Do We Still Need Humans in the Loop? Comparing Human and LLM Annotation in Active Learning for Hostility Detection

Anonymous ACL submission

Abstract

Instruction-tuned LLMs can annotate thousands of instances at low cost. This raises two questions for active learning (AL): can LLM labels replace human labels within the AL loop, and does AL remain necessary when entire corpora can be cheaply labeled? We investigate both on a new dataset of 277,902 German political TikTok comments (25,974 LLM-labeled, 5,000 human-annotated), comparing LLM and human annotation across seven conditions, four encoders, and 10 random seeds. Under a two-question interface that mirrors the human annotation task, LLM annotation at scale outperforms human-supervised classifiers at roughly one-tenth the cost ($28 for GPT-5.2 Batch API vs. $316 for Prolific). The advantage holds for both a closed-source (GPT-5.2) and an open-weight (Qwen3.5-122B-10B) LLM, is robust under soft-label evaluation, and is unlocked specifically by the two-question decomposition; a holistic single-prompt baseline only ties with human supervision. AL provides no reliable advantage over random sampling under either LLM annotator. However, error structure varies sharply: only GPT-5.2 under the two-question interface produces classifiers with near-human FP/FN balance, while other LLM variants over-flag border-control and economic-competition discourse. We release the dataset and code.

1 Introduction

Annotating data is expensive, a persistent bottleneck for supervised NLP. A recent survey finds that 94% of practitioners consider annotation a limiting factor, and that active learning (AL) remains widely used (Romberg et al., 2026). In AL, a classifier iteratively selects the most informative unlabeled instances, queries a human for labels, and retrains, aiming to reach target performance with fewer labels than random sampling (Settles, 2009).

Recent work places LLMs inside the AL loop as annotators (Zhang et al., 2023; Kholodna et al., 2024), but these studies do not compare human and LLM labels under matched conditions, nor do they examine how the annotation source affects classifier behavior. When an LLM can cheaply label an entire corpus, the practical question shifts: from how to allocate a human annotation budget (AL versus random sampling) to whether to spend one at all, with AL (human labels) now competing against full LLM annotation at scale. Apart from cost, LLM annotation does not expose workers to psychological harm (Steiger et al., 2021), may help overcome the scarcity of qualified workers in many languages, and enables rapid iteration over annotation guidelines. We use cost as the operational lens because it is the most directly measurable, and test whether AL retains a performance advantage that justifies its cost premium.

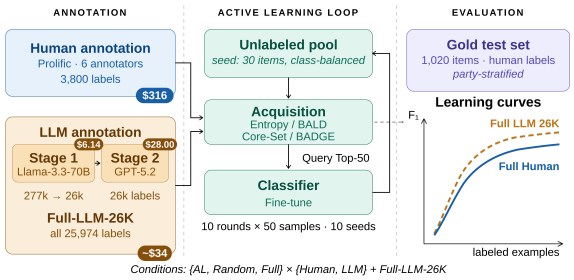


Figure 1: **Experimental framework.** Human and LLM annotations feed into the same active learning loop, evaluated across seven conditions, four German encoder models, and 10 seeds on a party-stratified test set.

We investigate this question on detecting anti-immigrant hostility in German TikTok comments collected from the official accounts of six major political parties. The task is challenging to annotate: it requires differentiating hostility directed at immigrant groups from criticism of immigration as a policy issue, a culturally specific distinction where annotation source matters (Sap et al., 2022). It also represents a common deployment scenario: training a lightweight supervised classifier from limited labels for inference at scale. Our experimental design (Fig. 1) compares seven labeling conditions

that vary annotation source, sampling strategy (AL vs. random vs. full pool), and label volume (530 to 25,974 instances) across four encoders and 10 random seeds. The LLM conditions use GPT-5.2 and Qwen3.5-122B with a two-question decomposition matching the human annotation interface; an interface ablation (GPT-5.2 with a holistic single-question prompt) and a reasoning-mode ablation (Qwen3.5-122B without reasoning) are reported in App. E.1 and App. E.2. The labeling pool comprises 25,974 immigration-relevant comments constructed via a two-stage LLM pipeline, of which 5,000 are independently annotated by six crowdworkers, yielding 1,200 gold evaluation items and 3,800 training labels. We organize the paper around four research questions:

RQ1: Can LLM annotation at scale match or exceed human-supervised classifiers, and how does this depend on the annotation interface and the choice of LLM?

RQ2: Does any AL acquisition strategy reliably outperform random sampling under LLM-supervised pools?

RQ3: Is the LLM-at-scale advantage robust to soft-label evaluation that accounts for human annotator disagreement?

RQ4: How do error structure and topical disagreement vary across LLM annotators?

Our central finding is that under a two-question interface that mirrors the human annotation task, LLM annotation at scale significantly outperforms human-supervised classifiers at roughly one-tenth the cost using a closed-source LLM (GPT-5.2), with the result holding under an open-weight alternative (Qwen3.5-122B) run on local hardware.

2 Related Work

2.1 Active Learning and LLM Annotators

AL reduces annotation cost by selecting the most informative instances for human labeling (Settles, 2009), but practical adoption remains limited by unreliable performance gains, hyperparameter sensitivity, and difficulty obtaining reliable uncertainty estimates for acquisition functions (Lowell et al., 2019; Romberg et al., 2026; Schröder and Niekler, 2020). AL strategies frequently fail to outperform random selection on text and multimodal tasks, with pool composition often the underlying cause (Karamcheti et al., 2021; Snijders et al., 2023). Instruction-tuned LLMs offer a complementary cost reduction: zero-shot ChatGPT achieves higher accuracy than crowd workers relative to trained-annotator labels across several text annotation tasks at roughly one-thirtieth the cost (Gilardi et al., 2023). These findings extend to open-source models and other domains (Alizadeh et al., 2025; Törnberg, 2025), including German migration discourse (Kostikova et al., 2024). Recent work places LLMs inside the AL loop as annotators (Zhang et al., 2023; Kholodna et al., 2024; He et al., 2024; Xia et al., 2025), though annotation quality is task- and prompt-sensitive and requires validation against human gold labels (Pangakis et al., 2023), with hate-speech zero-shot performance depending on the definition prompt (Melis et al., 2025). We address these gaps with a matched-condition comparison of human and LLM annotations that evaluates downstream classifier behavior (not just annotation agreement) and contrasts AL with human labels against LLM annotations at scale.

