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

Julia Wunderle Jan Pfister Andreas Hotho

{lastname}@informatik.uni-wuerzburg.de

Contribution

We have extended SuperGLEBer, a German language understanding benchmark, to achieve two main goals: (1) to participate in all shared tasks by applying simple methods across 38 diverse LLMs (100M–9B parameters); and (2) to contextualize the tasks within the broader German NLU task landscape.

GermEval 2025 Tasks

Candy Speech Detection

Detects „candy speech", i.e. positive attitudes in YouTube comments via classification (Flausch-Class) and span detection (Flausch-Span).

Harmful Content Detection

Detects harmful content in social media posts via 3 classification tasks: C2A DBO and VIO.

LLMs4Subjects

Multi-label classification task assigning library records to 28 subject domains. We only benchmarked subtask 1.

Understanding Sustainability Reports

Analyses sustainability reports via classification of 20 DNK criteria (Sustain-Class) and verifiability rating (Sustain-Reg).

Methodology

Task Framework: SuperGLEBer[1]

As backbone for our task implementation, we utilize the German NLU Benchmark SuperGLEBer

- 29 tasks covering four task types:
  - Classification
  - sequence tagging,
  - sentence similarity
  - QA
- Framework: flair and SentenceTransformers,
- Hyperparameters:
  - batch size 8
  - learning rate 5e-5
  - 5 epochs
  - Same seed
- QLoRA where supported, else LoRA
- 38 mostly German (L)LMs evaluated

Implementation

Classification: Add a linear layer on the [CLS] token representation.
Example input: "*ihr seid die Besten*" – Output class: *yes*

Tagging: Add a linear layer on each token's representation to predict its class (following the BIO-label scheme).

Token	Label
Ihr	*B-affection_declaration*

Regression: Apply RMSNorm, then a linear layer on the hidden states, followed by a sigmoid activation

- Input: "*Prävention Über das Kerngeschäft „Versicherung" hinaus…*" and "*Die [ORG] arbeitet hier eng mit den relevanten Institutionen…*"
  – Output value: *0.667*

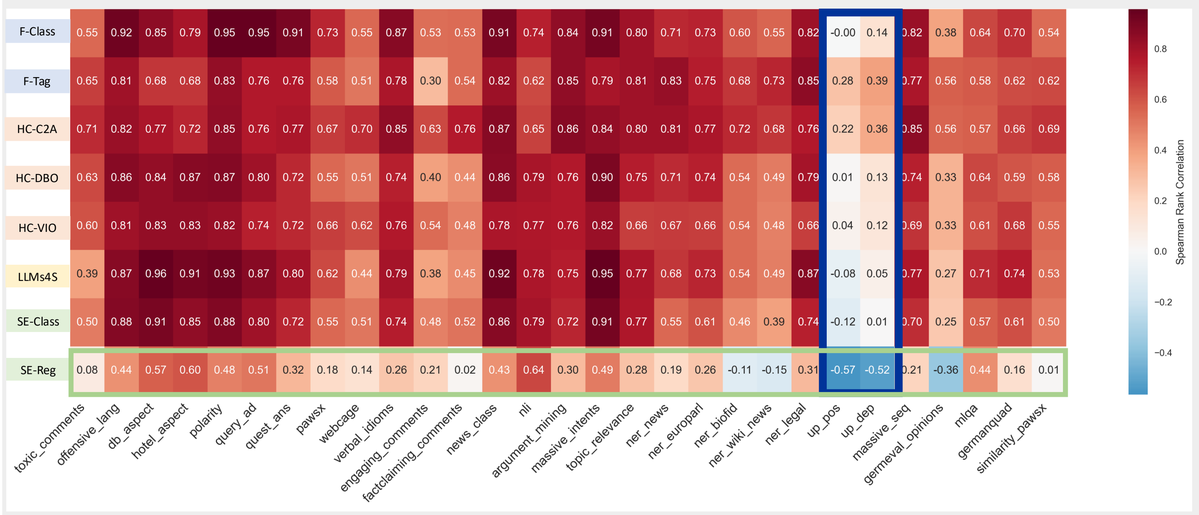
Results

Task	ModernGBERT 1B[2]	LLäMmlein 7B[3]	Llama-3.1-8B	Placement
Flausch-Class	**0.948**	0.947	**0.948**	3
Flausch-Tag	**0.662**	0.657	0.510	6
HC-C2A	**0.953**	0.943	0.936	1
HC-DBO	0.897	**0.900**	0.873	2
HC-VIO	**0.952**	0.947	0.949	1
LLMs4S	**0.787**	0.780	0.785	-
Sustain-Class	**0.659**	0.614	0.584	4
Sustain-Reg	0.118	0.350	**0.454**	1

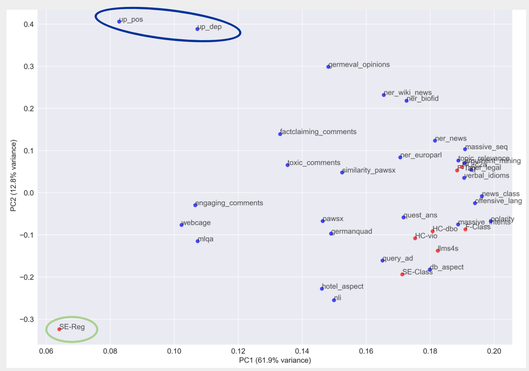
We submitted the three best performing models on the development dataset (depicted in table above) for the official evaluation.

