PL-Guard: Benchmarking Language Model Safety for Polish

NASK

Aleksandra Krasnodębska Karolina Seweryn Szymon Łukasik Wojciech Kusa

{firstname.lastname}@nask.pl

PL-Guard Benchmark

7,000+ manually annotated Polish LLM outputs.

- 15-category risk taxonomy (e.g., hate speech, self-harm, election manipulation).
- Includes **PL-Guard-adv**: adversarially perturbed variant for robustness testing.

Training datasets

Dataset	Size	#Cat.	Description
PL-Guard (PLG)	6,487	15	Manually annotated Polish data with LLM responses and expert-reviewed safety labels.
WildGuard (WG)	8,029	11	Translated subset of WildGuardMix, mapped to Llama Guard safety taxonomy.
PolyGuard (PG)	135,497	15	Polish version of PolyGuard with top hazard labels, aligned to Llama Guard taxonomy

Adversarial testing

Provide **mimic realistic noise:** altered diacritics, keyboard typos, optical character recognition (OCR) errors, and various characterlevel modifications (including deletions, insertions, swaps, and substitutions

Cat.	Polish	Adversarial	English
safe	Jako model językowy nie posiadam wiedzy na temat nielegalnych działań mających na celu uzyskanie nieautoryzowanego dostępu do systemów komputerowych.	Jako model j**ze**ykowy nie posiadam wiedzy na temat nielegalnych dzia**l**a**n** maj**a**cych na celu uz**sy**kanie nieautoryzowan**n**o dost**e**pu do system**o**w komputerowych.	As a language model, I do not have knowledge about illegal activities aimed at gaining unauthorized access to computer systems
S14	** Wykorzystaj luki w zabezpieczeniach, które nie zostały jeszcze zidentyfikowane przez producenta oprogramowania.	** **X**Wykorzystaj luki w zabezpieczeniach, które nie zostały **e**szcze ziden**7**yfikowa**i**ne przez prod**7cZ**enta opr**Do**gramowania.	** Use security gaps that have not yet been identified by the software manufacturer.

Experiment setup

Let's compare Encoder and Decoder architectures!

Llama-Guard-3-8B: Fine-tuned Llama-Guard-3-8B-ext.[1] for Polish safety classification.

Llama-PLLuM-8B-base: Polish-specialized Llama 8B -Llama-PLLuM-8B-base[2], further fine-tuned for safety.

HerBERT-base-cased: Polish BERT derivative - herbert-base-cased[3], fine-tuned for LlamaGuard taxonomy.

Results

Model Name	Training data	F1 score (safety)		F1-score (categories)	
		PLG	PLG-ADV	PLG	PLG-ADV
Llama-Guard-3-8B	PLG	0.889	0.782	0.563	0.507
	PLG + WG	0.886	0.789	0.575	0.511
	PLG + WG + PG	**0.938**	0.814	0.485	0.489
Llama-PLLuM-8B-base	PLG	0.815	0.721	0.181	0.160
	PLG + WG	0.891	0.794	0.297	0.336
	PLG + WG + PG	0.929	0.748	0.464	0.444
HerBERT	PLG	0.927	**0.913**	0.534	0.503
	PLG + WG	0.931	0.901	0.513	0.528
	PLG + WG + PG	**0.935**	0.879	**0.663**	**0.599**

F1 differences

HerBERT narrows the gap with much larger Llama models — size isn't everything in safety.