2.2 Subjective Annotation and Disagreement

Hate speech detection has been widely surveyed (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018); annotation difficulty increases when hostility toward a *group* must be distinguished from criticism of a *policy* (Waseem and Hovy, 2016) or when bias is implied rather than explicit (Sap et al., 2020), the central challenge in our task. Related work creates dehumanization annotation resources (Engelmann et al., 2024) and treats ambiguity as signal (Assenmacher et al., 2025). LLM hate-speech annotation is sensitive to extralinguistic context: party cues for political text (Vallejo Vera and Driggers, 2025) and geographical, persona, and anchoring cues (Masud et al., 2024). The most related German dataset (Fillies et al., 2025) covers far-right ecosystems rather than party accounts. Hate-speech annotation faces further challenges in distinguishing hate from offensive language (Davidson et al., 2017), cross-dataset and cross-cultural transfer (Antypas and Camacho-Collados, 2023; Lee et al., 2024), and dataset construction (Piot et al., 2024; Vidgen et al., 2021). Annotator disagreement on subjective tasks is increasingly understood as signal rather than noise (Plank, 2022; Röttger et al., 2022; Davani et al., 2022), with low inter-annotator agreement common even under detailed guidelines for German hate speech (Ross et al., 2017). Recent work proposes aggregating soft labels to capture this signal in both training and evaluation (Wright and Augenstein, 2025), raising

the question of whether LLMs can capture human disagreement patterns (Ni et al., 2025).

3 Methodology

3.1 Dataset Collection and Preprocessing

We collected public comments from the official TikTok accounts of six major German political parties (CDU/CSU, Bündnis 90/Die Grünen, SPD, FDP, AfD, Die Linke) via the TikTok Research API, covering January 2024 to September 2025, yielding 467,660 comments. Anti-immigrant hostility is a culturally embedded subjective task where annotation source plausibly matters, and our corpus is newly collected rather than drawn from an existing benchmark, reducing pretraining exposure for LLM annotators. We applied a four-stage preprocessing pipeline: (1) removal of duplicate comment IDs; (2) content filtering to remove non-informative comments (emoji-only, empty, or single-token sequences; (3) exact text deduplication within and across party accounts; and (4) language identification via GlotLID (Kargaran et al., 2023) to remove non-German. This yielded a corpus of **277,902** German comments (median 14 words); per-step/per-account counts are in App. A. This corpus forms the input to the LLM-based prefiltering stage (§3.3).

3.2 Label Space Definition

Following prior work on anti-immigrant hate speech (Basile et al., 2019; Sanguinetti et al., 2018), we define a binary classification task. A comment is labeled ANTI-IMMIGRANT if it expresses hostility towards immigrants (see next paragraph for detailed description). All other comments are labeled NOT ANTI-IMMIGRANT.

Two-Question Annotation Framework. We operationalize this label through the two sequential yes/no questions Q1/Q2 rather than a direct binary judgment (same decomposition for both human and LLM annotation). **Q1 (reference):** Does this comment refer to immigrants, refugees, asylum seekers, foreigners, or immigration policy? **Q2 (valence):** Are these groups portrayed negatively through hostility, stereotypes, deportation rhetoric, or attribution of social or economic problems?

The final label is computed as: ANTI-IMMIGRANT $\Leftrightarrow$ Q1 = YES $\wedge$ Q2 = YES; otherwise NOT ANTI-IMMIGRANT. The decomposition serves two purposes. For human annotators, it addresses low agreement in an initial pilot using a single holistic prompt (Fleiss κ=.175; App. B.2), which required integrating topic, target, and valence in a single step. Embedding each criterion in the interface forces explicit evaluation and enables diagnostic tracing of disagreements to either reference (Q1) or valence (Q2). The same decomposition is used for the LLM interface, supporting a direct comparison without interface asymmetry confounds. We refer to this as the *decomposed* (D) interface; a *holistic* (H) ablation is reported in App. E.1.

3.3 Pool Construction and Annotation

We construct the labeling pool through a two-stage LLM-based pipeline. Stage 1 uses Llama-3.3-70B-Instruct as a high-recall topic filter, chosen for high-throughput batched inference at low cost. Stage 2 produces labels under the decomposed interface from §3.2 for the more difficult subjective classification, using two frontier LLMs, closed-source GPT-5.2 and open-weight Qwen3.5-122B-A10B, to test whether findings generalize across LLMs. Implementation details are in App.C.4.

Stage 1: Topic Prefiltering. We applied `Llama-3.3-70B-Instruct` (Team, 2024) with a binary prompt asking whether each comment discusses immigration, migrants, refugees, asylum seekers, or related policies (prompt: App. C.1). This reduced the corpus to **25,974 comments**.

Stage 2: Annotation. GPT-5.2 (D). We applied `GPT-5.2` via the OpenAI Batch API with a single prompt eliciting Q1 and Q2 as structured JSON (prompt: App. C.3); all 25,974 comments received valid labels (Table 1), with the 21.8% ANTI-IMMIGRANT rate close to the 21.7% base rate in the human gold evaluation subset (§3.4).

Qwen3.5-122B-A10B (D). Qwen3.5-122B-A10B (Qwen Team, 2026) ran on the same 25,974 comments with reasoning enabled. 136 items hit the 16,384-token output budget before emitting valid JSON and were excluded from training; evaluation on the human-labeled test set is unaffected.

Ablations. Two variants isolate the contributions of the decomposed interface (GPT-5.2 holistic; prompt App. C.2, results App. E.1) and of reasoning (Qwen3.5 reasoning disabled; results App. E.2). We refer to the four variants throughout as **GPT-D**, **Qwen-DT**, **GPT-H**, and **Qwen-DNT**.

Prefilter validation. To verify that the Stage 1 filter does not systematically discard relevant content,

LLM	Interface / mode	ANTI	%
GPT-5.2	D	5,651	21.8
Qwen3.5	D + thinking	7,961	30.6
GPT-5.2	H	7,847	30.2
Qwen3.5	D, no-thinking	8,312	32.0

Table 1: ANTI-IMMIGRANT label counts and rates per LLM variant on the 25,974-comment prefiltered pool. Top two rows: main annotators; bottom two: ablations (App. E.1, App. E.2). Per-party breakdowns: App. C.5.

we re-annotated a party-stratified sample of 500 Llama-excluded comments with GPT-D. All 500 received Q1 = NO (no immigration reference) and were labeled NOT ANTI-IMMIGRANT, indicating that the prefilter discards off-topic content rather than dropping ANTI-IMMIGRANT cases.

Annotation Corpus. From the GPT-D labeled pool, we sampled 5,000 comments for the human annotation study (§3.4) using party-stratified sampling at a 40/60 ANTI/NOT target ratio, a moderate enrichment over the 21.8% base rate to ensure sufficient positive instances for AL. The same 5,000 comments are covered by all four LLM variants, enabling matched-pool comparison across sources.

3.4 Human Annotation Study

We recruited annotators through Prolific, restricted to proficient German speakers. Of eight recruited, six were retained after qualification. Annotations were collected via argilla.io on Hugging Face Spaces using the two-question interface (§3.2). The study comprised three phases with decreasing overlap: a 50-item qualification test (6-way), a 200-item pilot (6-way), and a 4,750-item main round (950 triple-annotated, 3,800 single-annotated). The 1,200 multiply-annotated items (test + pilot + triple-annotated main) form the **gold evaluation set**; the 3,800 single-annotated items form the **AL training pool**. See App. B for qualification criteria, attention checks, and item assignment.

