Die SuperGLEBer at GermEval 2025 Shared Tasks: Growing Pains - When More Isn't Always Better

Julia Wunderle 🍺 **and Jan Pfister** 🍺 **and Andreas Hotho**
Data Science Chair
Center for Artificial Intelligence and Data Science (CAIDAS)
Julius-Maximilians-Universität Würzburg (JMU)
`{lastname}@informatik.uni-wuerzburg.de`

Abstract

We participate in this year's GermEval 2025 Shared Tasks by extending SuperGLEBer, a comprehensive benchmark for evaluating German language understanding to the new tasks. Rather than focusing on optimizing task-specific performance, we adopt a complementary strategy: applying simple methods across 38 diverse (L)LMs (100M to 9B parameters) to analyze the tasks themselves, revealing that most models perform similarly on this year's tasks compared to existing SuperGLEBer tasks. Notably, the regression-based verifiability rating task diverges from this trend, emerging as substantially more difficult and methodologically distinct. Through our comprehensive analysis, we find that three new tasks, including Flausch-Erkennung subtask 2, rank among the top 10 most discriminating tasks of the benchmark, effectively distinguishing between model capabilities. Most remarkably, we demonstrate that just 2-3 strategically selected tasks can approximate the complete benchmark rankings with 97-99% correlation, potentially enabling more efficient large-scale model evaluation while maintaining ranking accuracy. Overall, our submissions achieved competitive results, placing 1st (out of one)-6th across different tasks, i.e. for Flausch-Erkennung subtask 1 and 2 we placed 3rd and 6th respectively.

1 Introduction

In an earlier work we introduced the first comprehensive benchmark for German language understanding models called SuperGLEBer (Pfister and Hotho, 2024). It contains 29 diverse tasks such as document classification, sequence tagging, sentence similarity, and question answering. This allowed us to systematically evaluate the performance of any existing HuggingFace model on German language understanding tasks.

In this work, we extend SuperGLEBer with all four shared tasks from GermEval 2025 to provide a comprehensive evaluation of current German models on all shared tasks: (**1**) **Flausch-/Candy Speech Detection** (Clausen et al.), which identifies "candy speech" in YouTube comments through coarse–grained binary classification and fine-grained span detection with category assignment (10 categories); (**2**) **SustainEval** (Prange et al.), which analyzes German sustainability reports through content classification of text snippets to predefined reporting criteria and verifiability rating of sentences on a 0.0-1.0 scale; (**3**) **Harmful Content Detection in Social Media** (Felser et al.), which detects harmful content in German social media posts through three binary/multi-class classifications: calls to action, stance toward democratic order (4 categories), and positive attitudes toward violence; and (**4**) **LLMs4Subjects** (D'Souza et al.), a multi-label classification task that assigns library records to subject domains (28 categories) using the GND taxonomy. We participate in all subtasks, except for LLMs4Subjects, where we only benchmark on subtask 1. However, we did not officially submit a run due to limitations requiring participation in both subtasks, and subtask 2 was not trivial to implement into the existing SuperGLEBer framework[1].

Our approach is partly inverse to traditional shared task participations, following the spirit of van der Goot (2022, 2023) at SemEval: we participate in multiple shared tasks at once, and rather than focusing solely on optimizing model performance, we aim to learn more about the tasks themselves by applying simple methods across a wide range of models. This systematic evaluation allows us not only to assess model performance but also to analyze the complexity and discriminative power of the new tasks in comparison to existing SuperGLEBer benchmark tasks.

🍺 These authors contributed equally to this work.

[1] https://github.com/LSX-UniWue/SuperGLEBer

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 479–493, 2025

2 Modelling the tasks in SuperGLEBer

SuperGLEBer is an open-source benchmark suite designed to evaluate German language understanding of language models across a wide range of tasks (Pfister and Hotho, 2024). It includes 29 diverse tasks such as document classification, sequence tagging, sentence similarity, and question answering, reflecting the breadth of natural language understanding challenges in German. We add the shared tasks to the SuperGLEBer framework, which already supports most task types. Following flair's implementation (Akbik et al., 2019) we finetune the text classification tasks using the standard approach of adding a linear layer on top of the output representation of the CLS token, while for sequence tagging tasks we use the same approach, but train the linear layer to predict the correct class on top of the output representation of each input token individually.

Flausch-Erkennung For the binary classification in subtask 1, the input sentence: *"ihr seid die Besten"* has to be classified as `yes`. For the span prediction in subtask 2, each token's class is predicted individually using a BIO-label scheme:

Token	Label
ihr	`B-affection_declaration`
seid	`I-affection_declaration`
die	`I-affection_declaration`
Besten	`I-affection_declaration`

SustainEval The multi-class classification subtask 1 is solved by assigning text snippets to predefined reporting criteria: *"Prävention Über das Kerngeschäft „Versicherung" hinaus fühlt sich . . . "* and *"Die [ORG] arbeitet hier eng mit den relevanten Institutionen und Einrichtungen. . . "* the model has to classify this as class `18`, i.e `Corporate Citizenship`. For the regression subtask 2 the verifiability of the last sentence given the previous context has to be predicted, in this case `0.667`. Here, we also follow flair's standard setup, using a simple linear layer on top of the model's hidden states. As we found that this can be unstable during training, we add two stabilizing measures:

1. An RMS-Norm before the linear layer, as some hidden state features can be very large, leading to training instabilities.

2. A sigmoid activation function after the linear layer, constraining outputs to the value range of the task: between 0 and 1.

Harmful Content Detection For harmful content detection, we perform binary or multi-class classification across three subtasks to detect harmful content in German social media posts: C2A and VIO use binary classification to detect calls to action and positive attitudes toward violence respectively. For example, given the input: *"In 300 Jahren sind dann alle wieder daheim"* the model classifies this as `False` for the violence subtask. DBO performs multi-class classification (4 categories) to determine stance toward democratic order. For example, given: *"Die schädliche Arbeitsagentur soll mal in den Spiegel schauen."* the correct label would be `criticism`.

LLMs4Subjects Subtask 1 Here, we perform multi-label classification (28 categories) to assign library records to subject domains. For example, given the input describing a cultural heritage site: *"Der Löhrerhof in Hürth : Denkmalgerechte Planung und Sanierung "Der Löhrerhof ist eine bäuerliche Hofanlage aus dem 19. Jh. im Stadtteil Alt-Hürth, die zuletzt hauptsächlich als Kunst- und Kulturzentrum genutzt wurde."* the model assigns the labels `arc ; bau`.

3 Experiments

We evaluate the performance of all models from the original SuperGLEBer benchmark, as well as models added after its official release using the flair internal default metric: F1-score for classification and sequence tagging, and mean squared error for regression.

