Toward Generalized Cross-Lingual Hateful Language Detection with Web-Scale Data and Ensemble LLM Annotations

Dang H. Dang, Jelena Mitrović, Michael Granitzer
University of Passau, Chair of Data Science
{hai-dang.dang, jelena.mitrovic, michael.granitzer}@uni-passau.de

Abstract

We study whether large-scale unlabelled web data and LLM-based synthetic annotations can improve multilingual hate speech detection. Starting from texts crawled via OpenWebSearch.eu (OWS) in four languages (English, German, Spanish, Vietnamese), we pursue two complementary strategies. First, we apply *continued pre-training* to BERT models by continuing masked language modelling on unlabelled OWS texts before supervised fine-tuning, and show that this yields an average macro-F1 gain of approximately 3% over standard baselines across sixteen benchmarks, with stronger gains in low-resource settings. Second, we use four open-source LLMs (`Mistral-7B`, `Llama3.1-8B`, `Gemma2-9B`, `Qwen2.5-14B`) to produce synthetic annotations through three ensemble strategies: mean averaging, majority voting, and a LightGBM meta-learner. The LightGBM ensemble consistently outperforms the other strategies. Fine-tuning on these synthetic labels substantially benefits a small model (`Llama3.2-1B`: +11% pooled F1), but provides only a modest gain for the larger `Qwen2.5-14B` (+0.6%). Our results indicate that the combination of web-scale unlabelled data and LLM-ensemble annotations is the most valuable for smaller models and low-resource languages.

Keywords: OpenWebSearch.eu, OWI, LLM ensembling, synthetic hate speech annotation, BERT continued pre-training, cross-lingual generalization

1. Introduction

A central bottleneck in building robust detectors for hateful and offensive language is the scarcity of high-quality labelled training data (Vidgen and et al., 2020; Fortuna and Nunes, 2018). While raw web text can be collected at scale, annotating it remains costly (Ross and et al., 2017) and human annotators inevitably introduce subjective biases (Caselli et al., 2021). Recent large language models (LLMs) have demonstrated strong performance on hate-speech benchmarks (Guo et al., 2024) and have therefore been explored as automated annotators (Hartvigsen et al., 2022). However, existing work in this space remains narrow in scope, typically covering a single language and lacking rigorous comparison with models trained on human labels.

Large web crawls such as OpenWebSearch.eu (OWS) (Granitzer et al., 2024) and OpenWebIndex (OWI) (Hendriksen et al., 2024) make billions of multilingual pages available, yet how to best leverage them for hate-speech detection is an open question. We address this gap along two axes:

1. We investigate whether *domain-adaptive continued pre-training* on large unlabelled OWS corpora improves the downstream performance of BERT-family models.
2. We study whether *ensemble-based LLM annotation* can replace or supplement human labelling for multilingual hate-speech detection.

Terminology. Throughout this paper, **continued pre-training** denotes an additional masked-language-modelling adaptation step applied to a general-purpose pre-trained BERT using domain-relevant but *unlabelled* OWS texts. This precedes the subsequent supervised **fine-tuning** on hate-speech labelled data. The term follows the domain-adaptive pre-training (DAPT) paradigm of Gururangan et al. (2020). Figure 1 summarises the end-to-end pipeline.

Concretely, this work is guided by two research questions:

- **RQ1:** To what extent can large-scale unlabelled web data improve multilingual hate-speech detection via continued pre-training of BERT?
- **RQ2:** How effective are LLM ensemble strategies for synthetic annotation in improving detection performance and cross-lingual generalisation?

Code, fine-tuned models, OWS data, and labelled splits are available at `https://github.com/HaiDangDang/HateOWS`.[1]

2. Related Work

Domain-specific continued pre-training for hate speech. Gururangan et al. (2020) established that *domain-adaptive pre-training* (DAPT), continuing masked language modelling on in-domain text before fine-tuning, consistently improves downstream performance. HateBERT (Caselli et al., 2021) applied DAPT to abusive language using

[1] Repository includes all training scripts, annotation prompts, and evaluation code.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 549–559
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

1.5M English Reddit posts, while Belay et al. (2025) extended it multilingually with AfroXLMR-Social across 19 African languages (F1 gains of 1–30%). None of these works use web-crawled data at OWS scale or cover our four-language setting.

Cross-lingual hate speech detection. Dmonte et al. (2024) showed via the GenOffense benchmark that single-dataset models generalise poorly, though their study was English-only. Usman et al. (2025) explored translation-based LLM approaches for multilingual detection, highlighting the difficulty of cross-lingual transfer. We extend these evaluations to four languages and sixteen benchmarks.

Ensemble approaches. Kashif et al. (Kashif et al., 2023) and Daouadi et al. (Daouadi et al., 2024) built multimodal and Arabic BERT ensembles respectively for downstream classification. Our work instead applies the ensemble paradigm to the *annotation* phase, combining token-probability outputs from four LLMs.

LLM-based annotation and synthetic data. Zhu et al. (2023) found ChatGPT achieves 60.9% average accuracy relabelling social-computing datasets, while Giorgi et al. (2025) showed that LLM annotation biases differ substantially from human biases, and Piot et al. (2025) demonstrated that LLMs can reliably reproduce model-ranking patterns as scalable proxy evaluators. Our work *ensembles* four open-source LLMs for annotation and compares three aggregation strategies, measuring how synthetic labels benefit models of varying capacity.

3. Methodology and Setup

Figure 1 provides a high-level overview of the pipeline. The following subsections detail each component.

3.1. OpenWebSearch.eu Data Collection

We used OWS (Granitzer et al., 2024) and OWI (Hendriksen et al., 2024) to collect large-scale web texts in English, German, Spanish, and Vietnamese. To increase the proportion of conversational and user-generated content, we filtered the OWS index to retain only URLs whose path contained at least one of the keywords *thread*, *forum*, *reply*, *post*, *status update*, or *quote*. We additionally required retrieved pages to conform to one of ten conversational schema.org types (e.g. `DiscussionForumPosting`, `SocialMediaPosting`, `Comment`; full list in Appendix A.5), to comply with structured human discourse. **No hate-specific keyword filtering was applied**: the unlabelled texts serve purely for domain-adaptive continued pre-training, not as a labelled hate-speech corpus.