Inter-annotator agreement is moderate: Krippendorff's $\alpha = 0.49$ on the pilot (6-way, 200 items) and $\alpha = 0.43$ on the main triple-annotated items (3-way, 950 items). Agreement is higher on Q1 (reference, $\alpha = 0.57$–0.59) than on Q2 (valence, $\alpha = 0.43$–0.49): disagreement concentrates on valence rather than topic identification, consistent with prior work on German hate speech (Ross et al., 2017). Hard labels are aggregated via majority vote (full pairwise agreement in App. B); the soft-label evaluation in §3.7 uses Bayesian posteriors over the raw annotator votes.

Condition	Annotator	n_{train}
AL-HUMAN	Human	$\leq$530
AL-LLM	LLM	$\leq$530
RANDOM-HUMAN	Human	$\leq$530
RANDOM-LLM	LLM	$\leq$530
FULL-HUMAN	Human	3,800
FULL-LLM	LLM	3,800
FULL-LLM-26K	LLM	25,974

Table 2: Seven experimental conditions. Prefixes: AL = active learning, RANDOM = random sampling, FULL = full-pool training. AL conditions use four acquisitions (Entropy, BALD, Core-Set, BADGE; §3.6) and report max n_{train} at round 10 (30 seed + 500 acquired).

3.5 Experimental Conditions

We compare seven conditions (Table 2) spanning three design axes: annotation source (human or LLM), sampling strategy (AL vs. random vs. full pool), and label volume. LLM conditions are instantiated under our two LLM annotators (GPT-D and Qwen-DT). Human-vs-LLM at matched sampling isolates label source; AL-vs-Random isolates acquisition value; FULL-LLM-26K tests whether scale compensates for label quality. All conditions share the classifier, training recipe, and evaluation set (§3.7); costs in §3.8.

3.6 Active Learning Design

We implement pool-based AL for AL-HUMAN and AL-LLM. Both start from a class-balanced, party-stratified seed set of 30 comments (5 per party), following Fairstein et al. (2024), and simulate the AL loop by revealing pre-existing labels when queried. In each of 10 rounds, the top 50 most informative instances are selected, yielding up to 530 labeled examples at round 10. After each round, the classifier is retrained from scratch.

We evaluate four acquisition strategies: ENTROPY and BALD (Houlsby et al., 2011; Gal and Ghahramani, 2016) are uncertainty-based, selecting instances on which the classifier is least confident. Two are diversity-based, selecting instances that span encoder (CORE-SET, Sener and Savarese (2018)) and gradient (BADGE, Ash et al. (2020)) embedding spaces. See App. D.2 for details.

3.7 Model, Training, and Evaluation

Models and Training. We evaluate four pretrained transformer encoders: ModernGBERT, german-bert, gbert-base, and xlm-r-base; details in App. D.1. xml-r-base is a multilingual baseline. Each is fine-tuned with a randomly initialized binary classification head, with comments

tokenized and truncated to 128 tokens. We use AdamW (Loshchilov and Hutter, 2019) with learning rate 2×10^{-5}, batch size 16, weight decay 10^{-2}, and linear warmup (10%) followed by decay. To address class imbalance, we apply focal loss (Lin et al., 2017) with $\gamma = 2.0$ and per-class weights recomputed at each AL round. Training runs for up to 10 epochs with early stopping on dev macro-F1 (patience 5 for full conditions and the AL seed round, 3 for subsequent AL rounds).

Evaluation. The 1,200-item gold evaluation set is partitioned into a development set (15%, for early stopping) and a test set (85%, for final reporting), both stratified by class. We report macro-F1 supplemented by class-specific F1 for ANTI-IMMIGRANT (F1-Anti); for AL conditions, we additionally report the Area under the Learning Curve (ALC), normalized by the number of AL rounds. All experiments use 10 random seeds; we report mean and standard deviation. To test robustness to majority-vote aggregation, we additionally evaluate against two Bayesian posteriors over per-item labels (Wright and Augenstein, 2025): **Beta-Binomial (BB)**, treating annotators as exchangeable Bernoulli voters; and **Dawid-Skene (DS)** (Dawid and Skene, 1979), estimating per-annotator competence via EM (details in App. F.1).

3.8 Cost-Effectiveness Analysis

Table 3 compares annotation costs across conditions. All costs are reported in USD. Human annotation totaled $416 across all 5,000 items (Prolific's "good rate" of £9/hour).[1] GPT-5.2 annotation of the full 25,974-comment pool cost $28 via the OpenAI Batch API, plus $6.14 for Llama prefiltering. Qwen3.5 ran on local A100 GPUs. Token estimates, GPU-hours, and per-LLM breakdowns are in App. C.4. The order-of-magnitude ($\sim$10$\times$) gap between human annotation and the GPT-5.2 reflects throughput differences rather than transient pricing; see §4.2 for cost-performance tradeoffs.

4 Results

We evaluate all seven conditions across four encoders and 10 seeds on the held-out test set. The relative ordering of annotation strategies is consistent across encoders; tables report all four backbones and headline figures use ModernGBERT (full per-backbone results in App. H). LLM-supervised conditions use GPT-D and Qwen-DT; the ablations interface GPT-H and reasoning-mode Qwen-DNT are in App. E.1 and App. E.2.

[1] Phase breakdown: £18 (test, 20 min) + £54 (pilot, 60 min) + £162 (main, 3 hours) + £77.92 platform fee. (£1 = $1.334)

Condition	n_{train}	Cost	$/label
AL-HUMAN	530	$44	$0.083
AL-LLM	530	$0.57	$0.001
FULL-LLM	3,800	$4.10	$0.001
FULL-HUMAN	3,800	$316	$0.083
FULL-LLM-26K	25,974	$28	$0.001
Full human study	5,000	$416	$0.083

Table 3: Annotation costs by condition. Human cost scales proportionally from the $416 study total; LLM costs are GPT-5.2, excluding a one-time $6.14 for Llama prefiltering. Qwen3.5 costs (local A100s) are in App. C.4. RANDOM conditions are omitted (same cost as AL counterparts).

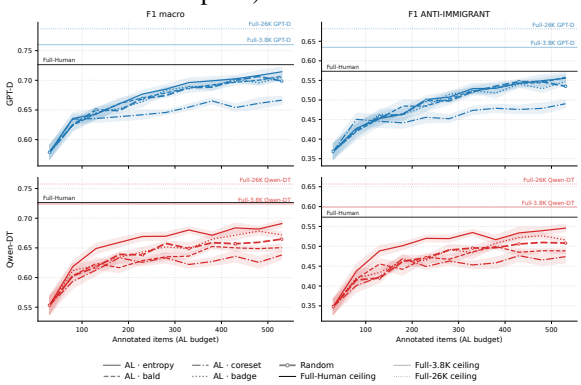


Figure 2: ModernGBERT learning curves. Top: GPT-D (blue); bottom: Qwen-DT (red). Horizontal lines: Full-LLM ceilings at 3.8K and 25,974 labels; black: FULL-HUMAN. Shaded bands: $\pm$1 SD across 10 seeds. Results for other encoders in App. H.