3.1 Models

This results in evaluating 38 models, all of which are listed alongside their results in Table 5. As SuperGLEBer is a model-architecture-agnostic framework, which allows us to evaluate most HuggingFace compatible models, we refer to these models using their HuggingFace identifiers, often omitting the leading organization name for simplicity (i.e. "(LSX-UniWue/)ModernGBERT_1B"). The models vary in size (from 100M to 9B parameters), and architecture (encoder and decoder), including recent multilingual models, German-specific models and models pretrained in languages other than German and later fine-tuned on German data (see Table 6 for more details).

3.2 Training Setup

For each of the task type, we follow SuperGLEBer (Pfister and Hotho, 2024) in implementing the training routine. For text classification, regression and sequence tagging we use flair (Akbik et al., 2019), applying consistent training procedures across all models: a batch size of 8 (without gradient accumulation), a learning rate of 5e-5, 5 epochs and a seed set to 42. In addition, we introduce a maximum input sequence length of 512 tokens, to enable fair comparisons for all models regardless of their individual context size and class weighting for all classification tasks during training.

Efficient Training with (Q)LoRA We consequently opt to use QLoRA-training (Dettmers et al., 2023) for all models where it is supported by the HuggingFace library, falling back to LoRA (Hu et al., 2022) for the GBERT family where quantization is not supported. Enabling (Q)LoRA for all models where possible ensures comparability between different models and rules out the possibility that the performance difference between models stems from different training procedures. Following the hyperparameters given by Dettmers et al. (2023) we use 4-bit quantization, double quantization and NormalFloat4, with a default LoRA rank of 8 and a dropout rate of 0.1. All models are trained on H100 GPUs.

4 Results

4.1 How does our simple approach perform?

Our full evaluation on the development datasets is presented in Table 5. For datasets without a predefined development split, we randomly split the training set into new train and dev subsets using an 80:20 ratio.

Flausch-Erkennung For Flausch-Erkennung, the binary classification subtask 1 proves relatively straightforward, with leo-hessian-7b achieving the highest performance (0.953), closely followed by DOSMo-7B-v0.2 (0.952). However, the span detection subtask 2 emerges as significantly more challenging, with ModernGBERT_1B leading at only 0.662, followed by LLäMmlein2Vec_7B (0.657) according to our metric. This substantial performance drop highlights the complexity of fine-grained span detection with category assignment compared to coarse-grained binary classification.

SustainEval Both SustainEval subtasks present considerable challenges for current models. The content classification subtask achieves moderate performance with ModernGBERT_1B leading (0.659) and LLäMmlein_7B following (0.633). The verifiability rating regression subtask proves particularly difficult, with even the best-performing Meta-Llama-3.1-8B reaching only 0.454, while Llama3-German-8B achieves 0.430. This regression task's low scores across all models suggest it requires fundamentally different capabilities than traditional classification tasks.

Harmful Content Detection The harmful content detection tasks show varied difficulty levels across subtasks. Call-to-action detection (c2a) achieves high performance with ModernGBERT_1B (0.953) and LLäMmlein_7B (0.952) performing similarly well. Stance detection toward democratic order (dbo) proves moderately more challenging, with LLäMmlein2Vec_7B leading (0.900), followed closely by ModernGBERT_1B and leo-hessian-7b (both 0.897). Violence detection (vio) represents the easiest subtask overall, with LLäMmlein_7B achieving the highest score (0.956) and EuroLLM-9B close behind (0.954).

LLMs4Subjects The multi-label subject classification task shows competitive performance between top models, with ModernGBERT_1B (0.787) and Llama3-German-8B (0.786) achieving nearly identical results. This close performance suggests that both encoder and decoder architectures can effectively handle this taxonomic classification challenge.

Overall Performance Patterns Remarkably, the German-only encoder ModernGBERT_1B demonstrates exceptional consistency, achieving best or second-best performance in 6 out of 8 subtasks despite being up to 9 times smaller than competing models. This pattern reinforces that specialized German training outweighs raw parameter count for these tasks. The performance hierarchy generally follows with decoder models leo-hessian-7b, LLäMmlein variants, and Meta-Llama-3.1-8B, most of which received explicit German fine-tuning, while even recent SotA model families like Qwen stay behind, highlighting the importance of specialized training for German. Despite this, many overall well performing models do not perform as well on SustainEval-Regression, with scores ranging from 0.029 to 0.454. This suggests that verifiability rating and the regression task present different challenges than the other SuperGLEBer tasks.

Our official submissions Overall, we submitted the three best overall performers according to our analyses - "ModernGBERT_1B", 'LLaMmlein_7B", and "Llama-3.1-8B" (details in Table 6) - to all official leaderboards except for LLMs4Subjects. Our simple SuperGLEBer framework approach achieved competitive results across most tasks: 3rd place out of 11 teams for Flausch-Erkennung subtask 1, with an F1 score of 0.883 (Table 1), 6th place out of 7 teams for Flausch-Erkennung subtask 2, with a strict F1 of 0.127, type F1 of 0.173, and span F1 score of 0.580 (Table 2). Notably, for subtask 2 we observed a significant discrepancy between the official span-based evaluations vs. our BIO-label token classification-based evaluation, which we aim to analyze in the future. For the SustainEval classification subtask we achieved 4th place out of 6 teams, with an accuracy score of 0.573 and 1st place out of 1 team for SustainEval regression subtask, with an kendall's tau score of 0.402 (see Table 3). Official Harmful-Content results have not yet been published[2]. For subtask 1 (c2a) we achieved a F1 score of 0.870, subtask 2 (dbo) 0.690 and subtask 3 vio (0.840).

4.2 How do models perform on the new tasks, compared to the old tasks?

As we augmented the SuperGLEBer benchmark with new tasks, we want to understand the performance relationship between the new shared tasks from this year and the already existing tasks. To shed light on this, we analyze the ranking consistency between tasks using Spearman rank correlation. Figure 1 shows the correlation matrix between all new tasks (rows) and existing tasks (columns) based on model performance rankings.

What we find is that most new tasks exhibit strong positive correlations (0.7-0.9) with many existing tasks, suggesting that models performing well on existing SuperGLEBer tasks also excel on these new tasks. In contrast, SE-Regression shows weak or inverse correlations, highlighting it as a distinct challenge that deviates from existing performance patterns. Interestingly, for the existing up-dep and up-pos tasks, we find a (strong) inverse correlation to most new tasks - except for Flausch-Erkennung Tagging and Harmful Content c2a, indicating that models that perform well on these tasks do not perform well on the new tasks. We assume up-dep and up-pos are special, because they are the only syntactic tasks in SuperGLEBer, focusing on predicting dependency parse labels and POS tags respectively, while all other tasks are semantic/pragmatic in nature, focusing on extracting meaning and content such as named entity recognition or text classification. This also becomes visible when embedding each task by interpreting each model's performance on a task as a "feature" (Figure 5) and subsequently calculating a principal component analysis (PCA) on the resulting space: as anticipatable the previously already discussed up-pos, up-dep and SE-Regression are most "out-of-distribution".