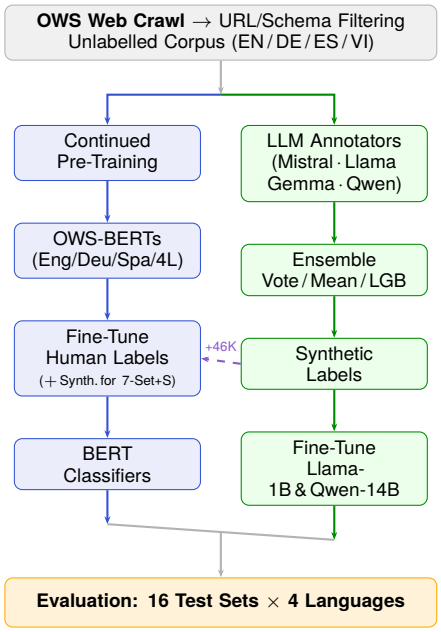

Figure 1: End-to-end pipeline. **Left (BERT track)**: OWS texts are used for domain-adaptive continued pre-training; models are then fine-tuned on human labels. **Right (LLM track)**: four LLMs generate token-probability annotations; three ensemble strategies produce synthetic labels used to fine-tune smaller LLMs. The dashed arrow shows the cross-track path: ≈46K `Qwen2.5-14B`-annotated texts augment the BERT training data (7-Set + Synth.). Both tracks are evaluated on the same sixteen test sets. LGB = LightGBM.

In total, we collected approximately 6M English, 3M German, 3M Spanish, and 500K Vietnamese texts. We sub-sampled differently per task: $\sim$3M mixed-language texts for BERT continued pre-training and $\sim$2M texts for LLM-based annotation.

3.2. Human-Labelled Datasets

We used sixteen publicly available hate-speech datasets (Table 1) spanning four languages. The datasets vary in annotation guidelines, topic coverage, and label granularity.

Label mapping. To enable unified training and evaluation, we mapped all annotations to a binary **Hate**/**Neutral** scheme. Three datasets (*HateXplain*, *AHSD*, *ViHSD*) originally distinguish *Hate* from *Offensive*; both were merged into **Hate**. The *AbusEval* labels *Implicit Abusive* and *Explicit Abusive* were similarly collapsed. The remaining twelve datasets already use binary labels.

We acknowledge that merging qualitatively dif-

Dataset	Train		Test			
	Inst#	% Hate	Inst#	% Hate	Lang	Ref
HateXplain	15,299	0.59	3,846	0.59	eng	(Mathew et al., 2022)
Sexism	10,904	0.13	2,632	0.13	eng	(Samory et al., 2021)
Covid	1,282	0.19	971	0.20	eng	(He et al., 2022)
US_election	1,283	0.12	1,117	0.13	eng	(2021)
HateEval-eng	9,000	0.42	1,000	0.43	eng	(Basile et al., 2019)
AbusEval	13,240	0.21	860	0.21	eng	(Caselli et al., 2020)
AHSD	21,783	0.83	3,000	0.83	eng	(Davidson et al., 2017)
Eng-All	72,791	0.49	13,426	0.19	eng	
GermEval21	2,071	0.33	2,085	0.37	deu	(Risch et al., 2021)
GermEval19	9,698	0.33	2,507	0.34	deu	(Struß et al., 2019)
GermEval18	5,009	0.34	3,532	0.34	deu	(Wiegand, 2019)
HASOC	2,373	0.28	526	0.25	deu	(Mandl et al., 2019)
Gahd	8,797	0.42	2,198	0.43	deu	(Goldzycher et al., 2024)
Deu-All	27,948	0.36	10,848	0.09	deu	
ViHSD	8,061	0.17	2,672	0.18	vie	(Luu et al., 2021)
Haternet	4,794	0.26	1,205	0.25	spa	(Basile et al., 2019)
HateEval-spa	5,309	0.41	1,286	0.43	spa	(2019)
Chileno	7,572	0.06	1,928	0.06	spa	(2022)
Spa-All	17,675	0.22	4,419	0.02	spa	

Table 1: Sixteen human-labelled hate-speech datasets across four languages. Inst# = number of instances; % Hate = fraction labelled hateful. Datasets in **bold** form the *7-Set* training subset.

ferent categories (hate, offensiveness, and abuse) into a single positive class is a simplification. It may inflate apparent recall on datasets whose original positive class was broader (e.g., *AbusEval*). Readers should keep this caveat in mind when interpreting cross-dataset comparisons.

Training configurations. We defined five configurations: (i) *7-Set*—seven datasets (bold in Table 1), excluding all Spanish data, used for `Llama3.2-1B` (Dubey et al., 2024) and `Qwen2.5-14B` (et al., 2025) fine-tuning; (ii) *Eng*, (iii) *Deu*, (iv) *Spa*—language-specific subsets for BERT experiments; and (v) *16-Mix*—all sixteen datasets combined. For evaluation, we used the test portions of all sixteen datasets.

3.3. BERT Continued Pre-Training and Fine-Tuning

We continually pre-trained four BERT variants on OWS texts:

- `OwsSpa`: 1.1M Spanish texts (≈110M tokens).
- `OwsDeu`: 650K German texts (≈52M tokens).
- `OwsEng`: 1.5M English texts (≈119M tokens).
- `Ows4L`: 2.78M multilingual texts—43% English, 32% German, 17% Spanish, 6.7% Vietnamese (≈136M tokens).