4.1 Annotation Source at Matched Volume and at Scale (RQ1)

Key finding. At matched volume (3,800 labels), FULL-LLM under GPT-D matches or slightly exceeds FULL-HUMAN. At scale (25,974 labels), FULL-LLM-26K consistently outperforms FULL-HUMAN across all encoders under both LLM annotators.

On 3,800 labels, Full-LLM under GPT-D matches Full-Human across all four backbones (Table 4; learning curves in Figure 2); Full-Human slightly outperforms Qwen-DT at this volume. Scaling to 25,974 labels changes the picture: Full-LLM-26K under GPT-D gains +0.05 to +0.06 F1-Macro over Full-Human across all four backbones; Qwen-DT gains +0.01 to +0.03. F1-Anti follows the same pattern (GPT-D: 0.682–0.686 vs. Full-Human 0.573–0.603). At $\sim$7$\times$ the volume and one-tenth the cost (§4.2), LLM annotation produces

Condition	germ.	Mod.	gbert	xlm-r
F1-Macro				
FULL-HUMAN	.740	.726	.735	.725
FULL-LLM, GPT-D	.736	.760	.736	.724
FULL-LLM-26K, GPT-D	**.789**	**.787**	**.790**	**.788**
FULL-LLM, Qwen-DT	.708	.724	.716	.692
FULL-LLM-26K, Qwen-DT	.752	.758	.755	.751
AL-HUMAN (Ent.)	.714	.703	.706	.704
AL-LLM, GPT-D (Ent.)	.713	.714	.720	.677
AL-LLM, Qwen-DT (Ent.)	.662	.691	.678	.660
RANDOM-HUMAN	.695	.677	.690	.702
RANDOM-LLM, GPT-D	.713	.699	.701	.703
RANDOM-LLM, Qwen-DT	.671	.665	.672	.662
F1-Anti Immigrant				
FULL-HUMAN	.603	.573	.599	.594
FULL-LLM, GPT-D	.609	.635	.610	.596
FULL-LLM-26K, GPT-D	**.685**	**.682**	**.686**	**.684**
FULL-LLM, Qwen-DT	.583	.599	.596	.563
FULL-LLM-26K, Qwen-DT	.651	.657	.654	.650
AL-HUMAN (Ent.)	.554	.537	.545	.562
AL-LLM, GPT-D (Ent.)	.561	.555	.577	.501
AL-LLM, Qwen-DT (Ent.)	.516	.546	.538	.486
RANDOM-HUMAN	.527	.492	.529	.548
RANDOM-LLM, GPT-D	.569	.535	.552	.564
RANDOM-LLM, Qwen-DT	.520	.508	.531	.514

Table 4: F1-Macro and F1-Anti at the final AL round (mean across 10 seeds). **Bold**/underline = best/second-best per encoder. AL conditions report Entropy; Random samples uniformly.

stronger classifiers on both metrics.

4.2 Cost-Effectiveness

Key finding. Full-LLM-26K achieves the strongest performance under both LLM annotators at roughly one-tenth the cost of Full-Human.

Figure 3 shows the cost-performance frontier for ModernGBERT. Three budget regimes: (i) Under $10, only LLM annotation is feasible; Full-LLM at matched volume costs $4 (F1-Macro 0.724-0.760). (ii) At $28, Full-LLM-26K dominates the frontier (F1-Macro 0.787-0.790). (iii) At $316, Full-Human reaches 0.725-0.740. Full-LLM-26K dominates Full-Human on F1 at every cost level.

4.3 AL vs. Random Sampling (RQ2)

Key finding. No acquisition strategy reliably outperforms random sampling. This holds for all four acquisition functions under both LLM and human annotators.

The four acquisition strategies cluster within 0.04 ALC of random sampling across encoders and label sources, with no consistent direction (App. H.2). Entropy edges above random most consistently: across the three source variants (Human, GPT-D, Qwen-DT), it produces the highest ALC in

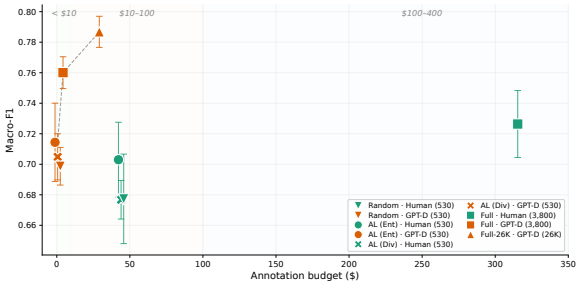


Figure 3: Annotation cost–performance frontier for ModernGBERT (mean ± 1 SD across 10 seeds). Three budget regimes: under $10 (LLM-only), $10–100 (LLM at scale), and $100–400 (human annotation).

11 of 12 cells, though the gap remains small (AL-Human: 0.658 vs. 0.647; AL-LLM GPT-D: 0.671 vs. 0.664; AL-LLM Qwen-DT: 0.635 vs. 0.628). Diversity-based methods diverge: BADGE consistently outperforms Core-Set by 0.01–0.04 ALC, but neither reliably beats random sampling.

AL vs. LLM at matched cost. AL-Human at round 10 acquires 530 labels at $44, while Full-LLM-26K (GPT-D) reaches 0.789 F1-Macro at $28: 49× more labels and 0.075–0.085 higher F1-Macro at comparable cost. AL reduces the cost of human annotation but not below LLM at scale.

4.4 Soft-Label Robustness (RQ3)

Key finding. Full-LLM-26K outperforms Full-Human under both majority-vote hard labels and soft-label evaluation against Beta-Binomial and Dawid-Skene posteriors. The result holds for both LLM annotators and all four encoders.

The hard-label result of §4.1 could be questioned: majority-vote ground truth obscures annotator disagreement on subjective items. We address this by evaluating against two Bayesian posteriors of the per-item label (§3.7) that weight items by annotator agreement. The pattern under both posteriors is consistent with the hard-label result (App. F.2). Under BB, Full-LLM-26K (GPT-D) exceeds Full-Human by 0.03–0.04 expected macro-F1 across encoders; under DS the gap widens further. Full-LLM-26K (Qwen-DT) shows the same direction with smaller magnitudes.