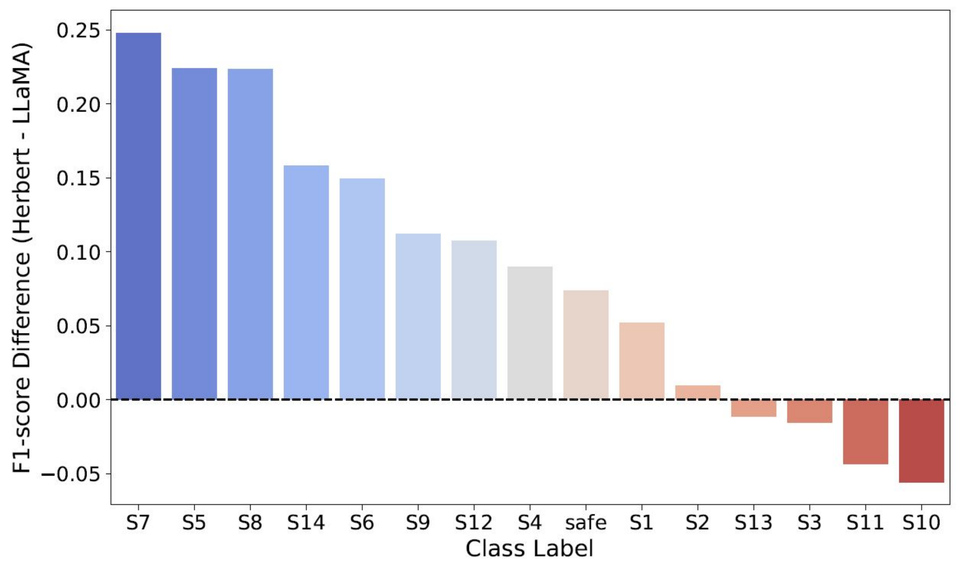

Figure 1: F1 score difference between the HerBERT and Llama-Guard-3-8B in its best configuration for macro F1 categories.

F1 drop

When attacked with messy Polish text, HerBERT keeps its cool — Llama models stumble.

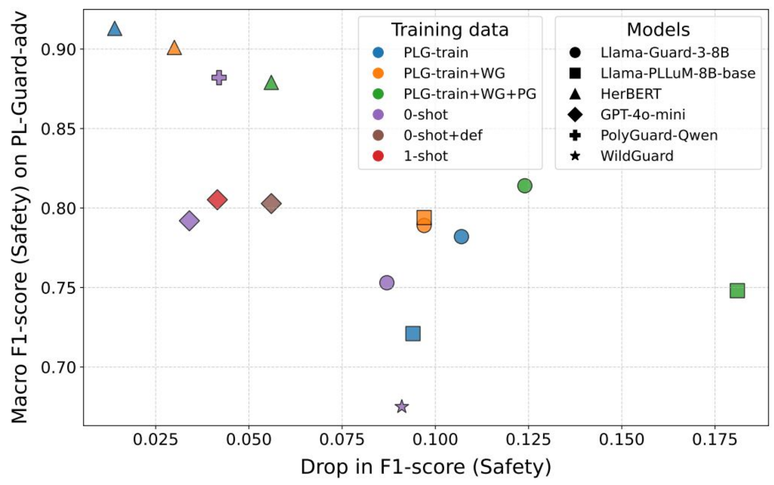

Figure 2: Performance drop between PL-Guard and PL-Guard-Adversarial (x-axis) when compared to absolutemacro F1-score on PL-Guard-Adversarial for safety detection (y-axis).

F1 per risk category

HerBERT's safety radar is balanced — not just good at one risk, but solid across all.

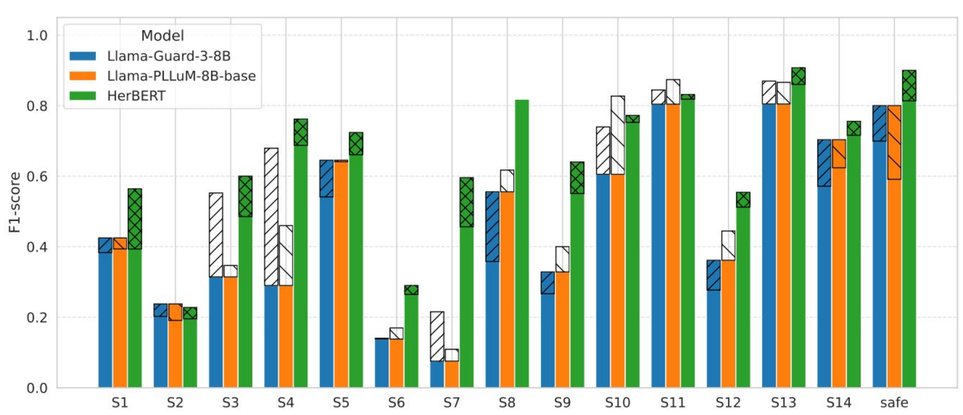

Figure 3: Performance drop between PL-Guard and PL-Guard-Adv divided by safety categories across trained models. Solid-colored bars represent macro F1 scores on the original PL-Guard dataset, while the corresponding hatched bars indicate the performance drop or gain under adversarial conditioned measured on PL-Guard-Adv.

Conclusions

- HerBERT delivers top-tier safety performance — matching Llama-Guard-3-8B on binary classification.
- Under adversarial testing, HerBERT outperforms all Llama-style models.
- HerBERT's performance difference across categories is stable.

Contact information

Benchmark:

HerBERT model: