), few-shot (fs) classification and fine-tuning (t) on the test set. The best F_1-scores are marked as follows: in one approach underlined, in both prompting approaches underlined twice, and in all approaches in **bold**.

3.2 rose drastically from 0.31 in FSC to 0.65 after fine-tuning. For binary classification, the values also improved, at least 0.1 points.

Comparing the prompting approaches, the LLMs show better performance in FSC than in ZSC. Only in four tasks, Llama 3.2 was better in ZSC. The instruction-tuned models were surprisingly good at the identification of offensive and toxic content (binary classification), scoring around 0.6 F_1 (already in ZSC) on the HASOC 2020, GermEval 2019 and 2021 datasets. However, when the number of classes increases to 4 (i. e., in the fine-grained tasks), the LLMs fail, and the F_1-scores are in the range of only 0.25-0.35.

Regarding the small-sized datasets, unexpectedly, the fine-tuned LLMs exhibit solid performance on the classification of informativeness and topic (COVID-19 Twitter Dataset) and of coarse-grained speech acts. However, for the identification of credibility (3 classes), we expected better results. As anticipated, the LLMs performed worst at the classification of 17 highly unbalanced fine-grained speech acts. Regarding the medium-sized datasets, the results with fine-grained classes must be described as moderate. Even the fine-tuned LLMs only reached a maximum of 0.51 F_1 on HASOC 2020 and 0.44 F_1 on GermEval 2019.

As far as the LLMs are concerned, it is impossible to generalise which of the models is superior. Depending on the task and approach, some LLMs provide comparable results, such as EuroLLM and Teuken at topic classification with zero-shot prompting or at informativeness classification with 8-shot. In some tasks, the differences are enormous and reached a gap of almost 0.2 F_1. In ZSC, Teuken had 0.38 F_1 on the COVID-19 Twitter dataset (credibility) and 0.46 F_1 on GermEval 2019 (offensive). In FSC on GermEval 2019 (coarse), EuroLLM achieved 0.66 F_1. In ZSC on GermEval 2021, Llama 3.2 had 0.6 F_1 (toxic) and 0.53 F_1 (fact claiming). However, as we can see from Table 2, Llama 3.2 often scored the best F_1 depending on the task and approach.

We can draw the following conclusions from the experiments and evaluation:

- Fine-tuning outperforms prompting and is better suited for small- and medium-sized datasets with fine-grained annotations.
- Prompting achieves good results when a task is well-known and defined as binary classification.
- Prompting with the use of examples exhibits better performance than zero-shot.
- Apart from the chosen approach, the LLMs fail on small-sized datasets with fine-grained annotations with only a few examples per class.

5 Conclusion

Across prompting and fine-tuning approaches, LLMs exhibit satisfactory performance as text classifiers for German. The scores decrease rapidly as the number of labels increases. The fine-tuned LLMs significantly outperform the instruction-tuned LLMs in a zero- and 8-shot prompting approach. Moreover, the instruction-tuned LLMs exhibit certain limitations and are challenging to use.

References

Ali, M., Fromm, M., Thellmann, K., Ebert, J., Weber, A. A., Rutmann, R., … Flores-Herr, N. (2024). *Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs.* Retrieved from `https://arxiv.org/abs/2410.03730`

Balloccu, S., Schmidtová, P., Lango, M., & Dusek, O. (2024, March). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the*

European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 67–93). St. Julian's, Malta: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2024.eacl-long.5/`

Bauer, N., Preisig, M., & Volk, M. (2024). Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets. In *Workshop on Trolling, Aggression and Cyberbullying.* Retrieved from `https://api.semanticscholar.org/CorpusID:269950799`

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs.* Retrieved from `https://arxiv.org/abs/2305.14314`

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024). *The Llama 3 Herd of Models.* eprint arXiv:2407.21783. Retrieved from `https://arxiv.org/abs/2407.21783`

He, J., Wang, L., Wang, J., Liu, Z., Na, H., Wang, Z., ... Chen, Q. (2024). Guardians of Discourse: Evaluating LLMs on Multilingual Offensive Language Detection. *ArXiv*, *abs/2410.15623*. Retrieved from `https://api.semanticscholar.org/CorpusID:273501874`

Mandl, T., Modha, S., Kumar M, A., & Chakravarthi, B. R. (2021). Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation* (p. 29–32). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3441501.3441517` doi: 10.1145/3441501.3441517

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., ... Martins, A. F. T. (2024). *EuroLLM: Multilingual Language Models for Europe.* Retrieved from `https://arxiv.org/abs/2409.16235`

Münker, S., Kugler, K., & Rettinger, A. (2024). Zero-shot Prompt-based Classification: Topic Labeling in Times of Foundation Models in German Tweets. *ArXiv*, *abs/2406.18239*. Retrieved from `https://api.semanticscholar.org/CorpusID:270737586`