4.5 Error Structure and Confidence (RQ4)

Key finding. GPT-D is the only LLM annotator whose classifiers preserve a roughly human-like FP:FN ratio. However, all LLM-supervised classifiers, even those with fewer total errors, produce errors at substantially higher confidence than

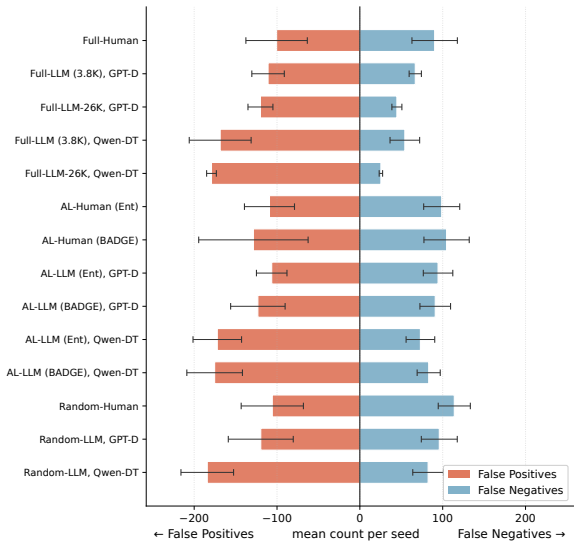


Figure 4: FP/FN counts per condition for ModernGBERT (mean ± 1 SD across 10 seeds; dots = individual seeds). Confidence-on-errors distributions in App. G.

human-supervised ones.

FP:FN asymmetry. Across four encoders (Fig. 4), FULL-HUMAN produces FP:FN ratios of 1.11–2.43 (mean 1.71). FULL-LLM-26K (GPT-D) reaches 2.66–2.74 (mean 2.70): FP counts are essentially identical to FULL-HUMAN (117–120 vs. 100–153) but FN counts roughly halve (44 vs. 63–90). The minority-class F1 gain of §4.1 is therefore a recall gain, not an FP artifact (App. G). FULL-LLM-26K (Qwen-DT) shows a different profile: FP counts 1.5–2× FULL-HUMAN (179–190) combined with low FN counts (23–26), yielding ratios of 7.05–8.14. Both LLM annotators improve minority-class recall, but only GPT-D preserves a human-like balance. Ablations show this depends on the decomposed interface: GPT-H has 7.54–9.89 and QWEN-DNT 6.51–6.91, both far from human-like (App. E.1, E.2).

Confidence on errors. Error counts tell only half the story; confidence matters too. FULL-HUMAN assigns mean predicted-class probability 0.691–0.782 to misclassified instances; FULL-LLM-26K (GPT-D) has 0.847–0.896 and (Qwen-DT) 0.841–0.936. The share of errors with confidence > 0.90 jumps from 10–38% under FULL-HUMAN to 52–69% (GPT-D) and 40–78% (Qwen-DT). The pattern is most pronounced for false negatives: under GPT-D on german_bert, 80% of FNs are predicted NOT ANTI-IMMIGRANT with confidence > 0.90 vs. 49% under FULL-HUMAN. This has practical consequences for moderation pipelines that triage items by confidence: under LLM-supervised classifiers, fewer items needing review surface as low-confidence cases. The ablations reverse on this dimension (GPT-H: 0.752–0.767 mean confidence on errors, similar to FULL-HUMAN; QWEN-DNT comparable), tying the confidence inflation to the decomposed-D / reasoning-on combination.

Annotation boundaries. A few instances are persistently misclassified (incorrect under 9 of 10 seeds across FULL-HUMAN, GPT-D, and Qwen-DT) and reflect genuine annotation ambiguity rather than model failure. Consider: *Wir nehmen die Menschen, wie sie sind und nicht wie sie sein sollten. Leitkultur* 🤡 ('We take people as they are and not as they should be. Leitkultur 🤡') This juxtaposes an emoji signalling sarcasm with a familiar quote and the term *Leitkultur* (a politically charged concept implying cultural assimilation). The sarcasm could target *Leitkultur* (mocking cultural assimilation, NOT ANTI-IMMIGRANT) or the acceptance statement (mocking tolerance, ANTI-IMMIGRANT). All three supervision regimes commit to the same incorrect reading with high confidence (.88, .94, .92). §4.6 examines: do such cases concentrate in specific discourse themes?

4.6 Topical Disagreement (RQ4)

Key finding. Per-topic human-LLM disagreement decomposes into three distinct pathways: an interface-driven component that the decomposed interface (GPT-D) resolves, a topic-intrinsic component that no LLM variant resolves, and a model-specific component on which the two LLM annotators diverge under the same interface.

We embed the 25,974 prefiltered comments using a multilingual sentence transformer and apply BERTopic (Grootendorst, 2022), yielding 22 named discourse topics (App. G.3). Per-topic agreement between human and LLM labels ranges from 37% to 96%; full breakdowns are in App. G.4.

Pathway 1: interface-driven. Several large clusters where the holistic interface over-detected anti-immigrant content are resolved under the decomposed interface: Welfare & Economic Competition (gap +28.7 under GPT-H drops to +0.7 under GPT-D), Rejection Rhetoric (+31.8 → 0.0), Islam & Islamization (+21.1 → −2.6), Directed Personal Hostility (+22.5 → +2.8), and Criminal Violence & Public Safety (+21.2 → +4.5). The Q1 (reference) / Q2 (valence) decomposition breaks a co-occurrence shortcut the holistic prompt did not,

directly accounting for a substantial fraction of the F1-Anti and FP:FN gains in §4.1 and §4.5.

Pathways 2 and 3: topic-intrinsic and model-specific. Two pathways remain unresolved. Border Control & Grenzschutz sees gaps of +49.0 (GPT-D), +54.9 (Qwen-DT), +45.1 (GPT-H), and +62.8 (Qwen-DNT): every LLM variant over-flags Border content, a topic-intrinsic source where surface border vocabulary co-occurs with negative valence regardless of stance. Independently, GPT-D and Qwen-DT diverge on metaphor-heavy topics under the same interface: on Rejection Rhetoric (n=22), GPT-D matches the human rate exactly while Qwen-DT over-detects by 36 percentage points, with the effect extending to large clusters (Welfare under Qwen-DT: gap +15.9).

5 Discussion

5.1 Why the Decomposed Interface Works

The decomposed interface, not just LLM scale, drives the result. GPT-H at 26K labels ties with Full-Human on F1; GPT-D exceeds Full-Human by 5–6 F1 points across encoders. The topical analysis (§4.6) suggests a mechanism: the holistic prompt applies a co-occurrence shortcut where immigration-adjacent words plus negative valence triggers an ANTI label; the Q1 (reference) / Q2 (valence) decomposition's forced separate judgment breaks this shortcut. Policy-critique or economic-anxiety comments (labeled by humans as NOT ANTI-IMMIGRANT), pass Q1 with a NO and are correctly classified. The same decomposition was originally introduced for human annotation to address low inter-annotator agreement (Fleiss $\kappa = 0.175$ under a holistic prompt; §3.2): the structural argument applies to both annotator types.

5.2 The Scale-Quality Trade-off

At costs relevant for resource-constrained research, LLM-at-scale is decisively favorable: Full-LLM-26K (GPT-D) at $28 outperforms Full-Human at $316 across all four encoders. At a comparable cost to AL-Human's 530 labels ($44), Full-LLM-26K provides 49× more labels and 0.075–0.085 higher F1-Macro. The advantage is robust under soft-label evaluation (§4.4): Full-LLM-26K (GPT-D) wins every cell under BB and DS posteriors, indicating alignment with the underlying distribution of human judgments rather than an artifact of majority-vote aggregation.