Each model was then *supervised-fine-tuned* on its corresponding training configuration. `Ows4L` was evaluated under three settings: *7-Set*, *16-Mix*, and *7-Set + Synthetic* (adding ≈46K `Qwen2.5-14B`-annotated OWS examples; see Section 3.4).

3.4. LLM Ensemble Annotation

3.4.1. Annotation Method

We collected token-level output probabilities from four instruction-tuned, 4-bit quantised LLMs via the Unsloth framework (UnslothAI, 2023): `Mistral-7B` (Jiang and AI, 2023), `Llama3.1-8B`, `Gemma2-9B` (Riviere et al., 2024), and `Qwen2.5-14B`. A simple zero-shot prompt was used to classify each OWS text as *Hate* or *Neutral*. From the raw probabilities, three ensemble strategies determine the final label:

1. **Majority Voting (`Vote`):** Each model's probability is thresholded to a hard label; a text is marked *Hate* if at least two of the four models vote for it.
2. **Mean Averaging (`Mean`):** For each class, the average probability across all models is computed; the class with the higher mean is assigned.
3. **LightGBM Meta-Learner (`LGB`):** A LightGBM classifier (Ke et al., 2017) is trained on the eight-dimensional probability vectors (two classes × four models) using the seven human-labelled training sets as supervision. Unlike `Vote` and `Mean`, which treat all models equally, `LGB` learns to weight each annotator differentially based on its reliability, and discovers confidence thresholds that align with human judgements. This discriminative calibration explains its consistent advantage (Section 4.2).

Label imbalance in synthetic data. A notable characteristic of the resulting annotations is severe class imbalance: `Neutral` comprises over 97% of annotated OWS texts across all three ensemble methods. This reflects the web's base rate (most discussion-forum content is not hateful) and directly constrains the minority-class detection ability of models trained on these sets.

3.4.2. LoRA Fine-Tuning of LLMs

We used Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tune `Llama3.2-1B` and `Qwen2.5-14B`. Human-labeled training used the *7-Set* (excluding Spanish). The synthetic subset contains **240,647** texts: 125,617 German, 108,375 English, and 6,655 Vietnamese. After ensemble annotation, `Vote` assigned 4,717 texts as hate, `Mean` 3,994, and `LGB` 3,122—confirming the pronounced class imbalance.

3.4.3. Computational Resources

All experiments ran on a single NVIDIA A6000 (48 GB). Table 2 summarises the cost per task.

Task / Model	Time	VRAM	Params
Annotation (240K, batch 128)			
`Mistral-7B`	4 h	13 GB	–
`Llama3.1-8B`	4 h 45 m	23 GB	–
`Gemma2-9B`	6 h 30 m	43 GB	–
`Qwen2.5-14B`	7 h	42 GB	–
LoRA Fine-Tuning			
`Llama3.2-1B`	<2 h	<5 GB	5.6M (0.45%)
`Qwen2.5-14B`	≈12 h	<15 GB	34.4M (0.23%)

Table 2: Cost per task on a single NVIDIA A6000 (48 GB).

4. Results

4.1. RQ1: OWS Continued Pre-Training with BERT

Table 3 reports macro-F1 across all sixteen test sets. All four OWS-continually-pre-trained models outperform both BERT and HateBERT on every per-language average and on the overall 16-set average under multilingual training (7-Set Mix, 7-Set + Synth., 16-Mix). Monolingual OWS models additionally beat both baselines on their target-language test sets.

4.1.1. Single-Language Continued Pre-Training

In-language performance. When fine-tuned on language-matched data, all three monolingual OWS models outperform both BERT and HateBERT on their target language (shaded cells in Table 3). `OwsEng` reaches an English average of 86.9% (+0.3% over BERT, +0.8% over HateBERT), `OwsDeu` a German average of 75.0% (+3.1% over BERT, +5.7% over HateBERT), and `OwsSpa` a Spanish average of 75.5% (+0.5% over BERT, +3.0% over HateBERT). HateBERT's English Reddit pre-training offers no advantage on German or Spanish, explaining the large OWS margins in those languages.

Cross-language effects and overall average. OWS pre-training also produces cross-lingual gains: `OwsDeu`, trained on only German, boosts English-test performance by +11.2% over BERT and Vietnamese by +6.4%, likely because the German OWS crawl contains multilingual web content. On the 16-set overall average, `OwsDeu` (66.9%) and `OwsEng` (72.3%) both surpass BERT and HateBERT. The sole exception is `OwsSpa` (43.6%), which leads on its three Spanish benchmarks but collapses on the thirteen non-Spanish test sets because neither its pre-training data nor its fine-tuning set contains other languages.

Where HateBERT outperforms OWS models. HateBERT's advantage is limited to one scenario: when fine-tuned on non-English data only, it retains strong English performance from its Reddit pre-training (e.g., English average of 55.4% under the Spanish config versus 35.5% for `OwsSpa`). In all other configurations the OWS models match or exceed HateBERT.

4.1.2. Multilingual Continued Pre-Training

`Ows4L` outperforms both BERT and HateBERT on all four per-language averages and on the overall average under every multilingual training configuration. Under *7-Set Mix*, the gain over BERT is +1.6% (Eng), +1.9% (Deu), +1.8% (ViHSD), and +1.9% (Spa), with an overall average of 73.1% versus 71.6% for BERT and 70.7% for HateBERT. The benefit is most pronounced in low-data and low-resource settings: ViHSD improves by +6.6% over HateBERT, and AHSD by +3.4%. Under *16-Mix*, the gap narrows but `Ows4L` still leads (82.1% versus 81.5% BERT and 79.9% HateBERT), achieving the best score on 6 of 16 test sets. These results indicate that OWS continued pre-training is *most valuable when supervised data is limited*.