Pan, R., García-Díaz, J. A., & Valencia-García, R. (2024). Comparing Fine-Tuning, Zero and Few-Shot Strategies with Large Language Models in Hate Speech Detection in English. *Computer Modeling in Engineering & Sciences.* Retrieved from `https://api.semanticscholar.org/CorpusID:269943133`

Pfister, J., Wunderle, J., & Hotho, A. (2024). *LLäMmlein: Compact and Competitive German-Only Language Models from Scratch.* Retrieved from `https://arxiv.org/abs/2411.11171`

Plakidis, M., Leitner, E., & Rehm, G. (2025). Automated Speech Act Classification in Offensive German Language Tweets. *Traitement Automatique des Langues*, *65*(3). (Special Issue on Abusive Language Detection)

Plakidis, M., & Rehm, G. (2022, June). A Dataset of Offensive German Language Tweets Annotated for Speech Acts. In N. Calzolari et al. (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4799–4807).

Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.513/

Remy, F., Delobelle, P., Avetisyan, H., Khabibullina, A., de Lhoneux, M., & Demeester, T. (2024). *Trans-Tokenization and Cross-lingual Vocabulary Transfers: Language Adaptation of LLMs for Low-Resource NLP.* Retrieved from https://arxiv.org/abs/2408.04303

Risch, J., Stoll, A., Wilms, L., & Wiegand, M. (2021, September). Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In J. Risch, A. Stoll, L. Wilms, & M. Wiegand (Eds.), *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments* (pp. 1–12). Duesseldorf, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.germeval-1.1/

Samuel, V., Zhou, Y., & Zou, H. P. (2025, January). Towards Data Contamination Detection for Modern Large Language Models: Limitations, Inconsistencies, and Oracle Challenges. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5058–5070). Abu Dhabi, UAE: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.coling-main.338/

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Conference on Natural Language Processing.* Retrieved from https://api.semanticscholar.org/CorpusID:208334401

Wang, Z., Pang, Y., Lin, Y., & Zhu, X. (2024). *Adaptable and Reliable Text Classification using Large Language Models.* Retrieved from https://arxiv.org/abs/2405.10523

Zampieri, M., Rosenthal, S., Nakov, P., Dmonte, A. E., & Ranasinghe, T. (2023). OffensEval 2023: Offensive language identification in the age of Large Language Models. *Natural Language Engineering*, *29*, 1416 - 1435. Retrieved from https://api.semanticscholar.org/CorpusID:265659717

A Task Overview

To evaluate the performance of LLMs, we selected five German datasets comprising 13 tasks. Table 3 lists the tasks and their definitions as well as illustrates the number of classes in a task and the minimum number of instances per task.

The German COVID-19 Twitter Dataset[2] is a novel credibility dataset consisting of 643 COVID-19-related texts extracted during the pandemic. Credibility is framed as informative and relevant content regarding a predefined set of topics and therefore each

[2] Due to X's content redistribution policy, the dataset is not published. A paper on the dataset is currently under review.

Dataset	Task	Identification and classification of	No.	Min.
German COVID-19 Twitter Dataset	*Informativeness*	Informative content related to COVID-19	3	55
	Topic	Topic-related content	6	15
	Credibility	Credible content related to COVID-19	3	10
German Speech Acts Dataset	*Coarse*	Coarse-grained speech acts	6	20
	Fine	Fine-grained speech acts	17	11
HASOC 2020	*Coarse*	Hate, offensive and profane content	2	907
	Fine	Hate, offensive and profane content	4	170
GermEval 2019	*Coarse*	Offensive language	2	2257
	Fine	Offensive language	4	263
	Offensive	Explicit and implicit offensive language	2	393
GermEval 2021	*Toxic*	Toxic comments on Facebook	2	1472
	Engaging	Engaging comments on Facebook	2	1118
	Fact claiming	Fact-claiming comments on Facebook	2	1417

Table 3: Short description of the tasks. "No." means the number of classes in a task, "Min." means minimum number of instances per class in a dataset.

tweet is annotated for informativeness, topic and credibility. In the informativeness task, texts are classified into informative (*informative*), non-informative (*none*), and tweets that report personal experience (*personal_experience*). In the topic task, main COVID-19-related topics are *case report*, *consequences*, *governmental decisions*, *risk reduction*, and *vaccination*. Tweets that are not topic-related are marked as *none*. In the credibility task, tweets that have high or low credibility are classified as *credible* or *non-credible*. If it is not possible to decide from the text whether the content is credible or not, tweets are assigned the class *none*.