5.3 Two LLMs, Two Distinct Error Structures

Despite similar aggregate F1, the two LLM annotators produce structurally different errors: GPT-D yields a near-human FP:FN ratio with halved FNs, while Qwen-DT skews FP-heavy (7.5) and over-flags metaphor-heavy topics. Both LLMs produce errors at higher confidence than human-supervised classifiers, roughly doubling high-confidence errors. This impacts deployment: pipelines that triage low-confidence predictions for human review will surface fewer errors needing review for classifiers trained on LLM labels. The topical analysis (§4.6) adds two qualifications: Border Control content is over-flagged by all four LLMs (a topic-intrinsic limitation no interface or model choice resolves), and GPT-D and Qwen-DT diverge by 25–36 points on metaphor-heavy clusters. The choice of LLM, not only the interface, materially affects which discourse themes are flagged.

5.4 Active Learning in Pre-Enriched Pools

The Llama-prefiltered pool already concentrates task-relevant content, reducing the marginal benefit of acquisition: Core-Set and BADGE in particular fared worse than random, consistent with the absence of meaningful diversity. When pool construction does relevance work, acquisition complexity provides limited additional benefit, though this may not generalize to unfiltered pools where relevance is itself the acquisition target.

6 Conclusion

We asked whether LLM annotation at scale can replace human-in-the-loop AL for training subjective classifiers. Under matched interfaces, the answer is yes: LLM annotation produces stronger classifiers than human supervision at one-tenth the cost, with both closed-source (GPT-5.2) and open-weight (Qwen3.5-122B) LLMs, and the result is robust under soft-label evaluation. Three conditions on this finding matter for practitioners. Interface is the prerequisite: decomposed unlocks the cost advantage that holistic does not. The LLM choice shapes error structure: GPT-D preserves a near-human FP:FN balance that Qwen-DT does not. And AL provides little advantage over random sampling in pre-filtered pools. Humans remain essential as gold-standard evaluators; for training labels at scale, the question has shifted from "human labels or LLM labels?" to "which LLM, which interface, and what shape of pool?".

7 Limitations

Pre-enriched annotation pool. Our pool is constructed via LLM-based topic prefiltering, which concentrates content around the target domain and reduces the marginal benefit of uncertainty- and diversity-based acquisition. Our AL-versus-random finding (§4.3) should be interpreted in this context and may not transfer to unfiltered pools.

Simulated active learning. We simulate the AL loop by pre-collecting labels and revealing them when queried, standard practice in pool-based AL evaluation. A live deployment may differ if annotators calibrate across rounds.

Single task and language. All experiments concern a single binary task in German on TikTok. The three-pathway taxonomy (§4.6) may transfer to other subjective tasks, but the specific clusters that fall into each pathway will not.

Two LLM annotators. Stage 2 annotation uses two frontier LLMs (closed-source GPT-5.2 and open-weight Qwen3.5-122B-10B). The model-specific pathway in §4.6 relies on this two-LLM contrast; additional models or alternative interface designs may surface further variation.

Crowdsourced rather than expert annotation. Human annotations were collected from Prolific crowdworkers, not domain experts. Crowdworker judgements may differ systematically from trained experts on borderline cases, and expert annotation could shift the decision boundary and alter the human-versus-LLM comparison.

Evaluation against the human gold standard. All conditions are evaluated against the majority-vote human label, which structurally disadvantages classifiers whose decision boundary diverges from the human one. Soft-label evaluation (§4.4) partially addresses this; expert adjudication on the contested clusters in §4.6 would disentangle genuine model errors from evaluation artifacts.

Annotator demographics. Our six annotators are predominantly white, all resident in Germany, and not pre-screened for political orientation (App. B). A more diverse panel might shift the gold standard, particularly on contested topics (Sap et al., 2022).

Single round of LLM annotation. Each LLM annotates each comment once at temperature 0. Aggregating multiple samples (e.g., self-consistency) might reduce model-specific topical bias but increases cost; we did not explore this trade-off.

In-distribution evaluation only. Our evaluation set is sampled from the same corpus and time window as the training pool. Performance on out-of-distribution data (later time slices, other platforms, other German political contexts) may differ. Whether the three-pathway taxonomy (§4.6) transfers to OOD evaluation is an open question.

Cost figures are a snapshot. Crowdwork rates and LLM API pricing both evolve; absolute dollar comparisons reflect late-2025 pricing and should be interpreted as illustrative of the current cost structure rather than as fixed ratios.

8 Ethical Considerations

Annotator exposure and compensation. Our annotation task required crowdworkers to read and classify comments containing xenophobic language, stereotypes, and dehumanizing rhetoric. Annotators were informed about the nature of the content before accepting the task, could pause or withdraw at any time without penalty, and were compensated at Prolific's recommended "good rate" of £9/hour regardless of completion speed (full cost breakdown in §4.2 and App. C.4). We did not use rejection-based payment models. Nonetheless, sustained exposure to hostile content carries documented psychological risks (Steiger et al., 2021), and reducing the need for human annotation of harmful content is an explicit motivation for this work.

Data collection and privacy. Comments were collected via the TikTok Research API, in accordance with its terms of service for academic research. All comments are publicly posted content. We do not collect or store user profile information beyond the comment text and the party account under which it appeared. Usernames and user IDs are stripped from the released dataset. The annotation study was conducted with informed consent from all annotators.

Risks of automated content classification. Classifiers trained to detect anti-immigrant content carry dual-use risks. In moderation contexts, they may improve platform safety; without appropriate human oversight, they may suppress legitimate political speech, including immigration-policy criticism, particularly given the system-

atic over-detection patterns we document in Sections 4.5 and 4.6 (high-confidence false positives concentrated on policy-critique content). Our findings on FP : FN asymmetry and confidence inflation are intended to inform deployment decisions, not to enable unsupervised filtering of political speech. Released code and data are intended for research; downstream deployment should include human review of flagged items, particularly in topic clusters with high human-LLM disagreement.

Use of generative tools. We used LLM-based tools to assist with literature retrieval in the related work section and as coding assistants during experiment implementation. All generated content was verified by the authors.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *J. Comput. Soc. Sci.*, 8(1):17.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035. SIAM.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dennis Assenmacher, Paloma Piot, Katarina Laken, David Jurgens, and Claudia Wagner. 2025. Beyond the explicit: A bilingual dataset for dehumanization detection in social media. *CoRR*, abs/2510.18582.

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.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. Assoc. Comput. Linguistics*, 10:92–110.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Paul Engelmann, Peter Brunsgaard Trolle, and Christian Hardmeier. 2024. A dataset for the detection of dehumanizing language. *CoRR*, abs/2402.08764.

Yaron Fairstein, Oren Kalinsky, Zohar Karnin, Guy Kushilevitz, Alexander Libov, and Sofia Tolmach. 2024. Class balancing for efficient active learning in imbalanced datasets. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 77–86.

Jan Fillies, Esther Theisen, Michael Hoffmann, Robert Jung, Elena Jung, Nele Fischer, and Adrian Paschke. 2025. A novel german tiktok hate speech dataset: far-right comments against politicians, women, and others. *Discover Data*, 3(1):4.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, JMLR Workshop and Conference Proceedings, pages 1050–1059. JMLR.org.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 165–190. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 6155–6218. Association for Computational Linguistics.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track - European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9-13, 2024, Proceedings, Part X*, Lecture Notes in Computer Science, pages 397–412. Springer.

Aida Kostikova, Dominik Beese, Benjamin Paassen, Ole Pütz, Gregor Wiedemann, and Steffen Eger. 2024. Fine-grained detection of solidarity for women and migrants in 155 years of german parliamentary debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5884–5907. Association for Computational Linguistics.

Woosuk Kwon. 2025. *vLLM: An Efficient Inference Engine for Large Language Models*. Ph.D. thesis, UC Berkeley.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 21–30. Association for Computational Linguistics.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.

Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521, Vienna, Austria. Association for Computational Linguistics.

Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Miserlis Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2025. Can large language models capture human annotator disagreements? *CoRR*, abs/2506.19467.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. *CoRR*, abs/2306.00176.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 2025–2039. AAAI Press.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10671–10682. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Julia Romberg, Christopher Schröder, Julius Gonsior, Katrin Tomanek, and Fredrik Olsson. 2026. Reassessing active learning adoption in contemporary NLP: A community survey. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 2621–2647. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 175–190. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5884–5906. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017*, pages 1–10. Association for Computational Linguistics.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *CoRR*, abs/2008.07267.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Burr Settles. 2009. Active learning literature survey.

Ard Snijders, Douwe Kiela, and Katerina Margatina. 2023. Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2187–2209, Dubrovnik, Croatia. Association for Computational Linguistics.

Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 341:1–341:14. ACM.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Petter Törnberg. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195.

Sebastián Vallejo Vera and Hunter Driggers. 2025. Llms as annotators: the effect of party cues on labelling decisions by large language models. *Humanities and Social Sciences Communications*, 12(1):1–11.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2025. Aggregating soft labels from crowd annotations improves uncertainty estimation under distribution shift. *PLoS One*, 20(6):e0323064.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, Zhengmian Hu, Yue Zhao, Nedim Lipka, Seunghyun Yoon, Ting-Hao Kenneth Huang, Zichao Wang, Puneet Mathur, Soumyabrata Pal, Koyel Mukherjee, Zhehao Zhang, Namyong Park, Thien Huu Nguyen, Jiebo Luo, Ryan A. Rossi,

and Julian J. McAuley. 2025. From selection to generation: A survey of llm-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14552–14569. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 13088–13103. Association for Computational Linguistics.

A Dataset and Preprocessing

Table 5 reports per-party raw video and comment counts from the TikTok Research API collection (January 2024 to September 2025), prior to preprocessing. Table 6 reports per-step removal counts during preprocessing.

Party	Videos	Comments	Comm./video
AfD	259	131,541	508
CDU	585	149,399	255
SPD	583	70,366	121
Linke	437	84,701	194
Grüne	360	23,599	66
FDP	326	8,054	25
Total	**2,550**	**467,660**	avg 183

Table 5: Per-party raw video and comment counts from the TikTok Research API, January 2024 to September 2025. Accounts: AfD = @afdfraktionimbundestag, CDU = @insidecdu, SPD = @deinespd, Linke = @die.linke, Grüne = @diegruenen, FDP = @fdpbt.

Step	Removed
Duplicate IDs	20,390
Emoji-only	86,883
Empty	1,989
Mention-only	1,273
URL-only	447
Hashtag-only	406
Number-only	783
Laughter-only	270
Duplicate text (within-party)	28,543
Duplicate text (cross-party)	4,068
Non-German (GlotLID)	44,706
Total removed	**189,758**
Remaining	**277,902**

Table 6: Per-step preprocessing counts applied to the 467,660 raw comments. GlotLID (Kargaran et al., 2023) is used for language identification to retain only German-language comments.

Comments are short: median 5 words in the raw corpus (mean 8.3, 95th percentile 27), rising to a median of 13 words (mean 16.3, 95th percentile 34) in the 25,974-comment immigration-relevant subset retained after Llama prefiltering. The 95th percentile in the prefiltered subset motivates the 128-token truncation used in §3.7.

B Human Annotation Study

Table 7 summarises the phase structure introduced in §3.4.

Phase	Overlap	Items	Use
Test	6-way	50	Eval
Pilot	6-way	200	Eval
Main (triple)	3-way	950	Eval
Main (single)	1-way	3,800	Train
Total eval		1,200	
Total train		3,800	

Table 7: Annotation study structure with 6 annotators. Multiply-annotated items form the gold evaluation set; single-annotated items form the AL training pool.

B.1 Annotator Demographics

The six annotators retained across all study phases (§3.4) were recruited through Prolific. They are referred to throughout this appendix as A1, A2, A3, A4, A5, and A8 (Annotator_6 and Annotator_7 were excluded after the qualification round, see App. B.3; original identifiers are preserved for transparency). Ages ranged from 19 to 60 (mean: 40). Four annotators identified as female and two as male; five identified as White and one as Asian. All six were resident in Germany at the time of the study. Four were native German speakers, and two were highly proficient non-native speakers with first languages Portuguese and Japanese. Employment status comprised three full-time workers, one part-time worker, one unemployed, and one not in paid work. Prolific approval counts ranged from 31 to 1102 (median: 158), indicating a mix of relatively new and experienced platform workers.

Political orientation was not included in the pre-screening criteria. This is a limitation, since annotator political leanings may correlate with how anti-immigrant content is interpreted (Vallejo Vera and Driggers (2025) for parallel findings on LLM annotators). Future replications should consider stratified pre-screening on political orientation to characterize this potential source of bias.

B.2 Holistic-Prompt Pilot and the Two-Question Framework

The two-question decomposition (§3.2) was adopted following an initial holistic-prompt pilot. Ten Prolific annotators received the binary labeling task ("Is this comment anti-immigrant: YES/NO?") with the same written guidelines later used in the main study, at a scale comparable to the qualification round. The pilot yielded a Fleiss $\kappa = 0.175$, indicating very low agreement. Manual review suggested that annotators inconsistently applied the boundary between policy critique and group-directed hostility; the single-question format did not surface this distinction during annotation. We therefore designed the main study with the label decomposed into two sequential yes/no questions (Q1: presence of immigrant reference; Q2: negative portrayal), scaffolding the decision and forcing engagement with each criterion. The main study used a freshly recruited cohort of 8 annotators, of whom 6 were retained after qualification (App. B.3). Pilot costs are not included in the cost analysis of §3.8; they were comparable to the qualification round.

B.3 Attention Checks

Instruction manipulation checks (IMCs) were interleaved throughout all phases at densities of approximately 10 per 100 items (test), 5 per 100 (pilot), and 3 per 100 (main). Each check instructs the annotator to select specific Q1 and Q2 responses, testing whether comment text is being read. A minimum of two checks was enforced per annotator per phase.