Benefit of adding synthetic data. Comparing *7-Set + Synth.* against *7-Set Mix* isolates the effect of the ≈46K `Qwen2.5-14B`-annotated OWS texts. All three models improve on the overall average: BERT rises from 71.6% to 75.2% (+3.6%), HateBERT from 70.7% to 74.7% (+4.0%), and `Ows4L` from 73.1% to 75.7% (+2.6%). The gains concentrate on English (Eng Avg: +7.0% to +8.5% across models), whereas German regresses (-1.4% to -3.6%) and Vietnamese drops slightly (-1.2% to -2.7%). Notably, `Ows4L` still leads under this setting (75.7% versus 75.2% BERT and 74.7% HateBERT), but the gap narrows compared to 7-Set Mix, suggesting that synthetic data partially compensates for the advantage that OWS continued pre-training provides.

4.2. RQ2: LLM Ensemble Annotations

4.2.1. Small-Scale Models: Llama3.2-1B and BERT

Table 4 shows that synthetic data substantially benefits smaller models. `Llama3.2-1B` gains 2–11% pooled F1 over the human baseline depending on the ensemble strategy. The strongest configuration is `H+LGB` (human + LGB-labelled data), reaching 65.4% overall F1 (+10.6%), with per-dataset gains up to +25% on AHSD and +19% on Gahd (Table 8).

English and German. For the three synthetic-only strategies, English and German pooled F1

BERT Models

Train Set	Spanish (Spa)			German (Deu)			English (Eng)			7-Set Mix			7-Set Mix + Synth.			16-Set Mix		
Models	BERT	HateBERT	OwsSpa	BERT	HateBERT	OwsDeu	BERT	HateBERT	OwsEng	BERT	HateBERT	Ows4L	BERT	HateBERT	Ows4L	BERT	HateBERT	Ows4L
HateXplain	30.3	54.7	28.9	52.6	55.5	59.7	77.7	77.5	77.9	77.7	77.6	77.6	**78.9**	77.8	78.4	78.0	77.1	77.5
Sexism	47.7	49.9	46.4	51.8	56.6	56.5	85.1	84.9	**86.7**	85.9	85.8	85.4	85.2	84.9	85.1	84.5	85.3	86.4
Covid	44.6	48.2	44.6	54.6	52.9	63.1	77.5	75.2	79.5	**81.1**	79.6	81.0	77.4	76.8	76.9	78.1	76.3	78.8
US_election	46.6	47.9	47.4	46.6	51.1	46.6	66.5	67.7	**70.4**	64.2	63.2	68.8	64.4	62.4	64.7	67.5	68.5	68.7
HateEval-eng	37.4	54.0	36.4	47.2	56.4	58.8	**77.3**	75.6	75.9	58.7	60.5	60.0	60.2	62.2	61.4	77.0	76.3	76.2
AbusEval	44.2	53.5	44.2	48.6	58.6	51.4	67.0	67.1	70.7	52.1	55.1	52.5	51.7	56.7	55.9	68.6	67.1	**72.1**
AHSD	16.1	39.5	14.4	28.8	48.5	45.3	91.7	90.6	91.7	49.4	50.6	54.0	78.4	79.1	79.4	91.4	91.0	**91.9**
Eng Avg	36.8	55.4	35.5 -1.3	52.7	61.4	63.9 +11.2	86.6	86.1	86.9 +0.3	74.6	75.0	76.2 +1.6	83.1	83.0	83.2 +0.1	86.6	86.2	86.8 +0.2
GermEval21	39.4	38.9	38.8	59.9	58.4	**60.3**	39.3	39.9	39.2	56.4	57.5	58.1	51.5	54.7	54.4	59.8	59.4	60.0
GermEval19	40.0	39.9	39.9	74.8	70.8	78.8	41.7	41.7	41.3	73.4	69.8	74.6	72.0	71.0	73.5	77.5	73.4	**79.1**
GermEval18	40.1	39.8	39.7	76.4	72.8	81.4	42.0	42.4	41.0	80.5	75.5	**83.1**	76.3	74.6	78.6	80.5	75.7	82.6
HASOC	42.6	42.7	42.7	73.5	71.3	72.5	47.2	48.3	46.3	68.0	68.5	69.3	69.1	67.7	71.9	**75.4**	73.7	74.8
Gahd	37.7	36.5	36.9	71.0	70.0	**73.2**	39.2	40.8	38.9	53.2	53.9	55.5	45.7	46.5	45.9	72.9	71.2	**73.2**
Deu Avg	39.7	39.2	39.2 -0.5	71.9	69.3	75.0 +3.1	41.1	41.8	40.7 -0.4	68.3	66.0	70.2 +1.9	64.8	64.6	66.6 +1.8	74.4	71.3	75.5 +1.1
ViHSD	45.0	45.7	45.0	49.1	50.9	55.5	45.2	45.6	45.0	72.2	67.4	**74.0**	71.0	64.7	72.8	72.3	67.9	73.8
Haternet	67.3	64.9	68.1	45.7	48.7	56.8	43.0	45.1	44.4	45.2	46.7	48.0	47.2	45.2	49.1	69.3	67.7	**71.9**
HateEval-spa	75.9	73.5	76.6	38.3	38.4	46.9	36.2	36.7	36.3	36.9	36.7	38.6	38.5	36.5	38.2	78.6	76.7	**79.0**
Chileno	55.4	49.4	56.8	50.1	49.7	48.8	48.5	48.5	48.5	50.2	49.0	50.2	49.8	49.1	49.6	**57.1**	54.0	56.0
Spa Avg	75.0	72.5	75.5 +0.5	46.0	46.8	52.9 +6.9	44.0	44.9	44.4 +0.4	45.2	45.5	47.1 +1.9	46.7	44.9	47.3 +0.6	76.9	75.1	77.4 +0.5
Avg	44.1	57.2	43.6 -0.5	61.3	65.0	66.9 +5.6	72.1	71.9	72.3 +0.2	71.6	70.7	73.1 +1.5	75.2	74.7	75.7 +0.5	81.5	79.9	**82.1** +0.6

Table 3: Macro-F1 (%) for **BERT**, **HateBERT**, and four **OWS-BERT** variants across six training configurations. Bold-underlined = best per row. Darker-shaded cells = *in-domain* results (training language matches test language). Per-language averages show the OWS – BERT delta (Δ) on the OWS column.

consistently exceeds the human baseline: up to +8% for English and +22% for German, with `LGB` yielding the largest margins in both languages.