In the German Speech Acts Dataset[3] (Plakidis & Rehm, 2022), 1,959 sentences are annotated for six coarse- and 23 fine-grained speech acts. In the coarse-grained task, sentences shall be classified into following classes: *assertive*, *directive*, *expressive*, *commissive*, *unsure* and *other*. In the fine-grained task, assertive, directive, expressive, and commissive speech acts are split into fine-grained ones. Similarly to Plakidis, Leitner, and Rehm (2025), due to sparse occurrences in the dataset, we modified a few fine-grained classes reducing the number of classes from 23 to 17.

The HASOC 2020 Dataset for German[4] (Mandl et al., 2021) consists of 2,899 tweets including binary and fine-grained annotations regarding the classification of hate-offensiveness. In the coarse-grained task, the goal is to identify hate, offensive and profane content and classify tweets into two classes: hate and offensive (HOF) or non hate-offensive (NOT). In the fine-grained task, a distinction is made between texts that contain hate speech ($HATE$), offensive content ($OFFN$), profane words ($PRFN$) and texts that do not contain hate speech, profane, offensive content (NOT).

[3] `https://github.com/MelinaPl/speech-act-analysis`

[4] `https://hasocfire.github.io/hasoc/2020/call_for_participation.html`

The GermEval 2019 Dataset[5] (Struß et al., 2019) originates from a shared task on the identification of offensive language. As for HASOC 2020, the first task deals with the coarse-grained binary classification of offensive language (*OFFENSE* and *OTHER*), and the second task – with the fine-grained classification containing four classes (*PROFANITY*, *INSULT*, *ABUSE*, *OTHER*). The third task focuses on the classification of explicit and implicit offensive language using the classes *EXPLICIT* and *IMPLICIT*.

The GermEval 2021 Dataset[6] (Risch et al., 2021) consists of 4,188 Facebook posts and addresses three classification problems. The first task deals with the classification of toxic comments. The second task on the engaging comment classification focuses on rational, respectful, and reciprocal comments. Due to the spread of misinformation and fake news, the third task is dedicated to the classification of fact-claiming comments and conceived as a pre-processing step for manual fact-checking. All three tasks belong to binary classification and are marked with *1* and *0*.

B Models Overview

In our experiments, we utilise several recently released LLMs such as Llama 3.2, EuroLLM, Teuken, LLäMmlein and BübleLM. Meta Llama 3.2-3B is a smaller and more efficient version of the Llama3 family (Grattafiori et al., 2024) trained on approx. 9 trillion tokens from publicly available online data. EuroLLM-9B (Martins et al., 2024) is an open-weight multilingual LLM trained on 4 trillion tokens divided across official European Union languages (and several additional languages). Teuken-7B (Ali et al., 2024) is also a European LLM developed by the OpenGPT-X project. It is trained on 4 trillion tokens where 60% of data is non-English (8.72% data is German). LLäMmlein (Pfister et al., 2024) is a German Tinyllama LM trained on only high-quality German data from RedPajama V2. The last model, BübleLM, is a small German LM based on Gemma-2-2B and trained on 3.5B tokens from the Occiglot-FineWeb project. The model is characterized by using trans-tokenization – a cross-lingual vocabulary transfer strategy – for language adaptation of LLMs (Remy et al., 2024).

The instruction-tuned LLMs used in prompting experiments are as follows:

- Llama-3.2-3B-Instruct[7]
- EuroLLM-9B-Instruct[8]
- Teuken-7B-instruct-research-v0.4[9]
- *several LLäMmlein chat models[10]

The pre-trained base LLMs used in fine-tuning experiments are as follows:

[5] https://fz.h-da.de/iggsa/

[6] https://germeval2021toxic.github.io/SharedTask/

[7] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[8] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[9] https://huggingface.co/openGPT-X/Teuken-7B-instruct-research-v0.4

[10] https://huggingface.co/collections/LSX-UniWue/llammlein-chat-preview-6734b15176c7f079f72a9291

- Llama-3.2-3B[11]
- EuroLLM-9B[12]
- Teuken-7B-base-v0.6[13]
- *LLaMmlein_1B[14]
- bueble-lm-2b[15]

As reported in Section 3, the experiments with various instruction-tuned LLäMmlein models, as well as with the pre-trained base LLäMmlein model failed. Therefore, these LLMs are marked with * in both lists. BübleLM has no instruction-tuned version and was excluded from the prompting experiments.

[11] https://huggingface.co/meta-llama/Llama-3.2-3B

[12] https://huggingface.co/utter-project/EuroLLM-9B

[13] The model is available upon request.

[14] https://huggingface.co/LSX-UniWue/LLaMmlein_1B

[15] https://huggingface.co/flair/bueble-lm-2b

Correspondence

Elena Leitner

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Berlin, Germany
elena.leitner@dfki.de

Georg Rehm

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Berlin, Germany
georg.rehm@dfki.de