GeistBERT: Breathing Life into German NLP

Raphael Scheible-Schmitt[1,2] and Johann Frei[3]

[1]School of Computation, Information and Technology, Technical University of Munich,
[2]IS[2]E - Intelligent Systems, Science and Engineering, LIACC polo on Azores University, Ponta Delgada, Portugal,
[3]Chair of IT Infrastructure for Translational Medical Research,
Faculty of Applied Computer Science, University of Augsburg

Correspondence: raphael.scheible@tum.de

Abstract

Advances in transformer-based language models have highlighted the benefits of language-specific pre-training on high-quality corpora. In this context, German NLP stands to gain from updated architectures and modern datasets tailored to the linguistic characteristics of the German language. GeistBERT seeks to improve German language processing by incrementally training on a diverse corpus and optimizing model performance across various NLP tasks. We pre-trained GeistBERT using fairseq, following the RoBERTa base configuration with Whole Word Masking (WWM), and initialized from GottBERT weights. The model was trained on a 1.3 TB German corpus with dynamic masking and a fixed sequence length of 512 tokens. For evaluation, we fine-tuned the model on standard downstream tasks, including NER (CoNLL 2003, GermEval 2014), text classification (GermEval 2018 coarse/fine, 10kGNAD), and NLI (German XNLI), using F_1 score and accuracy as evaluation metrics. GeistBERT achieved strong results across all tasks, leading among base models and setting a new state-of-the-art (SOTA) in GermEval 2018 fine text classification. It also outperformed several larger models, particularly in classification benchmarks. To support research in German NLP, we release GeistBERT under the MIT license.

1 Introduction

The advancement of neural language modeling (LM) in natural language processing (NLP) has been driven by the development of contextual pre-trained word representations, particularly through transformer-based architectures. Models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have significantly impacted the field by providing robust, generalized representations that can be fine-tuned for specific downstream tasks, enhancing performance across various NLP applications. While much of the early work focused on English and multilingual models, it has become clear that single-language models, particularly those trained on large, high-quality corpora, can outperform their multilingual counterparts when applied to their target language.

Building on this understanding, the German NLP community has seen the introduction of models like GottBERT (Scheible et al., 2024), which leveraged the German portion of the OSCAR (Ortiz Suárez et al., 2020) corpus to create a high-performance RoBERTa-based (Liu et al., 2019) model tailored specifically for the German language. However, as the field evolves, so too must the approaches to model training. Recent developments in pre-training methodologies, such as Whole Word Masking (WWM) (Cui et al., 2021) and the availability of newer, more extensive corpora like OSCAR23 (Jansen et al., 2022), OPUS (Tiedemann, 2012), and mC4 (Xue et al., 2021), present opportunities to further refine and enhance German language models.

To fully leverage these developments for German NLP, we introduce GeistBERT, a German Enhanced Incremental Semantically Tuned BERT model. GeistBERT builds on the foundation laid by the best checkpoint of the filtered GottBERT model (i.e. $^{\text{f}}\text{GottBERT}_{base}$) through continued pre-training (Gururangan et al., 2020a), extending it with modern German datasets including OSCAR23 and mC4 from CulturaX (Nguyen et al., 2023), Wikipedia, and several OPUS corpora. Since CulturaX already applies both deduplication and filtering, it provides a strong backbone of high-quality German text, while the additional corpora enrich the model with broader linguistic and domain diversity. By introducing Whole Word Masking (WWM) and leveraging the scale and variety of these sources, GeistBERT seeks to establish a new benchmark for German language models, with

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 42–50, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-006

improved performance across various NLP tasks.

Our contributions are as follows:

- We incrementally trained GeistBERT on top of $^{f}\mathrm{GottBERT}_{base}$ using a combination of modern German corpora (OSCAR23, OPUS, mC4), OpenLegal and Wikipedia.
- We integrated WWM into the pre-training process to enhance the model's ability to capture semantic relationships within the German language.
- We provide GeistBERT as base model to the community, accessible under an open-source license for further usage.

GeistBERT represents a step forward in the development of German-specific transformer models, offering enhanced capabilities through modern training techniques and high-varying data.