Vietnamese. All three synthetic-only strategies slightly trail the human baseline at the pooled level (F1 58–59 vs. 59.3), though `H+LGB` recovers strongly (+6.8%). Vietnamese comprises only ≈2.8% of the synthetic corpus (≈6,600 of 240K texts); the limited volume constrains standalone LLM annotation quality.

Spanish (zero-shot). No training configuration includes Spanish data. At the pooled level, `Mean` and `Vote` underperform the human baseline by 10–11% F1, because the synthetic corpus contains no Spanish texts and equal-weight aggregation cannot compensate. By contrast, `LGB` is the only synthetic-only strategy to exceed the human baseline (+1.6% pooled F1), confirming that the meta-learner acquires a degree of language-agnostic calibration.

Why LightGBM outperforms Vote and Mean. `Vote` and `Mean` treat all four annotators equally, amplifying any systematic biases shared across them. `LGB` learns *which* LLMs are reliable on *which* patterns and discovers decision boundaries that align with human annotations. The result is fewer false negatives in low-resource languages and better-calibrated probabilities overall.

4.2.2. Large-Scale Model: Qwen2.5-14B

`Qwen2.5-14B` fine-tuned on the 7-Set already approaches the best BERT models trained on all sixteen sets. Adding synthetic data yields a modest +0.6% pooled F1 with `H+LGB` (76.7 vs. 76.1). Language-level trends diverge:

	#	Base	Hum.	Mean	Vote	LGB	H+LGB
Llama3.2-1B							
EN	7	51.7	59.7	61.5	61.9	67.3	**75.4**
DE	5	44.6	42.7	55.6	54.3	**64.4**	57.8
VI	1	47.1	59.3	58.5	57.6	58.3	**66.1**
ES	3	33.0	46.5	36.7	35.2	48.1	**47.4**
7-Set	7	50.2	60.2	62.6	62.2	66.9	**69.7**
Rest	9	43.0	48.2	49.7	49.2	59.1	**60.3**
All	16	47.4	54.8	56.9	56.5	63.1	**65.4**
Qwen2.5-14B							
EN	7	69.4	**78.6**	78.5	74.2	76.5	**79.4**
DE	5	71.7	77.0	75.0	69.9	74.9	**76.1**
VI	1	71.9	76.0	76.0	70.9	75.9	**77.5**
ES	3	43.7	61.3	62.7	**67.9**	64.3	63.9
7-Set	7	67.4	76.6	76.0	73.8	76.0	**77.2**
Rest	9	67.6	75.3	75.3	71.5	73.8	**75.7**
All	16	67.8	76.1	75.9	72.8	75.1	**76.7**

Table 4: Pooled macro-F1 (%) by language and training split for LoRA-tuned LLMs. **#** = number of test sets; **7-Set**/**Rest** = training vs. held-out datasets. Scores are computed on the pooled predictions within each group, not averaged across datasets. The classification threshold is the mean predicted probability of the *Hate* class, not the default 0.5. Full per-dataset breakdown in Table 8 (Appendix).

- **English (+0.8%):** Small but consistent gains; per-dataset highlights include Covid (+5% F1) and HateEval-eng (+2.5%) (Table 8).
- **Spanish (+2.6%):** The clearest cross-lingual benefit—despite no Spanish training data, all ensemble strategies improve pooled F1, with `Vote` reaching +6.6%.
- **German (−0.9%):** Pooled F1 dips slightly; under `Vote` the drop reaches −7.1%. The model already achieves high German F1 on human data

(77.0); the imbalanced synthetic set (<3% hate) shifts its decision boundary towards *Neutral*, and `Vote` amplifies correlated LLM errors on German text.

5. Discussion

5.1. Answers to Research Questions

RQ1: Value of unlabelled web data. OWS continued pre-training reliably improves BERT-family models, especially in multilingual low-data settings. The effect is largest for `Ows4L` when training data is scarce (+3% average F1 under 7-Set) and diminishes as supervised data grows (+1% under 16-Mix). Neither monolingual OWS models nor HateBERT provide broad cross-lingual generalisation; only `Ows4L` achieves consistent gains across all four language groups.

RQ2: LLM ensemble annotations. The value of synthetic annotations depends strongly on (a) downstream model capacity and (b) language coverage of the synthetic set. For `Llama3.2-1B`, `H+LGB` labels yield +10.6% pooled F1 over the human baseline, effectively distilling knowledge from larger annotation LLMs into a small downstream model. For `Qwen2.5-14B`, gains are modest (+0.6% pooled F1); its strong priors quickly saturate on the imbalanced synthetic data. The LightGBM meta-learner is the most effective strategy because it learns model-reliability weights from human labels, making it robust to the shared biases of individual LLMs.

5.2. Limitations

Corpus quality. The URL/schema-based OWS filter is efficient but cannot guarantee conversational content quality. Manual inspection of a random sample and more precise classifier-based filtering would yield a cleaner pre-training corpus.

Label space collapse. Mapping hate, offensive, and abusive labels to a single positive class may conflate qualitatively different phenomena and reduce generalisation to datasets that maintain fine-grained distinctions.

LLM annotation bias. Synthetic labels inherit biases from the four annotating LLMs, all predominantly English-centric. The under-performance of `Mean` and `Vote` on Vietnamese and Spanish reflects this.

Ethical considerations. During data collection, no personally identifiable information (usernames, author profiles, or metadata) was gathered; we retained only raw texts, comments, and posts from publicly indexed web pages.

5.3. Future Work

We plan to (i) scale the OWS corpus substantially, increasing both volume and language diversity, to strengthen continued pre-training; (ii) adopt more capable recent SLMs and LLMs to assess whether newer architectures widen the gap between small and large models; (iii) extend coverage to additional low-resource languages; (iv) explore improved prompting strategies (chain-of-thought, calibration prompts) to reduce annotation bias and label imbalance; and (v) investigate hate-focused corpus filtering to increase the proportion of relevant content in OWS samples.

6. Conclusion

We presented a large-scale benchmark study examining how unlabelled web data and ensemble LLM annotations can improve multilingual hate-speech detection across sixteen benchmarks in four languages. Our key findings are:

1. Domain-adaptive continued pre-training on OWS data provides consistent F1 gains, with the multilingual `Ows4L` achieving the best average macro-F1 (77.0%) across all configurations.
2. The LightGBM ensemble outperforms mean averaging and majority voting and is the only synthetic strategy that reliably avoids regressions on unseen languages.
3. The benefit of synthetic data scales inversely with model capacity: +10.6% pooled F1 for `Llama3.2-1B`, but only +0.6% for `Qwen2.5-14B`.
4. Severe class imbalance in OWS-derived synthetic sets remains a critical bottleneck, especially for low-resource languages.

These findings motivate future work on scaling OWS data, adopting updated models, and developing richer annotation strategies.

Acknowledgements

This work has received funding from the Bavarian State Ministry of Economic Affairs, Regional Development, and Energy (StMWi).

Bayerisches Staatsministerium für
Wirtschaft, Landesentwicklung und Energie

7. References

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. Seattle, Washington (Hybrid). Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Idris Abdulmumin, Abinew Ali Ayele, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. AfroXLMR-Social: Adapting pre-trained language models for African languages social media text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Kheir Eddine Daouadi, Yaakoub Boualleg, and Kheir Eddine Haouaouchi. 2024. Ensemble of pre-trained language models and data augmentation for hate speech detection from arabic tweets.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Alphaeus Dmonte, Tejas Arya, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Towards generalized offensive language identification.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models.

Qwen et al. 2025. Qwen2.5 technical report.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and LLM biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the 19th AAAI Conference on Web and Social Media (ICWSM)*.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset. 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 4405–4424, Mexico City, Mexico. Association for Computational Linguistics.

Michael Granitzer, Stefan Voigt, Noor Afshan Fathima, Martin Golasowski, Christian Guetl, Tobias Hecking, Gijs Hendriksen, Djoerd Hiemstra, Jan Martinovič, Jelena Mitrović, Izidor Mlakar, Stavros Moiras, Alexander Nussbaumer, Per Öster, Martin Potthast, Marjana Senčar Srdič, Sharikadze Megi, Kateřina Slaninová, Benno Stein, Arjen P. de Vries, Vít Vondrák, Andreas Wagner, and Saber Zerhoudi. 2024. Impact and development of an open web index for open web search. *Journal of the Association for Information Science and Technology*, 75(5):512–520.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. Online. Association for Computational Linguistics.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. 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. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2022. Racism is a virus: anti-asian hate and counter-speech in social media during the covid-19 crisis. ASONAM '21, New York, NY, USA. Association for Computing Machinery.

Gijs Hendriksen, Michael Dinzinger, Sheikh Mastura Farzana, Noor Afshan Fathima, Maik Fröbe, Sebastian Schmidt, Saber Zerhoudi, Michael Granitzer, Matthias Hagen, Djoerd Hiemstra, Martin Potthast, and Benno Stein. 2024. The open web index: Crawling and indexing the web for public use. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V*, page 130–143, Berlin, Heidelberg. Springer-Verlag.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jonathan Jiang and Mistral AI. 2023. Mistral 7b. Accessed: 2024-05-01.

Mohammad Kashif, Mohammad Zohair, and Saquib Ali. 2023. Lexical squad@multimodal hate speech event detection 2023: Multimodal hate speech detection using fused ensemble approach.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. *A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts*, page 415–426. Springer International Publishing.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. FIRE '19, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21).

Paloma Piot, David Otero, Patricia Martín-Rodilla, and Javier Parapar. 2025. Can LLMs evaluate what they cannot annotate? Revisiting LLM reliability in hate speech detection. *arXiv preprint arXiv:2512.09662*.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12. Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. Gemma 2: Improving open language models at a practical size.

Björn Ross and et al. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *Workshop on Natural Language Processing for Computer-Mediated Communication*.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Floeck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. pages 352 – 363.

UnslothAI. 2023. Unsloth: A lightweight and fast framework for large-scale nlp models. Accessed: 2025-02-18.

Muhammad Usman, Muhammad Ahmad, M. Shahiki Tash, Irina Gelbukh, Rolando Quintero Tellez, and Grigori Sidorov. 2025. Multilingual hate speech detection in social media using translation-based approaches with large language models.

Bertie Vidgen and et al. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*.

Michael Wiegand. 2019. GermEval-2018 Corpus (DE).

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT reproduce human-generated labels? A

study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A. LLM Annotation Details

A.1. Annotation Prompt

Each OWS text was classified by the four instruction-tuned LLMs using the following zero-shot prompt:

```
You are tasked with annotating
speech.  Your response must be a
single valid number:

1 for
Hate/Offensive/Sexism/Toxic/Political/
COVID-related Hate Speech,

2 for Neutral Speech.

Provide only the number
corresponding to the category.
Do not include any explanation or
additional text.

Do you think the following
comment is
Hate/Offensive/Sexism/Toxic/Political/
COVID-related Hate Speech, speech
or Neutral speech?

"{comment}"

Your Answer:
```

Token-level probabilities for the target labels were extracted from the output distribution of each model. For the 240K ensemble annotation pool (Section 3.4), all four models were queried; for the 46K BERT augmentation subset (7-Set + Synth.), only `Qwen2.5-14B` was used.