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

As we extended the SuperGLEBer benchmark to new tasks, we examined how model performance on these tasks relates to existing tasks. Using Spearman rank correlation, the figure shows a matrix of correlations between new tasks (rows) and existing tasks (columns).

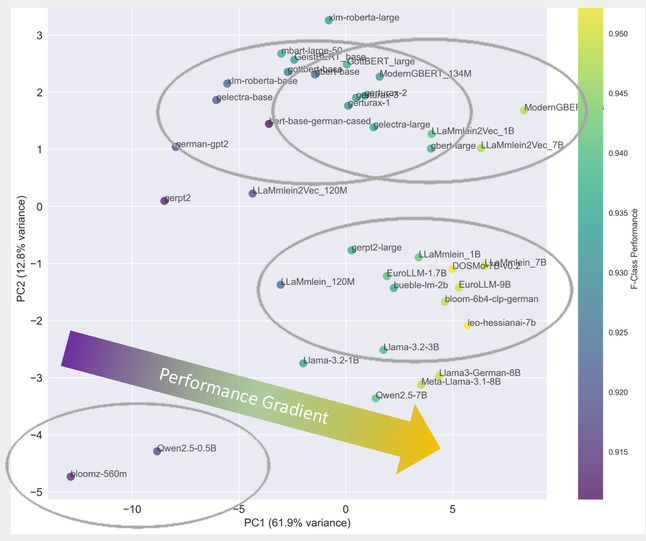


Most **new tasks strongly correlate (0.7-0.9) with existing tasks** suggesting similar performance patterns. In contrast, SE-Reg and the existing syntactic up-dep and up-pos, show weak/negative correlations, highlighting them as a distinct challenge that deviates from existing performance patterns. PCA embeddings further highlight these tasks as out-of-distribution.



How does task performance help in discriminating between model quality?

After comparing old and new tasks, we investigate whether a small subset of tasks is sufficient to serve as a model description for judging overall model quality. To explore this, we reverse the PCA analysis from the bottom left: task performance is treated as model embedding features, and models are visualized according to subtask performance (e.g., F-Class with similar patterns for other tasks). We observe clear patterns:

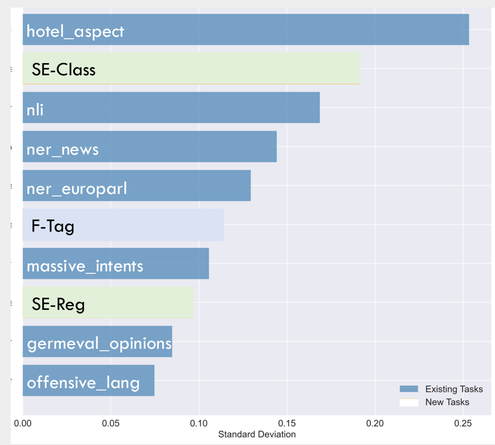


- Small multilingual models (bloomz-560m and Qwen 0.5B): **lower-left**
- Decoder models (e.g. bueble-lm-2b, EuroLLM, leo-hessianai-7b): **lower-right**
- Small German-focused encoder (e.g., gberts and gelectra): **upper-center**
- High-performing models (i.e. large LLäMmlein, ModernGBERTs): **upper-right**

Model architecture and language focus create distinct performance patterns.

This reveals a clear gradient in the PCA space along which the performance evolves rather monotonously, indicating that a **few tasks drive most of the discriminative signal.**
To identify highly discriminative tasks, we compute the standard deviation of model performance per task: **high deviation indicates strong model discrimination**.

- Three new tasks, **SE-Class** **F-Tag** **SE-Reg** achieve a standard deviation of 0.1 or higher, placing them among the top 10 **most discriminative tasks** for distinguishing model quality.
- A commonality of all tasks is a large number of target classes: i.e. existing tasks like hotel_aspect (15) and massive_intents (60), as well as new SE-Class (20) and Flausch-Tagging (21).



Task	Correlation	MAD	Exact Rank matches	Models in Ties
Massive_intents	*0.969*	*2.00*	*3/38*	*10*
Offensive_lang, db_aspect, ner_news, germanquad, HC-DBO	*0.994*	*0.69*	*22/38*	*0*
37 tasks	*1.00*	*0.00*	*38/38*	*0*

This indicates that a smaller set of highly discriminative tasks can capture most model differences, enabling a more compact benchmark.

[1] Pfister, Jan and Hotho, Andreas (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 (NAACL-HLT 2024)*. Mexico City, Mexico.
[2] Ehrmanntraut, Anton, Wunderle, Julia, Pfister, Jan, Jannidis, Fotis, and Hotho, Andreas (2025). ModernGBERT: German-only 1B Encoder Model Trained from Scratch. *arXiv preprint*.
[3] Pfister, Jan and Wunderle, Julia and Hotho, Andreas (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)*. Vienna, Austria