4.3 How well do the tasks discriminate between models?

Inspired by these insights, we reverse the process by interpreting task performance as embedding features for the models in Figure 4. We observe distinct clustering patterns: smaller multilingual models without a German focus (i.e., "bloomz-560m" and "Qwen2.5-0.5B") group in the lower-left quadrant, while high-performing models like ModernGBERTs and larger LLäMmlein variants tend to place in the upper-right region. Smaller German-focused encoder models (e.g., gbert variants and gelectra models) cluster together in the upper-center area, suggesting similar performance profiles despite architectural differences. On the other hand, decoder models (e.g. "bueble-lm-2b", EuroLLM, "leo-hessianai-7b") are grouped in the lower-right area. This pattern suggests that model architecture families and language-specific training create recognizable performance signatures that transcend individual task results.

In this figure the first two principal components already capture 74.7% of the total variance (PC1: 61.9%, PC2: 12.8%), indicating that model performance can be largely characterized by these two dominant factors, each resembling an axis in the "task performance space". Similarly, when we color the models in this graph according to the performance of the subtasks in Figure 3, we can identify a clear "gradient" in this PCA space along which the performance evolves rather monotonously. This is an indicator that most of the model performance discrimination might be driven by a small subset of tasks in the SuperGLEBer.

To further analyze this behavior we calculate

[2] although it looks like atleast a Top-3 finish at the moment

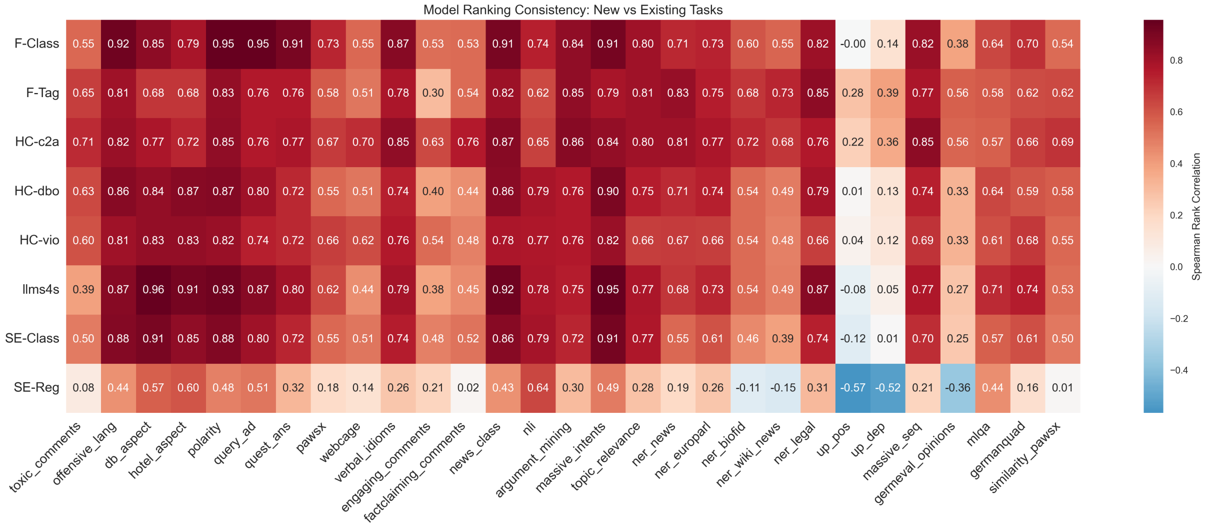


Figure 1: Model ranking consistency between new GermEval 2025 tasks and existing SuperGLEBer tasks. Each cell shows the Spearman rank correlation coefficient between model rankings on the corresponding task pair. Darker red indicates higher positive correlation (similar model rankings), while blue indicates negative correlations. The analysis reveals which new tasks are most predictive of performance on the existing tasks.

the standard deviation of model performances for each task. Tasks with higher deviation in model performance scores are more effective at distinguishing between models, while tasks where all models perform similarly provide less discriminative power (Figure 2). We find that three of the new tasks (SE-Class, Flausch-Tagging and SE-Regression) are able to create a standard deviation across models of about 0.1 or more, placing them in the top 10 most discriminating tasks overall. A more detailed list of task discrimination for new and existing tasks separately is provided in Figure 6. When examining these most discriminative tasks, we identify a common characteristic: most of these tasks have many target classes, like for the existing tasks “hotel_aspect” (15 labels, one for each combination of an aspect-class and sentiment class) or “massive_intents” (60 labels). A similar pattern holds for the new tasks: “SE-class” (20 labels), and “Flausch-tagging” (21 labels).

4.4 Do we need the entire SuperGLEBer benchmark to reproduce the rankings?

Motivated by the findings in Sections 4.2 and 4.3, we investigate whether a smaller, carefully chosen subset of tasks can approximate the full benchmark's model rankings. This could greatly reduce evaluation costs while maintaining ranking accuracy. Using a greedy selection strategy, we construct subsets of increasing size by iteratively adding tasks that maximize alignment with the full 37-task ranking across all 38 models. We measure the subset quality using Spearman rank correlation, mean absolute rank difference (MAD), exact rank matches (number of models that keep their rank), and the number of models involved in ranking ties.

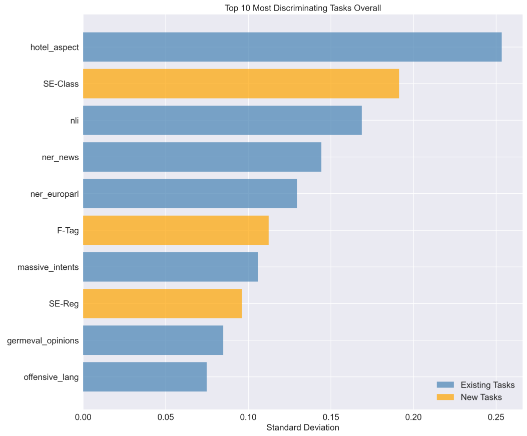


Figure 2: Top 10 most discriminating tasks, performance score standard deviation across all 38 evaluated models. Higher deviation indicates better discrimination between model capabilities: three of the new tasks from this year rank among the most discriminating overall.

As shown in Table 4, remarkably few tasks are needed to approximate the full benchmark rankings. Starting with a single task (“massive_intents”), we achieve 96.9% correlation with the complete 37-task ranking, though this results in 10 models having identical performance scores and thus being tied for the same rank position.

When we expand to two selected existing SuperGLEBer tasks - “verbal_idioms” and “massive_intents” - the correlation improves to 97.8% while reducing ranking ambiguity: only 2 models remain tied in placement, compared to 10 with a single task. This means that 36 out of 38 models can be clearly distinguished and ranked using these two tasks. Expanding further to three tasks (“toxic_comments”, “db_aspect”, and “ner_legal”) achieves 99.0% correlation with the full benchmark. The MAD, which measures the average difference in rank positions between the subset and full rankings, drops to 1.21 ranks. This means that on average, a model’s rank using these three tasks differs by only about one position from its rank using all 37 tasks. Importantly, this subset eliminates all ties, providing a clear ranking for all models.