A.2. LightGBM Meta-Learner Configuration

The LightGBM meta-learner (`LGB`) was trained with the following hyperparameters:

Two separate binary classifiers were trained—one for the Hate class and one for the Neutral class—using the eight-dimensional probability vectors (two classes $\times$ four models) from the human-labelled training sets as features.

A.3. 240K Ensemble Annotation Statistics

Table 6 summarises the 240,647-text annotation pool used for LLM ensemble labelling.

Table 5: LightGBM training configuration for the ensemble meta-learner.

Parameter	Value
objective	binary
metric	binary_logloss
boosting_type	gbdt
num_leaves	34
learning_rate	0.05
feature_fraction	0.9
bagging_fraction	0.8
bagging_freq	5

Table 6: Per-model and ensemble statistics of the 240K annotation pool, broken down by language. "% Hate" = fraction of texts with $P(\text{Hate}) > 0.5$ (per-model) or positive label (ensemble).

	deu	eng	vie	All
	125 617	108 375	6 655	240 647
Per-model mean $P(\text{Hate})$				
Qwen2.5-14B	0.027	0.021	0.031	0.024
Gemma2-9B	0.107	0.039	0.113	0.080
Llama3.1-8B	0.002	0.001	0.001	0.011
Mistral-7B	0.026	0.015	0.023	0.021
Per-model % classified Hate ($P > 0.5$)				
Qwen2.5-14B	2.69	1.96	3.56	2.38
Gemma2-9B	8.41	3.25	6.40	6.03
Llama3.1-8B	0.21	0.19	0.03	0.19
Mistral-7B	2.43	1.60	2.48	2.05
Ensemble % Hate				
Mean	1.98	1.27	1.89	1.66
LGB	1.57	0.94	2.01	1.30
Vote	2.37	1.45	2.42	1.96

A.4. 46K BERT Augmentation Statistics

Table 7 summarises the 45,940-text subset annotated solely by `Qwen2.5-14B` and used as synthetic training data for the BERT 7-Set + Synth. configuration.

Table 7: Label distribution of the 46K BERT augmentation set, annotated by `Qwen2.5-14B`, broken down by language.

	deu	eng	vie	spa	All
	26 084	11 635	6 871	1 350	45 940
Qwen2.5-14B label counts					
Neutral	22 702	4 043	6 341	1 143	34 229
Hate	3 284	3 943	474	149	7 850
Offensive	98	3 649	56	58	3 861
Qwen2.5-14B label %					
Neutral	87.03	34.75	92.29	84.67	74.51
Hate	12.59	33.89	6.90	11.04	17.09
Offensive	0.38	31.36	0.82	4.30	8.40
Mean $P(\text{Hate})$					
Qwen2.5-14B		—			0.451
Gemma2-9B		—			0.527

A.5. OWS Schema Types

The following ten `schema.org` types (matched under both `http` and `https` schemes) were used to filter the OpenWebIndex for conversational and user-generated content:

1. `DiscussionForumPosting`
2. `SocialMediaPosting`
3. `BlogPosting`
4. `Article`
5. `Comment`
6. `UserComments`
7. `QAPage`
8. `Question`
9. `Review`
10. `Blog`

B. Full Per-Dataset LLM Results

Llama3.2-1B

Dataset	Base		Human		Mean		Voting		LGB		Human+LGB	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
HateXplain	53.6	53.3	58.4	58.4	60.8 +02.4	60.7 +02.3	61.3 +02.9	61.3 +02.9	67.7 +09.3	67.0 +08.6	70.4 +12.0	**67.3** +08.9
Sexism	74.2	47.5	78.5	59.8	82.3 +03.8	51.7 -08.1	81.1 +02.6	50.9 -08.9	82.6 +04.1	56.3 -03.5	83.5 +05.0	**68.5** +08.7
Covid	38.2	37.7	75.0	63.0	77.3 +02.3	65.1 +02.1	76.1 +01.1	64.6 +01.6	69.6 -05.4	62.7 -00.3	76.3 +01.3	**69.6** +06.6
US_election	30.3	29.9	83.4	51.6	82.3 -01.1	59.8 +08.2	81.7 -01.7	60.0 +08.4	79.9 -03.5	**62.3** +10.7	85.2 +01.8	56.8 +05.2
HateEval-eng	49.5	49.5	58.2	51.5	61.9 +03.7	56.8 +05.3	62.1 +03.9	57.6 +06.1	64.0 +05.8	62.1 +10.6	67.2 +09.0	**66.2** +14.7
AbusEval	64.5	56.2	75.9	52.7	77.7 +01.8	51.5 -01.2	77.7 +01.8	52.2 -00.5	80.3 +04.4	**66.1** +13.4	78.7 +02.8	61.7 +09.0
AHSD	44.0	40.8	41.5	40.5	44.4 +02.9	43.6 +03.1	45.5 +04.0	44.5 +04.0	51.7 +10.2	49.7 +09.2	73.6 +32.1	**66.3** +25.8
GermEval21	46.8	46.2	41.5	39.3	55.5 +14.0	55.4 +16.1	54.9 +13.4	54.8 +15.5	57.5 +16.0	**57.3** +18.0	53.9 +12.4	53.9 +14.6
GermEval19	42.8	40.8	44.8	43.5	55.4 +10.6	55.4 +11.9	53.3 +08.5	53.1 +09.6	67.9 +23.1	**66.8** +23.3	59.2 +14.4	59.1 +15.6
GermEval18	42.5	40.0	44.7	42.8	55.0 +10.3	54.9 +12.1	53.5 +08.8	53.3 +10.5	68.6 +23.9	**67.2** +24.4	58.8 +14.1	58.7 +15.9
HASOC	40.5	40.2	56.5	54.1	56.7 +00.2	55.9 +01.8	53.6 -02.9	53.2 -00.9	69.8 +13.3	**65.9** +11.8	61.6 +05.1	59.8 +05.7
Gahd	54.9	54.7	45.1	37.8	56.9 +11.8	55.3 +17.5	57.3 +12.2	55.7 +17.9	61.4 +16.3	**61.3** +23.5	57.9 +12.8	56.8 +19.0
ViHSD	50.8	47.1	75.0	59.3	65.9 -09.1	58.5 -00.8	64.5 -10.5	57.6 -01.7	64.6 -10.4	58.3 -01.0	76.3 +01.3	**66.1** +06.8
Haternet	35.3	34.9	44.2	44.0	38.6 -05.6	38.5 -05.5	36.3 -07.9	35.9 -08.1	52.3 +08.1	49.8 +05.8	50.9 +06.7	**50.5** +06.5
HateEval-spa	42.6	38.8	53.7	53.7	46.3 -07.4	42.9 -10.8	47.2 -06.5	43.3 -10.4	49.9 -03.8	49.8 -03.9	54.4 +00.7	**53.9** +00.2
Chileno	25.4	23.7	47.6	37.7	29.0 -18.6	25.9 -11.8	26.6 -21.0	24.2 -13.5	51.6 +04.0	**39.0** +01.3	42.9 -04.7	35.6 -02.1
Avg	47.5	47.4	55.6	54.8	58.1 +02.5	56.9 +02.1	57.4 +01.8	56.5 +01.7	64.6 +09.0	63.1 +08.3	65.7 +10.1	**65.4** +10.6