2 Related Work

The rise of transformer-based models like BERT (Devlin et al., 2019) marked a major shift in NLP, enabling significant performance improvements. Originally introduced as an English model and later as a multilingual version (mBERT), BERT's success led to monolingual adaptations tailored to specific languages. For German, models like GermanBERT[1] and dbmdz BERT[2] emerged, trained on datasets of 12GB–16GB, sourced from Wikipedia, news articles, and legal texts.

RoBERTa enhanced BERT by training on a larger 160GB corpus, optimizing the architecture, and removing next sentence prediction. This strategy was applied to other languages, resulting in models like CamemBERT (Martin et al., 2020) for French and RobBERT (Delobelle et al., 2020) for Dutch, highlighting the benefits of large, diverse training corpora and the use of language-specific vocabularies.

In German NLP, GBERT and GELECTRA (Chan et al., 2020) built on this progress by training on 145GB of the OSCAR corpus (Ortiz Suárez et al., 2020) and additional sources, surpassing earlier German BERT models. These advancements underscored the impact of larger, well-curated datasets on model performance. GottBERT further extended this development as one of the first German RoBERTa models, trained on the German OSCAR corpus. Its results demonstrated the importance of data diversity but also noted that excessive data cleaning might reduce corpus variance and affect downstream performance. GeistBERT refines this lineage by increasing data variance, optimizing pre-training strategies, and achieving strong performance without increasing model size, making it a robust and accessible model for German NLP.

3 Methodology

3.1 Training Data and Pre-training

Compared to GottBERT, GeistBERT was trained on a substantially larger corpus, totaling approximately 1.3TB of text data. Training data was shuffled to support uniform sampling and minimize order effects during pre-training. GeistBERT was pre-trained using the same byte-level BPE tokenizer as GottBERT, following the GPT-2 design with a vocabulary size of 52k. While the tokenizer architecture mirrors GPT-2, the vocabulary itself was trained from scratch on German text. fairseq (Ott et al., 2019) was employed to compute the binary format for pre-training. Unlike GottBERT's TPU-based setup, which processed text as a continuous stream, GeistBERT's GPU training respected natural sentence boundaries. This preserves linguistic structure during pre-training and avoids cutting sequences in the middle of sentences.

Using fairseq, we pre-trained the GeistBERT model on a highly variant corpus consisting of 1.3TB plain text data on 8 NVIDIA A40 GPUs. The model was trained with the RoBERTa base architecture for 100k update steps using a batch size of 8k, initializing the weights with $^{f}\mathrm{GottBERT}_{base}$. We largely adhered to RoBERTa's default training configuration (Liu et al., 2019), including dynamic masking, optimizer settings, and fixed sequence lengths (512 tokens). A 10k iteration warmup was applied, gradually increasing the learning rate to a peak of 0.0007, followed by a polynomial decay to zero.

3.2 Downstream Tasks

We fine-tuned pre-trained BERT models using Huggingface (Wolf et al., 2019) scripts, optimizing batch size and learning rate via grid search. NER and classification (CLS) tasks were trained for up to 30 epochs, while NLI tasks ran for up to 10 epochs using fairseq-adapted hyperparameters. Each task was executed 24 times with varied hyperparam-

[1] https://www.deepset.ai/german-bert

[2] https://huggingface.co/dbmdz/bert-base-german-uncased

Table 1: Overview of datasets used for training. The table lists the individual corpora, their sizes in gigabytes, their data sources, and whether they were deduplicated or filtered. The final corpus aggregates all listed datasets, resulting in approximately 1.3 TB of training data.