A four-task subset maintains similar correlation (99.2%) with slightly better rank precision (MAD of 1.16), while a five-task subset that includes one of this year’s new tasks (“HC-dbo”) achieves 99.4% correlation with a MAD of just 0.68 ranks. At this point, 22 out of 38 models maintain their exact rank positions compared to the full benchmark.

These results demonstrate that comprehensive model evaluation can be effectively approximated using a small, strategically chosen subset of tasks, making large-scale model comparison more computationally efficient - confirming that more isn’t always better when it comes to benchmark design.

5 Related Work

GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) established the paradigm of comprehensive language understanding benchmarks with 11 and 10 diverse NLU tasks respectively. The English-only nature of these benchmarks led to the development of similar multilingual efforts for e.g. Russian (Shavrina et al., 2020), Persian (Khashabi et al., 2021), and Bulgarian (Hardalov et al., 2023).

While cross-lingual benchmarks like XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) include German tasks, their focus on cross-lingual transfer rather than monolingual capabilities makes them less suitable for comprehensive German model evaluation. Previous German evaluation efforts have been task-specific, focusing on individual capabilities like sentiment analysis (Cieliebak et al., 2017) or coreference resolution (Schröder et al., 2021), rather than providing comprehensive benchmarking frameworks.

6 Conclusion

We successfully extended SuperGLEBer with four new GermEval 2025 shared tasks, providing comprehensive evaluation of 38 models ranging from 100M to 9B parameters. Our simple approach achieved competitive results across tasks, placing 1st-6th in the shared task rankings, and on average performed best when using a “small” 1B model.

The analysis reveals that most new tasks correlate well with existing SuperGLEBer tasks (0.7-0.9 correlation), except for the regression-based verifiability rating which presents distinct challenges. Three new tasks rank among the top 10 most discriminating tasks, effectively distinguishing between model capabilities. We demonstrate that strategic task selection can reduce model benchmarking costs: just a small subset of 2-3 carefully chosen tasks approximate complete benchmark rankings consisting of 37 tasks with 97-99% correlation, enabling efficient large-scale model evaluation while maintaining ranking accuracy.

Regarding the Flausch-Erkennung (Candy Speech Detection) task specifically, we achieved solid performance with 3rd place in the binary classification subtask but found the span detection subtask more challenging, placing 6th with notable discrepancies between span-based and BIO-token evaluation approaches. Interestingly, the Flausch-Erkennung tagging subtask emerged as one of the top 10 most discriminating tasks in our benchmark, effectively distinguishing between model capabilities across the 38 evaluated models, suggesting its value for comprehensive model evaluation beyond just detecting candy speech in social media.

Limitations

Our evaluation relies on a simple approach that may (likely) not capture the full potential of more sophisticated methods for these tasks. The regression task implementation required additional stabilization measures, indicating potential challenges in the dataset or our approach - as no other team participated in this subtask, we lack points of comparison for further investigations.

Acknowledgements

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich Alexander-Universität Erlangen-Nürnberg (FAU) under the

NHR project b185cb. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Furthermore, we gratefully acknowledge the HPC resources provided by the JuliaV2 cluster at the Universität Würzburg (JMU), which was funded as DFG project as "Forschungsgroßgerät nach Art 91b GG" under INST 93/1145-1 FUGG. The project staff is partially funded by the DFG – 529659926. The data science chair is part of the CAIDAS, the Center for Artificial Intelligence and Data Science, and is supported by the Bavarian High-Tech Agenda, which made this research possible. This publication was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the project LitBERT, project no. 529659926. Furthermore, we would like to thank the task hosts for organizing the shared tasks and providing the data.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *Preprint*, arXiv:2503.05500.

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.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Yulia Clausen, Tatjana Scheffler, and Michael Wiegand. 2025. Overview of the GermEval 2025 Shared Task on Candy Speech Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Pieter Delobelle, Alan Akbik, and 1 others. 2024. Büblelm: A small german lm.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

DiscoResearch and Occiglot. 2024. Llama3-discoleo-instruct-8b-v0.1. Instruction-tuned German language model developed with support from DFKI and hessian.AI.

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. The GermEval 2025 2nd LLMs4Subjects Shared Task Dataset.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. Moderngbert: German-only 1b encoder model trained from scratch. *Preprint*, arXiv:2505.13136.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the GermEval 2025 Shared Task on harmful content detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. HsH Applied Academics.

GERTuraX. 2025a. gerturax-1 (revision 18de094).

GERTuraX. 2025b. gerturax-2 (revision 50ede60).

GERTuraX. 2025c. gerturax-3 (revision e0f62ac).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian general language understanding evaluation benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Maximilian Idahl. 2024. DOSMo-7B: A Large Language Model Trained Exclusively on German Text. In *Proceedings of the Konferenz der deutschen KI-Servicezentren 2024 (KonKIS 24)*. Accessed: 2024-12-10.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT2.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning. *arXiv preprint*.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German language understanding evaluation benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Björn Plüster. 2023. LeoLM: Igniting German-Language LLM Research. Accessed: 2024-11-15.

Jakob Prange, Charlott Jakob, Patrick Göttfert, Raphael Huber, Pia Wenzel Neves, and Annemarie Friedrich. 2025. Overview of the SustainEval 2025 Shared Task: Identifying the topic and verifiability of sustainability report excerpts. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. HsH Applied Academics.

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.

Raphael Scheible-Schmitt and Johann Frei. 2025. Geistbert: Breathing life into german nlp. *Preprint*, arXiv:2506.11903.

Fynn Schröder, Hans Ole Hatzel, and Chris Biemann. 2021. Neural end-to-end coreference resolution for German in different domains. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 170–181, Düsseldorf, Germany. KONVENS 2021 Organizers.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4717–4726, Online. Association for Computational Linguistics.

Bayerische Staatsbibliothek. 2025. bert-base-german-cased (revision 43cce13).

Bayerische Staatsbibliothek and Stefan Schweter. 2025. german-gpt2 (revision ab6efd0).

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rob van der Goot. 2022. MaChAmp at SemEval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task multi-lingual learning for a pre-selected set of semantic datasets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1695–1703, Seattle, United States. Association for Computational Linguistics.

Rob van der Goot. 2023. MaChAmp at SemEval-2023 tasks 2, 3, 4, 5, 7, 8, 9, 10, 11, and 12: On the effectiveness of intermediate training on an uncurated collection of datasets. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 230–245, Toronto, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. RedPajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 116462–116492. Curran Associates, Inc.

Team	F1	Precision	Recall
AIxcellent Vibes	0.891	0.927	0.857
HHUflauschig	0.887	0.900	0,875
Die SuperGLEBer	0.883	0.915	0.853
NLPSuedwestfalen	0.880	0.911	0.850
TUM NLP Group	0.879	0.887	0.871
nlp-augsburg-04	0.836	0.903	0.778
StickyBeardAux	0.819	0.868	0.775
Quabynar	0.754	0.710	0.804
Robert-Dennis-UniAugsburg03	0.746	0.817	0.687
Flauschgummi	0.697	0.888	0.574

Table 1: Leaderboard of Flausch-Erkennung subtask 1

Team	F1 (strict)	P (strict)	R (strict)	F1 (type)	P (type)	R (type)	F1 (span)	P (span)	R (span)
AIxcellent Vibes	0.631	0.658	0.605	0.769	0.803	0.738	0.676	0.705	0.648
HHUflauschig	0.615	0.629	0.601	0.766	0.785	0.749	0.668	0.684	0.653
Georg Hofmann	0.498	0.475	0.524	0.680	0.648	0.715	0.567	0.541	0.596
nlp-augsburg-04	0.334	0.241	0.543	0.492	0.355	0.801	0.365	0.264	0.594
Quabynar	0.159	0.149	0.171	0.408	0.381	0.438	0.257	0.240	0.276
Die SuperGLEBer	0.127	0.136	0.120	0.173	0.185	0.162	0.580	0.620	0.544
StickyBeardAux	0.039	0.044	0.036	0.562	0.629	0.507	0.050	0.056	0.045

Table 2: Leaderboard of Flausch-Erkennung subtask 2

Team	Subtask 1 Accuracy	Subtask 2 Kendall's Tau
22520474	0.626	-
1234566	0.586	-
s1nbo	0.579	-
janpf (our)	0.573	0.402
wangkongqiang	0.505	-
supachoke	0.486	-

Table 3: Leaderboard of SustainEval subtask 1 - Content Classification and subtask 2 - Verifiability Rating

Size	Task Subset	Correlation	MAD[†]	Exact Rank Matches	Models in Ties
1	existing_massive_intents	0.969	2.00	3/38	10
2	existing_verbal_idioms existing_massive_intents	0.978	1.71	8/38	2
3	existing_toxic_comments existing_db_aspect existing_ner_legal	0.990	1.21	8/38	0
4	existing_toxic_comments existing_db_aspect existing_verbal_idioms existing_ner_legal	0.992	1.16	6/38	0
5	existing_offensive_lang existing_db_aspect existing_ner_news existing_germanquad new_HC-dbo	0.994	0.68	22/38	0
Full	All 37 tasks (29 existing + 8 new)	1.000	0.00	38/38	0

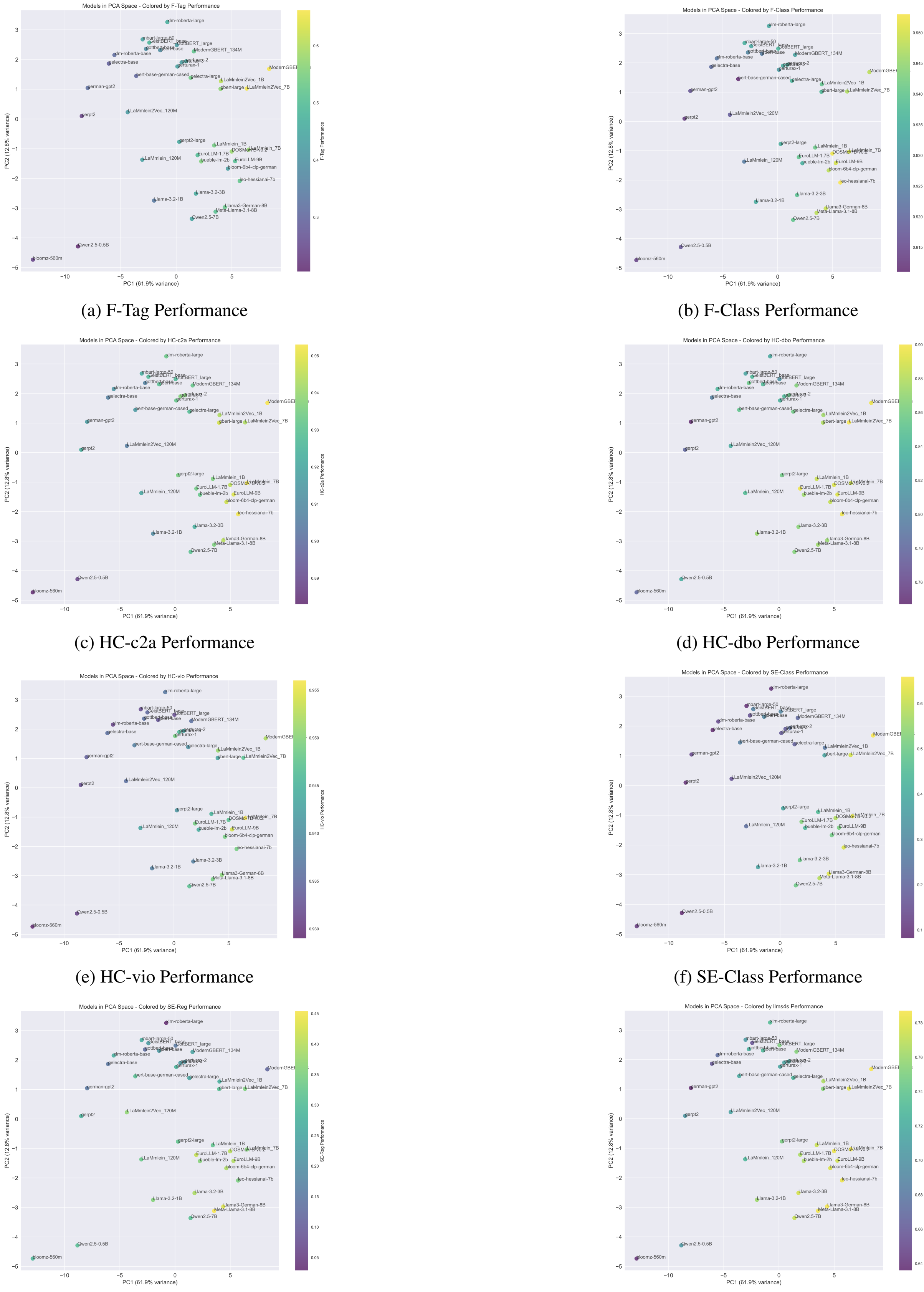
Table 4: Minimal task subsets for reproducing SuperGLEBer model rankings. We evaluate the correlation between subset-based rankings and the full 37-task ranking across 38 models. Results show that just 2 tasks ("existing_verbal_idioms" + "existing_massive_intents") achieve 97.8% correlation with the complete benchmark while reducing ties from 10 to 2 models. This demonstrates that comprehensive model evaluation can be approximated with carefully selected task subsets. [†]MAD = Mean Absolute Rank Difference.