Qwen2.5-14B

Dataset	Base		Human		Mean		Voting		LGB		Human+LGB	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
HateXplain	66.3	54.6	73.9	70.0	74.9 +01.0	72.2 +02.2	73.8 -00.1	72.4 +02.4	73.2 -00.7	69.5 -00.5	75.6 +01.7	**73.1** +03.1
Sexism	75.0	65.0	86.1	**73.8**	85.8 -00.3	64.9 -08.9	85.7 -00.4	62.9 -10.9	86.5 +00.4	69.3 -04.5	88.0 +01.9	71.1 -02.7
Covid	39.9	39.8	66.5	62.3	65.9 -00.6	62.4 +00.1	72.6 +06.1	67.6 +05.3	69.7 +03.2	65.8 +03.5	72.6 +06.1	**67.9** +05.6
US_election	49.6	46.0	86.8	66.1	83.1 -03.7	**67.1** +01.0	85.3 -01.5	62.3 -03.8	85.4 -01.4	66.8 +00.7	86.3 -00.5	63.3 -02.8
HateEval-eng	63.0	61.7	67.2	66.9	69.7 +02.5	69.5 +02.6	69.0 +01.8	68.3 +01.4	70.0 +02.8	**69.8** +02.9	70.0 +02.8	69.4 +02.5
AbusEval	66.9	63.4	82.1	**68.7**	81.5 -00.6	68.2 -00.5	81.4 -00.7	67.2 -01.5	80.5 -01.6	67.1 -01.6	82.1 +00.0	66.2 -02.5
AHSD	91.9	**84.9**	81.9	74.9	81.5 -00.4	74.8 -00.1	63.8 -18.1	59.6 -15.3	72.8 -09.1	67.1 -07.8	79.5 -02.4	72.9 -02.0
GermEval21	64.4	**63.8**	68.9	**63.8**	68.1 -00.8	58.5 -05.3	66.2 -02.7	51.3 -12.5	67.3 -01.6	58.3 -05.5	68.6 -00.3	59.9 -03.9
GermEval19	72.2	71.8	81.3	78.9	81.9 +00.6	79.0 +00.1	79.5 -01.8	74.0 -04.9	81.2 -00.1	78.3 -00.6	82.2 +00.9	**79.1** +00.2
GermEval18	75.8	75.3	82.8	**80.9**	82.6 -00.2	79.5 -01.4	80.1 -02.7	75.0 -05.9	82.6 -00.2	79.9 -01.0	83.2 +00.4	80.5 -00.4
HASOC	67.7	66.6	81.7	78.1	82.7 +01.0	**78.7** +00.6	83.1 +01.4	76.8 -01.3	81.6 -00.1	75.5 -02.6	83.1 +01.4	78.3 +00.2
Gahd	73.5	73.5	77.7	**77.7**	75.8 -01.9	75.0 -02.7	71.6 -06.1	69.3 -08.4	75.5 -02.2	75.0 -02.7	77.3 -00.4	77.0 -00.7
ViHSD	78.7	71.9	85.0	76.0	87.1 +02.1	76.0 +00.0	86.4 +01.4	70.9 -05.1	87.3 +02.3	75.9 -00.1	86.9 +01.9	**77.5** +01.5
Haternet	46.4	46.3	69.5	67.3	77.4 +07.9	74.0 +06.7	81.2 +11.7	**76.0** +08.7	75.0 +05.5	72.2 +04.9	77.0 +07.5	73.5 +06.2
HateEval-spa	57.9	54.6	65.2	64.6	65.2 +00.0	65.1 +00.5	67.8 +02.6	**67.8** +03.2	67.2 +02.0	67.0 +02.4	67.5 +02.3	67.4 +02.8
Chileno	32.6	29.2	60.1	44.1	60.3 +00.2	44.3 +00.2	74.1 +14.0	**49.6** +05.5	63.4 +03.3	46.8 +02.7	62.1 +02.0	44.8 +00.7
Avg	67.8	67.8	77.2	76.1	77.5 +00.3	75.9 -00.2	76.1 -01.1	72.8 -03.3	76.7 -00.5	75.1 -01.0	78.3 +01.1	**76.7** +00.6

Table 8: Full per-dataset accuracy and macro-F1 (%) of `Llama3.2-1B` (top) and `Qwen2.5-14B` (bottom) under six training conditions. `Base` = zero-shot; `Human` = 7-Set human labels; `Mean`/`Voting`/`LGB` = synthetic-only; `Human+LGB` = combined. Signed values (±) are relative to `Human`. Bold = best F1 per row. The classification threshold is the mean predicted probability of the *Hate* class, not the default 0.5.