Corpus	Documents	Size (GB)	Data Source	Deduplicated	Filtered
mC4 & OSCAR23	6,064,736,930	1316.57	CulturaX	Yes	Yes
ELRC-4244, ELRC-4240, ELRC-4258, ELRC-4217, ELRC-4189, ELRC-4171, ELRC-4149	14,919,003	2.34	OPUS	Yes	No
ECB	1,732,472	0.29	OPUS	No	No
EUbookshop	18,203,612	2.34	OPUS	No	No
Europarl	2,234,583	0.36	OPUS	No	No
EuroPat	19,387,517	3.52	OPUS	No	No
OpenSubtitles	41,612,280	1.35	OPUS	No	No
TildeMODEL	5,059,688	0.79	OPUS	No	No
German Wikipedia	4,767,776	7.23	Wikipedia	No	No
OpenLegalData	209,526	2.48	OpenLegal	No	No
Final corpus	**6,172,863,387**	**1337.28**			

eters, selecting the best checkpoint based on the highest F_1 score (accuracy for NLI). Performance was evaluated analogously to Scheible et al. (2024) and compared with results from that study. The parameter search space used for the grid search is summarized in Table 2. All tasks were processed using two Nvidia RTX 3090 GPUs, leveraging Huggingface's Transformers library (v4.34.1).

Table 2: Hyperparameters used in the grid search of the downstream tasks.

Parameter	Values
Learning Rate	5e-5, 2e-5, 1e-5, 7e-6, 5e-6, 1e-6
Batch Size	16, 32, 48, 64
Epochs	30

NLI We evaluated NLI on the German XNLI dataset (Conneau et al., 2018), an extension of MultiNLI (Williams et al., 2018), with 122k training, 2490 development, and 5010 test examples per language. Performance was measured by accuracy.

Named Entity Recognition NER evaluation used the German CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) and GermEval 2014 (Benikova et al., 2014) datasets. CoNLL 2003 includes four entity types, while GermEval 2014 provides fine-grained categories and supports nested annotations. Both were evaluated using the F_1 score, with GermEval using an adapted metric accounting for label and span equality.

Text Classification We evaluated classification on GermEval 2018 (Risch et al., 2018) (German tweet sentiment analysis) and 10kGNAD (Schabus et al., 2017) (German news categorization). GermEval 2018 followed the data splits defined by Chan et al. (2020), while 10kGNAD used a predefined 90%-10% train–test split, with 10% of the training set further held out for validation. Both tasks were evaluated using the mean F_1 score.

3.3 Model Properties

Table 3 lists the vocabulary sizes and total parameter counts of all models included in our evaluation. While most German BERT-style base models, such as $\text{GBERT}_{\text{base}}$, dbmdzBERT, and $\text{GELECTRA}_{\text{base}}$, contain approximately 110 million parameters, GeistBERT and $^{\text{f}}\text{GottBERT}_{base}$ are slightly larger at around 126 million parameters due to their RoBERTa-based architecture and a larger vocabulary of 52,009 tokens.

Large-scale German models such as $\text{GBERT}_{\text{large}}$, $\text{GELECTRA}_{\text{large}}$, and $^{\text{f}}\text{GottBERT}_{large}$ contain between 335 and 357 million parameters. Among the multilingual models, $\text{XLM-RoBERTa}_{\text{base}}$ and $\text{XLM-RoBERTa}_{\text{large}}$ are substantially larger, with 278 million and 560 million parameters respectively. The vocabulary sizes vary across models and are influenced by tokenizer design and pretraining data. GeistBERT uses the same tokenizer as GottBERT, which is based on byte-level BPE trained on German text.

4 Results

4.1 Training Dynamics

During the model pre-training the perplexity of the model is computed based on a test set for each optimization cycle (see Figure 1). After an initial sharp decrease, perplexity briefly increased for several steps before gradually declining until the final step. We assume that, given more training time, it would have continued to decrease further. The entire pre-

Table 3: The size of the vocabulary and the size of the parameters are shown for the model types used in this study. This table does not show other design differences of the models. Values were extracted using Huggingface's transformers library.

Model	Vocab Size	#Params
XLM-R$_{\text{large}}$	250002	559890432
$^{\text{f}}\text{GottBERT}_{large}$	52009	357145600
GBERT$_{\text{large}}$	31102	335735808
GELECTRA$_{\text{large}}$	31102	334686208
XLM-R$_{\text{base}}$	250002	278043648
mBERT	119547	177853440
GeistBERT	52009	125985024
$^{\text{f}}\text{GottBERT}_{base}$	52009	125985024
GBERT$_{\text{base}}$	31102	109927680
dbmdzBERT	31102	109927680
GELECTRA$_{\text{base}}$	31102	109337088
GermanBERT	30000	109081344

training process required approximately 8.3 days of computation time.