	Flausch		Harmful Content			Llms4s	SustainEval		SuperGLEBer
model	Class	Tag	c2a	dbo	vio	Class	Class	Reg	Old Avg
LLäMmlein_120M [Pfister et al.]	0.924	0.433	0.920	0.845	0.943	0.728	0.225	0.330	0.676
LLäMmlein_1B [Pfister et al.]	0.941	0.541	0.936	0.886	0.945	0.773	0.449	0.321	0.733
LLäMmlein_7B [Pfister et al.]	0.950	0.575	0.952	0.887	**0.956**	0.784	0.633	0.357	0.747
LLäMmlein2Vec_120M [Ehrmanntraut et al.]	0.918	0.419	0.905	0.832	0.936	0.712	0.161	0.368	0.684
LLäMmlein2Vec_1B [Ehrmanntraut et al.]	0.940	0.609	0.943	0.874	0.951	0.765	0.251	0.270	0.762
LLäMmlein2Vec_7B [Ehrmanntraut et al.]	0.947	0.657	0.943	**0.900**	0.947	0.780	0.614	0.350	0.787
ModernGBERT_134M [Ehrmanntraut et al.]	0.932	0.520	0.932	0.865	0.936	0.753	0.210	0.243	0.749
ModernGBERT_1B [Ehrmanntraut et al.]	0.948	**0.662**	**0.953**	0.897	0.952	**0.787**	**0.659**	0.118	**0.808**
german-gpt2 [Staatsbibliothek and Schweter]	0.917	0.318	0.916	0.747	0.934	0.636	0.135	0.151	0.642
gerpt2 [Minixhofer]	0.913	0.237	0.923	0.779	0.932	0.702	0.094	0.281	0.619
gerpt2-large [Minixhofer]	0.935	0.450	0.935	0.858	0.941	0.749	0.382	0.321	0.708
gbert-base [Chan et al.]	0.924	0.386	0.929	0.843	0.933	0.729	0.292	0.222	0.718
gbert-large [Chan et al.]	0.937	0.567	0.948	0.878	0.943	0.753	0.378	0.307	0.768
gelectra-base [Chan et al.]	0.921	0.291	0.906	0.798	0.936	0.659	0.082	0.148	0.666
gelectra-large [Chan et al.]	0.936	0.536	0.928	0.858	0.940	0.726	0.191	0.248	0.734
bert-base-german-cased [Staatsbibliothek]	0.911	0.300	0.915	0.852	0.938	0.717	0.273	0.309	0.700
gottbert-base [Scheible et al.]	0.926	0.387	0.907	0.856	0.936	0.722	0.157	0.131	0.708
GeistBERT-base [Scheible-Schmitt and Frei]	0.933	0.481	0.927	0.842	0.935	0.656	0.277	0.228	0.703
GottBERT-large [Scheible et al.]	0.936	0.419	0.927	0.811	0.932	0.751	0.288	0.154	0.724
gerturax-1 [GERTuraX]	0.930	0.433	0.932	0.836	0.949	0.714	0.176	0.281	0.740
gerturax-2 [GERTuraX]	0.935	0.444	0.937	0.855	0.943	0.716	0.232	0.214	0.744
gerturax-3 [GERTuraX]	0.934	0.428	0.937	0.843	0.946	0.713	0.184	0.222	0.740
DOSMo-7B-v0.2 [Idahl]	0.952	0.577	0.945	0.891	0.947	0.783	0.580	0.401	0.759
bueble-lm-2b [Delobelle et al.]	0.935	0.553	0.932	0.873	0.944	0.764	0.438	0.358	0.741
bloom-6b4-clp-german [Ostendorff and Rehm]	0.948	0.427	0.947	0.884	0.949	0.778	0.502	0.391	0.752
leo-hessianai-7b [Plüster]	**0.953**	0.521	0.952	0.897	0.949	0.784	0.622	0.349	0.758
Llama3-German-8B [DiscoResearch and Occiglot]	0.949	0.527	0.944	0.871	0.950	0.786	0.626	0.430	0.746
Llama-3.2-1B [Grattafiori et al.]	0.934	0.342	0.910	0.872	0.938	0.760	0.371	0.357	0.710
Llama-3.2-3B [Grattafiori et al.]	0.939	0.455	0.923	0.867	0.938	0.776	0.494	0.390	0.733
Meta-Llama-3.1-8B [Grattafiori et al.]	0.948	0.510	0.936	0.873	0.949	0.785	0.584	**0.454**	0.744
EuroBERT-210m [Boizard et al.]	0.906	0.296	0.888	0.785	0.931	0.711	0.120	0.122	-
EuroBERT-610m [Boizard et al.]	0.907	0.494	0.878	0.817	0.931	0.753	0.101	0.209	-
EuroBERT-2.1B [Boizard et al.]	0.898	0.389	0.894	0.773	0.931	0.770	0.127	0.151	-
EuroLLM-1.7B [Martins et al.]	0.940	0.496	0.934	0.896	0.949	0.764	0.528	0.410	0.728
EuroLLM-9B [Martins et al.]	0.950	0.504	0.949	0.892	0.954	0.775	0.513	0.395	0.753
xlm-roberta-base [Conneau et al.]	0.923	0.313	0.913	0.839	0.931	0.672	0.139	0.271	0.689
xlm-roberta-large [Conneau et al.]	0.933	0.465	0.933	0.830	0.937	0.733	0.101	0.029	0.730
Qwen2.5-0.5B [Team]	0.917	0.205	0.888	0.832	0.931	0.703	0.086	0.305	0.661
Qwen2.5-7B [Team]	0.940	0.441	0.930	0.865	0.947	0.772	0.498	0.308	0.728
Qwen3-0.6B [Team]	0.924	0.344	0.905	0.848	0.928	0.745	0.172	0.253	-
Qwen3-1.7B [Team]	0.932	0.390	0.904	0.866	0.931	0.768	0.191	0.393	-
Qwen3-4B [Team]	0.941	0.441	0.922	0.864	0.938	0.774	0.412	0.358	-
bloomz-560m [Muennighoff et al.]	0.914	0.242	0.883	0.780	0.929	0.638	0.109	0.288	0.622
mbart-large-50 [Liu et al.]	0.935	0.438	0.922	0.826	0.932	0.705	0.090	0.262	0.651

Table 5: Model performance across tasks and a single, seeded run (Pfister and Hotho (2024) show SuperGLEBer to produce stable results even across seeds). Where no development set was available, we split the train set 80/20 into a new train and dev set. Best models in bold. "Old Avg" contains the overall average score on the old, already existing SuperGLEBer tasks, a dash here means this model is missing from the official SuperGLEBer rankings. This table can be viewed as an addition to the offical leaderboard: `https://lsx-uniwue.github.io/SuperGLEBer-site/leaderboard_v1`. Contained in this table are the 38 models mentioned in the paper, and an additional 6 models which are not yet contained in the official SuperGLEBer. Details for these models in Table 6.

	Model	Parameters	Description
German only	LLäMmlein 120M, 1B & 7B	100M | 1.036B | 6.612B	German-only decoder models based on LLaMA2 architecture trained entirely from scratch, on the further deduplicated and filtered German portion of RedPajama V2 (Weber et al., 2024).
	ModernGBERT 134M & 1B	136M | 1.068B	German-only encoder models based on ModernBERT (Warner et al., 2025), trained from scratch, on German-only using the same dataset as the LLäMmlein models. Native context length up to 8192 tokens.
	LLäMmlein2Vec 120M, 1B & 7B	100M | 1.036B | 6.612B	Encoder versions of the LLäMmlein decoder models, created via LLM2Vec and additional context extension to up to 8192 tokens.
	DOSMo-7B	7.114B	German-only decoder built from scratch using the Mistral architecture. Trained on approximately 1 trillion German samples from RedPajama V2, MADLAD-400, OSCAR, mC4, Wikipedia, textbooks, and YouTube subtitles.
	german-gpt2	125M	German-only decoder based on GPT-2 architecture, trained from scratch on 16 GB of German subtitles CommonCrawl web crawls.
	gbert-base/-large	110M | 337M	German-only encoder models based on BERT architecture trained on 163.4 GB of German text, consisting of OSCAR, OPUS, Wikipedia, legal documents.
	bert-base-german-cased	109M	German-only encoder model based on BERT architecture trained from scratch on a 12GB dataset of wikipedia, legal documents and news.
	gelectra-base/-large	110M | 335M	German-only encoder family based on Electra architecture trained from scratch, on the same 163.4GB dataset as gbert-base and gbert-large.
	gottbert-base	126M	German-only encoders based on the RoBERTa architecture, trained on 145 GB of text from OSCAR, Wikipedia and a book corpus.
	GottBERT-large	358M	German-only encoder model based on RoBERTa architecture, trained on 121GB of filtered texts sourced from the first released OSCAR dataset.
	GeistBERT-base	126M	German-only encoder model family building on GottBERT trained on 1.3 trillion tokens sourced from OSCAR23, OPUS, MC4, German Wikipedia, OpenLegalData, Europarl, EUbookshop, ECB, and EuroPat, OpenSubtitles and TildeMODEL.
	gerturax-1/-2/-3	135M	German-only encoder family based on Electra architecture trained on 147GB to 1.1TB of texts from the CulturaX corpus.
German Finetune	gerpt2(-large)	125M | 776M	German-only GPT-2-style decoder family trained from scratch/ with weights initialized from the english GPT2 model on data from the CC-100 corpus.
	bloom-6b4-clp-german	6.251B	Decoder model preinitialized from multilingual bloom 7b checkpoint transfer learned to German using CLIP-Transfer.
	bueble-lm-2b	2.072B	Decoder-only model preinitialized from multilingual gemma checkpoint trans-tokenized to German. Trained on 3.8B tokens from Occiglot-FineWeb.
	leo-hessianai-7b	6.611B	Decoder-only model preinitialized from LLaMA2 checkpoint finetuned using German texts mostly sourced from OSCAR.
	Llama-German-8b	7.508B	Decoder-only model preinitialized from the multilingual Llama3 8B and finetuned on German via continual pretraining on 65 billion tokens, sourced from occiglot-fineweb-0.5.
Multilingual	EuroBERT-210m/-610m/-2.1B	212M | 609M | 2.110B	Multilingual encoder family based on EuroBERT architecture, supporting 15 languages.
	EuroLLM-1.7B/-9B	1.396B | 8.633B	Multilingual decoder family based on LLaMa architecture, supporting 35 languages.
	xlm-roberta-base/-large	278M | 561M	Multilingual encoder model family pretrained on 2.5TB of CommonCrawl data in 100 languages.
	mbart-large-50	612M	Multilingual encoder model trained on 50 languages, including German using a sequence-to-sequence translation objective.
	bloomz-560m	560M	Multilingual decoder model initially trained on a 1.5 TB corpus spanning 45 natural and 12 programming languages, followed by supervised multitask fine-tuning across multiple languages.
Other Models	Qwen2.5-0.5B/-7B	495M | 7.073B	Multilingual decoder model family trained on 18 trillion tokens supporting 29 languages.
	Qwen3-0.6B/-1.7B/-4B	597M | 1.722B | 4.025B	Multilingual decoder model family trained on approximately 26 trillion tokens supporting over 100 languages and dialects
	Llama-3.2-1B/-3B	1.237B | 3.215B	Multilingual decoder only model family trained on up to 9 trillion tokens of multilingual text and code.
	Meta-Llama-3.1-8B	7.508B	Multilingual decoder only model trained on over 15 trillion tokens including multilingual text.

Table 6: Short description of the models mentioned in Table 5. Parameter counts reflect total number of parameters after applying (Q)LoRA.



(a) F-Tag Performance

(b) F-Class Performance

(c) HC-c2a Performance

(d) HC-dbo Performance

(e) HC-vio Performance

(f) SE-Class Performance

(g) SE-Reg Performance

(h) LLMs4s Performance

Figure 3: Models in PCA space colored by performance on individual GermEval 2025 tasks. Each subplot shows the same PCA projection of models but colored by their performance on different tasks: (a) Flausch-Erkennung tagging, (b) Flausch-Erkennung classification, (c-e) Harmful content detection subtasks, (f) SustainEval classification, (g) SustainEval regression, and (h) LLMs4Subjects classification. The visualization reveals task-specific performance patterns and model clustering behaviors across different evaluation metrics. Model placement is identical to fig. 4.

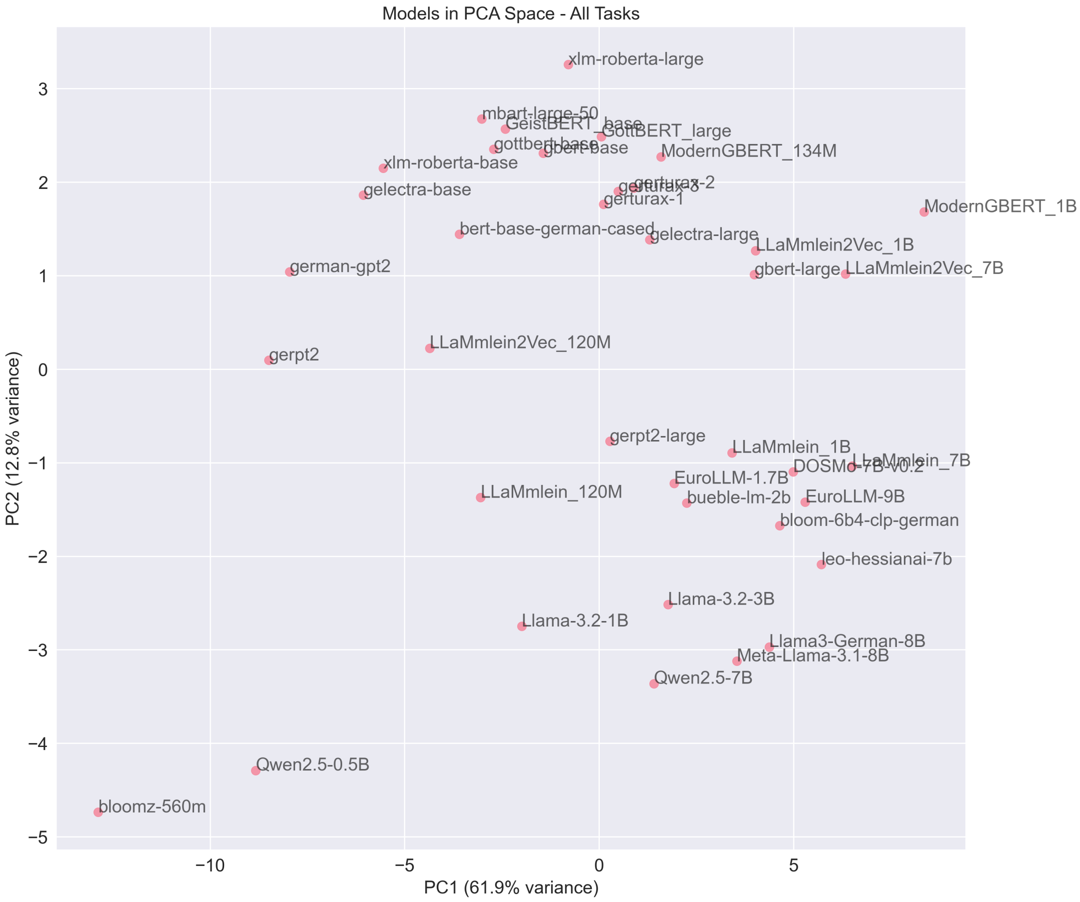


Figure 4: PCA visualization of model performance. Models positioned in PCA space based on their performance vectors across all tasks, showing clustering patterns of similar performing models. The visualization uses the complete set of 37 tasks (29 existing + 8 new GermEval 2025 tasks). Inspired by this plot, we also plotted the inverse loadings in Figure 5.

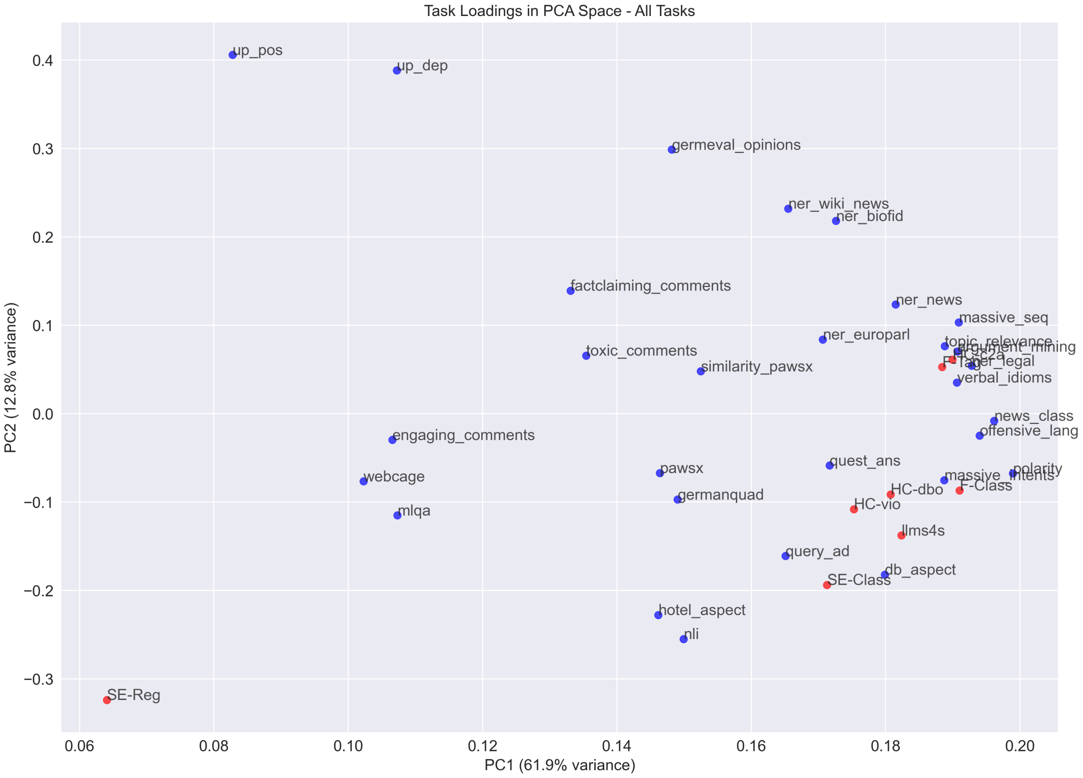
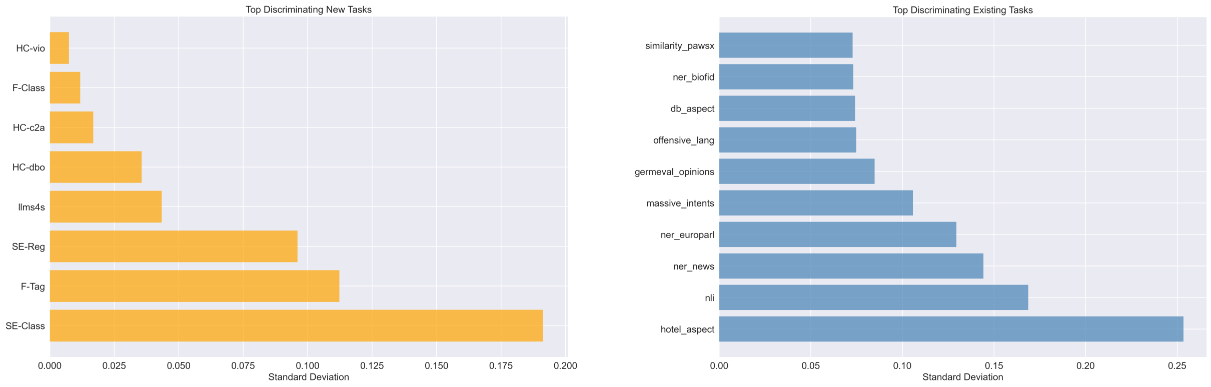


Figure 5: PCA visualization of task performance. Tasks positioned in PCA space based on each model's performance showing clustering patterns of tasks, where models perform similar. The visualization is based on the complete set of models, older SuperGLEBer tasks are blue, new ones from GermEval 2025 are red.



(a) New GermEval 2025 tasks

(b) Existing SuperGLEBer tasks

Figure 6: Task discrimination analysis showing standard deviation of model performance across tasks. (6a) New GermEval 2025 tasks ranked by their ability to discriminate between models. (6b) Top 10 most discriminating existing SuperGLEBer tasks. Higher standard deviation indicates better discrimination between model capabilities, making these tasks valuable for benchmark evaluation.