Importantly, GeistBERT started from a relatively low perplexity due to continued pre-training. In comparison, $^{\text{f}}\text{GottBERT}_{base}$ (trained entirely from scratch) started with a perplexity of about 52,592 and converged to around 4, whereas GeistBERT began at 35.17 and converged down to approximately 11. This illustrates the potential stability and efficiency benefits of continued pre-training in reaching useful representations quickly.

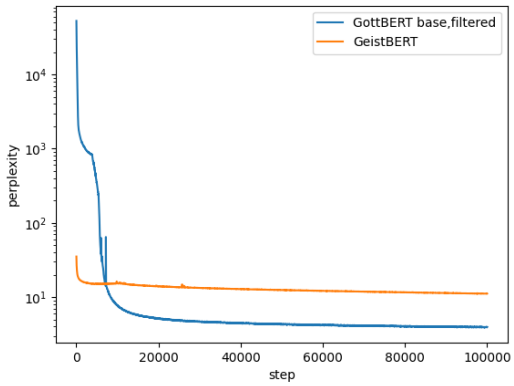

Figure 1: Perplexity of $^{\text{f}}\text{GottBERT}_{base}$ and GeistBERT, evaluated on the validation set after each optimization cycle; values are plotted on a logarithmic y-axis.

4.2 Downstream Tasks

GeistBERT sets a new state-of-the-art among base models for German NLP, outperforming all comparable models and closely approaching large-scale model performance across tasks. It even achieves absolute SOTA in GermEval 2018 fine-grained classification (see Table 6).

The optimal hyperparameters selected per task are summarized in Table 5, extending the original GottBERT setup (Scheible et al., 2024) by including GeistBERT models. The total computation time for all downstream evaluations amounted to 517 hours and 24 minutes (≈21.6 days) on two Nvidia RTX 3090 GPUs, detailed per task in Table 4.

Table 4: Computation time in hours and minutes for the downstream tasks summing up to 517 hours and 24 minutes, which are approximately 21.6 days.

Task		Computation Time
XNLI		47:52
GermEval 2014		235:36
CoNLL03		92:45
GermEval 2018	coarse	45:46
	fine	43:25
10kGNAD		85:49

NLI GeistBERT$_{\text{base}}$ achieves an accuracy of 82.67% on the German NLI task, outperforming all other base models in our evaluation. While it does not surpass top-scoring large-scale models such as GELECTRA$_{\text{large}}$ (86.33%) or GBERT$_{\text{large}}$ (84.21%), it performs competitively and even surpasses GottBERT_{large} (82.46%) and nearly matches $^{\text{f}}\text{GottBERT}^{\dagger}_{large}$ (82.79%), narrowing the performance gap despite its smaller size.

Named Entity Recognition GeistBERT achieves strong F_1 scores on both CoNLL 2003 (86.17%) and GermEval 2014 (88.47%), outperforming all other base models in our evaluation. It also surpasses all large-scale GottBERT variants on GermEval 2014 and comes remarkably close on CoNLL 2003, with only a 0.11% gap to the lowest-scoring large variant. While top-performing large models such as GBERT$_{\text{large}}$ (87.19% on CoNLL) and XLM-R$_{\text{large}}$ (88.83% on GermEval) remain ahead, GeistBERT narrows the performance gap significantly, demonstrating robust entity representation capabilities despite its compact size.

Text Classification GeistBERT$_{\text{base}}$ achieves strong performance across all classification tasks, ranking first in GermEval 2018 fine-grained classification (66.42%), second in 10kGNAD (90.89%), and third in GermEval 2018 coarse (79.67%). It consistently outperforms all other base models and surpasses several large-scale models, particularly

Table 5: Hyperparameters of the best downstream task models for each task and pre-trained model. This table extends the original GottBERT setup by including GeistBERT models. BS refers to batch size, and LR denotes the learning rate.

Model	**GermEval 2014**		**CoNLL 03**		**GermEval 2018**				**10kGNAD**	
					coarse		**fine**			
	BS	LR	BS	LR	BS	LR	BS	LR	BS	LR
GeistBERT	16	5 E-06	32	2 E-05	48	5 E-05	32	2 E-05	16	1 E-06
GottBERT$_{base}$	16	1 E-05	32	2 E-05	48	7 E-06	32	5 E-06	32	5 E-06
GottBERT$^{\dagger}_{base}$	48	2 E-05	32	5 E-05	48	1 E-05	64	7 E-06	32	5 E-06
^{f}GottBERT$_{base}$	16	7 E-06	16	1 E-05	16	1 E-05	48	2 E-05	16	5 E-06
^{f}GottBERT$^{\dagger}_{base}$	16	1 E-05	64	5 E-05	16	1 E-05	16	2 E-05	16	1 E-05
GELECTRA$_{\text{base}}$	32	5 E-05	64	5 E-05	16	2 E-05	48	5 E-05	48	5 E-05
GBERT$_{\text{base}}$	16	2 E-05	64	2 E-05	32	1 E-05	16	5 E-05	16	2 E-05
dbmdzBERT	48	2 E-05	48	5 E-05	16	5 E-06	64	2 E-05	16	2 E-05
GermanBERT	32	2 E-05	16	1 E-05	16	1 E-05	32	1 E-05	32	5 E-05
XLM-R$_{\text{base}}$	64	2 E-05	16	1 E-05	48	5 E-05	64	5 E-05	48	2 E-05
mBERT	48	1 E-05	16	2 E-05	16	2 E-05	64	5 E-05	64	2 E-05
GottBERT$_{large}$	64	5 E-06	16	5 E-06	64	5 E-06	32	7 E-06	64	1 E-06
^{f}GottBERT$_{large}$	32	5 E-06	48	2 E-05	32	5 E-06	32	7 E-06	16	5 E-06
^{f}GottBERT$^{\dagger}_{large}$	16	5 E-06	48	1 E-05	48	1 E-05	32	5 E-06	64	2 E-05
GELECTRA$_{\text{large}}$	16	7 E-06	16	5 E-06	64	1 E-05	32	2 E-05	32	2 E-05
GBERT$_{\text{large}}$	16	7 E-06	32	5 E-06	16	2 E-05	64	2 E-05	64	5 E-05
XLM-R$_{\text{large}}$	16	7 E-06	48	1 E-05	32	1 E-05	32	1 E-05	16	5 E-06

Table 6: All the results of the experiments are shown in percent. They are all based on the test set and the best score out of 24 runs (selection based on validation set). While NLI is measured by accuracy, all the other metrics are F_1 measures. Per model size, best results are **bold**, second-best underlined. Results for GottBERT are reported on both the unfiltered and filtered corpora, the latter indicated by ^{f}. For each GottBERT model, we include both the best and last checkpoint of the pre-training, with the last denoted by †. Values for non-GeistBERT models are taken from Scheible et al. (2024).

Model	**XNLI**	**GermEval 2014**	**CoNLL 03**	**GermEval 2018**		**10kGNAD**
				coarse	**fine**	
GeistBERT	**82.67**	**88.47**	**86.17**	**79.67**	**66.42**	**90.89**
GottBERT$_{base}$	80.82	87.55	85.93	78.17	53.30	89.64
GottBERT$^{\dagger}_{base}$	81.04	87.48	85.61	78.18	53.92	90.27
^{f}GottBERT$_{base}$	80.56	87.57	86.14	78.65	52.82	89.79
^{f}GottBERT$^{\dagger}_{base}$	80.74	87.59	85.66	78.08	52.39	89.92
GELECTRA$_{\text{base}}$	81.70	86.91	85.37	77.26	50.07	89.02
GBERT$_{\text{base}}$	80.06	87.24	85.16	77.37	51.51	90.30
dbmdzBERT	68.12	86.82	85.15	77.46	52.07	90.34
GermanBERT	78.16	86.53	83.87	74.81	47.78	90.18
XLM-R$_{\text{base}}$	79.76	86.14	84.46	77.13	50.54	89.81
mBERT	77.03	86.67	83.18	73.54	48.32	88.90
GottBERT$_{large}$	82.46	88.20	86.78	79.40	54.61	90.24
^{f}GottBERT$_{large}$	83.31	88.13	86.30	79.32	54.70	90.31
^{f}GottBERT$^{\dagger}_{large}$	82.79	88.27	86.28	78.96	54.72	90.17
GELECTRA$_{\text{large}}$	**86.33**	88.72	86.78	**81.28**	56.17	**90.97**
GBERT$_{\text{large}}$	84.21	88.72	**87.19**	80.84	**57.37**	90.74
XLM-R$_{\text{large}}$	84.07	**88.83**	86.54	79.05	55.06	90.17

in the fine-grained setting. The results indicate that GeistBERT performs competitively across diverse classification benchmarks, despite being a base-sized model.

5 Discussion

5.1 Principal Findings

The continued pre-training of GottBERT on a broader and partially deduplicated and filtered German corpus consisting of OSCAR23, OPUS, mC4, Wikipedia, and OpenLegal, together with the use of WWM, leads to clear improvements across multiple language modeling tasks. GeistBERT establishes a new state of the art among base models and achieves competitive results with larger models across multiple German NLP benchmarks.

5.2 Training Considerations and Data Quality

In contrast to the TPU-based training used for GottBERT, GPU training also enabled more flexible preprocessing, such as sentence-aware segmentation. This made it possible to preserve natural sentence structure during training, even when using fixed-length sequences. Nevertheless, hyperparameter tuning remains a crucial factor for achieving strong downstream performance (Dodge et al., 2020). WWM contributed to improved tokenization, aligning with previous findings (Martin et al., 2020; Chan et al., 2020). However, we did not perform a dedicated ablation study comparing WWM with standard subword masking, as this would have required training an additional baseline model. Nevertheless, the consistently strong downstream results of GeistBERT suggest that WWM contributed positively, in line with earlier findings. Moreover, we were able to adopt a higher peak learning rate (0.0007), which may also have been facilitated by initializing from the $^{\text{f}}\text{GottBERT}_{base}$ checkpoint.

While deduplication and filtering were applied to CulturaX, other subcorpora (e.g., OPUS, Wikipedia, OpenLegal) were only partially processed or left unfiltered. This means that some redundant or lower-quality data may still be present. Prior work suggests that models benefit from increased corpus diversity (Martin et al., 2020), and GeistBERT's use of many different corpora likely contributed to its robustness. Additionally, vocabulary size plays a role in performance (Toraman et al., 2023), though ours remains well-optimized.

We did not perform ablation experiments per subcorpus, as this would have required multiple additional large-scale pre-training runs. Nevertheless, we expect that improvements are not only attributable to the sheer size of the training data (1.3 TB), but also to the increased heterogeneity of the sources. The OSCAR23+mC4 portion clearly contributed the majority of the volume, while smaller corpora such as OpenLegal, Wikipedia, and OPUS are likely to have increased linguistic and domain diversity. Prior findings from CamemBERT (Martin et al., 2020) indicate that variance of a corpus matter and impacts downstream robustness, which suggests that the mix of sources in GeistBERT was similarly beneficial.

5.3 Continued Pre-training and Outlook

We chose to continue pre-training from GottBERT rather than training GeistBERT from scratch, as it is common practice with domain-specific adaptations (Lentzen et al., 2022; Lee et al., 2019; Arefeva and Egger, 2022; Gururangan et al., 2020b). This allowed us to reuse German-specific tokenization and pre-trained weights, and to focus on training and evaluating a single, well-defined setup within time constraints. While training from scratch with a custom vocabulary may yield more tailored embeddings (El Boukkouri et al., 2022), prior work suggests that continued pre-training often achieves comparable results. A direct comparison between continued pre-training and training from scratch on the same architecture and corpus remains an interesting avenue for future work.

Following the broad adoption of GottBERT in German NLP (Scherrmann, 2023; Bressem et al., 2024; Lentzen et al., 2022; Xu et al., 2021; Frei et al., 2022; Frei and Kramer, 2023), we hope GeistBERT will be similarly received and applied.

6 Conclusion

In this work, we introduced GeistBERT, a German RoBERTa-based language model trained on a diverse as well as partially deduplicated and filtered corpus, incorporating WWM to enhance pre-training. GeistBERT achieves SOTA performance among base models and even outperforms several larger models across multiple tasks. These results underscore the importance of corpus diversity and WWM in improving downstream performance. GeistBERT is released under the MIT license on Huggingface, with fairseq checkpoints provided.

Limitations

Several limitations should be acknowledged in this study. First, while deduplication and filtering were applied to CulturaX (OSCAR23 + mC4) and deduplication to selected OPUS corpora, other parts of the dataset (e.g., Wikipedia, OpenLegal) were not processed, potentially leaving redundant or noisy data.

Second, GeistBERT's training data, though diverse, remains specific to the selected corpora (OSCAR23, OPUS, mC4, Wikipedia, OpenLegal). Its generalization to other datasets or domains remains uncertain, and performance on dialects and cultural nuances within German may be limited. Further fine-tuning could improve adaptability to regional language variations.

Third, we did not include a detailed error analysis of model predictions. While such an analysis could provide additional insights into systematic failure modes, our focus in this work was on efficiency and establishing strong baselines for German NLP.

Finally, due to efficiency constraints and limited computational resources, we did not train a large version of GeistBERT, as pretraining based on GottBERT estimates would have required approximately 4.75 times more compute. While our results demonstrate the strong performance of the base model, larger architectures could potentially achieve even better results.

Ethical Considerations

Like all large-scale language models, GeistBERT may inherit biases from its training data, which can influence downstream tasks such as classification or decision-making. While deduplication reduces redundancy and noise, it does not remove deeper societal or representational biases. Furthermore, training on large web-based corpora raises privacy concerns, as models may inadvertently retain sensitive information. Responsible deployment is especially important in high-stakes domains like legal, medical, or financial NLP.

Despite optimizations for efficiency, pre-training and evaluating transformer models remain computationally demanding, contributing to energy use and carbon emissions. These environmental costs highlight the need for balancing model performance with sustainable development goals.

References

Veronika Arefeva and Roman Egger. 2022. When bert started traveling: Tourbert—a natural language processing model for the travel industry. *Digital*, 2(4):546–559.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014. GermEval 2014 Named Entity Recognition Shared Task: Companion Paper. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 104–112.

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J.W.L. Aerts, and Alexander Löser. 2024. medbert.de: A comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3504–3514.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv:2002.06305 [cs]*. ArXiv: 2002.06305.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or train from scratch? comparing pre-training strategies of BERT in the medical domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2626–2633, Marseille, France. European Language Resources Association.

Johann Frei, Ludwig Frei-Stuber, and Frank Kramer. 2022. Gernermed++: Transfer learning in german medical nlp.

Johann Frei and Frank Kramer. 2023. Annotated dataset creation through large language models for non-english medical nlp. *Journal of Biomedical Informatics*, 145:104478.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020a. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020b. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Manuel Lentzen, Sumit Madan, Vanessa Lage-Rupprecht, Lisa Kühnel, Juliane Fluck, Marc Jacobs, Mirja Mittermaier, Martin Witzenrath, Peter Brunecker, Martin Hofmann-Apitius, Joachim Weber, and Holger Fröhlich. 2022. Critical assessment of transformer-based AI models for German clinical notes. *JAMIA Open*, 5(4):ooac087.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *arXiv:1904.01038 [cs]*. ArXiv: 1904.01038.

Julian Risch, Eva Krebs, Alexander Löser, Alexander Riese, and Ralf Krestel. 2018. Fine-Grained Classification of Offensive Language. In *Proceedings of GermEval 2018 (co-located with KONVENS)*, pages 38–44.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. 2024. GottBERT: a pure German language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21237–21250, Miami, Florida, USA. Association for Computational Linguistics.

Moritz Scherrmann. 2023. German finbert: A german pre-trained language model.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, USA. Association for Computational Linguistics. Event-place: Edmonton, Canada.

Cagri Toraman, Eyup Halit Yilmaz, Şahinuç Furkan, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.