In the qualification test, Annotator_7 failed to meet the $\geq$80% accuracy threshold on attention checks (2/3, 67%) and was excluded. Annotator_6 passed attention checks but showed extremely low inter-annotator agreement on the valence question (average pairwise $\kappa = 0.13$ on Q2) and was also excluded. The remaining six annotators achieved 83–100% accuracy on attention checks across all phases, with all six scoring 100% in the main round.

B.4 Item Assignment

In the main round, 950 items received triple annotation for the gold evaluation set. Items were distributed via round-robin assignment over all $\binom{6}{3} = 20$ possible 3-annotator combinations to ensure balanced workload and annotator-pair coverage. The remaining 3,800 items received single annotation and were distributed evenly across annotators.

B.5 Inter-Annotator Agreement

Tables 8 and 9 report agreement statistics for the pilot and main phases. Tables 10 and 11 provide full pairwise Cohen's κ matrices for the derived binary label.

Metric	Q1	Q2	Label
Krippendorff's α	0.574	0.493	0.494
Cohen's κ (avg)	0.580	0.488	0.490
Cohen's κ range	.356–.807	.310–.758	.320–.741

Table 8: Inter-annotator agreement on the pilot phase (200 items, 6-way overlap).

Metric	Q1	Q2	Label
Krippendorff's α	0.591	0.432	0.433
Cohen's κ (avg)	0.599	0.437	0.439
Cohen's κ range	.384–.704	.221–.588	.221–.588

Table 9: Inter-annotator agreement on the main phase (950 triple-annotated items, 3-way overlap). Krippendorff's α is used as Fleiss' κ is not applicable for variable annotator subsets.

	A1	A2	A3	A4	A5	A8
A1	—	.470	.515	.518	.386	.320
A2	.470	—	.738	.741	.545	.408
A3	.515	.738	—	.615	.551	.395
A4	.518	.741	.615	—	.425	.351
A5	.386	.545	.551	.425	—	.363
A8	.320	.408	.395	.351	.363	—

Table 10: Pairwise Cohen's κ for the derived binary label on the pilot phase (200 items, 6-way overlap).

	A1	A2	A3	A4	A5	A8
A1	—	.535	.419	.272	.358	.357
A2	.535	—	.588	.471	.560	.480
A3	.419	.588	—	.536	.487	.521
A4	.272	.471	.536	—	.416	.359
A5	.358	.560	.487	.416	—	.221
A8	.357	.480	.521	.359	.221	—

Table 11: Pairwise Cohen's κ for the derived binary label on the main phase (950 triple-annotated items, 3-way overlap). Each pair is computed over items where both annotators provided labels.

B.6 Per-Annotator Statistics

Table 12 reports per-annotator average κ (with all other annotators) and ANTI-IMMIGRANT label

rates across the pilot and main phases. Annotator_1 shows a consistently higher positive label rate (34–37%) than other annotators (16–29%), reflecting a more inclusive interpretation of what constitutes anti-immigrant content.

	Pilot		Main	
Annotator	$\bar{\kappa}$	ANTI%	$\bar{\kappa}$	ANTI%
A1	.442	34.0	.388	37.0
A2	.581	26.5	.527	28.7
A3	.563	29.0	.510	21.6
A4	.530	30.0	.411	16.5
A5	.454	18.5	.408	16.3
A8	.367	15.5	.388	17.0

Table 12: Per-annotator average Cohen's κ (on the derived binary label, averaged over all pairwise comparisons) and ANTI-IMMIGRANT label rate, by study phase.

B.7 Class-Conditional Agreement

Agreement is higher for the majority class across both phases. On the pilot (200 items, majority vote), items labeled NOT ANTI-IMMIGRANT (n=156) show 90.2% average agreement with 63% unanimity, while items labeled ANTI-IMMIGRANT (n=44) show 81.4% average agreement with 30% unanimity. On the main round (950 triple-annotated items), agreement is high overall: 96.3% for ANTI-IMMIGRANT items (89% unanimous) and 98.5% for NOT ANTI-IMMIGRANT items (96% unanimous), reflecting that the main round's 3-way overlap yields higher raw agreement than the pilot's 6-way overlap.

C LLM Annotation Pipeline

C.1 Stage 1: Topic Prefiltering Prompt

Llama-3.3-70B-Instruct Topic Prefiltering

```
Does this German TikTok comment discuss immigration, migrants, refugees, asylum seekers, foreigners, or related policies?

Output JSON: {"label": "YES" or "NO"}

Comment: {comment}
```

Figure 5: Prefiltering prompt applied to all 277,902 comments (temperature = 0, JSON output).

C.2 Stage 2: Holistic Annotation Prompt (Interface Ablation)

GPT-5.2 Anti-Immigrant Annotation (Holistic)

```
You are an expert in sentiment analysis and political discourse, specializing in German-language text. Your task is to classify TikTok comments posted under videos from German political parties.

Classify the comment as either "ANTI-IMMIGRANT" or "NOT ANTI-IMMIGRANT".

Definitions:
- ANTI-IMMIGRANT: Expresses negative sentiments, hostility, fear, or opposition towards immigrants, refugees, asylum seekers, or immigration policies. Includes xenophobic language, stereotypes, calls for stricter borders, deportation advocacy, or blaming immigrants for social/economic issues. Subtle implications count if they clearly lean negative.
- NOT ANTI-IMMIGRANT: Neutral, positive towards immigrants/immigration, or unrelated to immigration altogether.

Guidelines:
- Analyze the comment in its original German.
- Base classification solely on the comment's content, not party context.
- If ambiguous or sarcastic, default to "NOT ANTI-IMMIGRANT".
- Base decisions only on visible text; no external inference.

Output JSON: {"label": "ANTI-IMMIGRANT" or "NOT ANTI-IMMIGRANT"}

Comment: {comment}
```

Figure 6: Holistic classification prompt applied to 25,974 prefiltered comments via GPT-5.2 Batch API (temperature = 0, structured JSON output). Used for the GPT-H interface ablation.

C.3 Stage 2: Decomposed Annotation Prompt

This prompt is used by GPT-D, Qwen-DT, and Qwen-DNT. The system message specifies the JSON schema; the user message provides the annotation guidelines (translated from the Argilla interface used by human annotators in §3.4) followed by the comment to annotate. Serving details for each model (endpoint, sampling parameters, reasoning-mode flag) are in §C.4.

Label derivation. The binary label is derived deterministically from the JSON response, outside the model: ANTI-IMMIGRANT $\Leftrightarrow q_1 = \text{YES} \wedge q_2 = \text{YES}$; otherwise NOT ANTI-IMMIGRANT. Items for which the model exceeds the output token budget before emitting valid JSON are excluded from downstream classifier training; evaluation on the human-labeled test set is unaffected. Per-model counts are reported in §C.4.

C.4 Infrastructure Details

Llama-3.3-70B-Instruct (Stage 1). The model was run locally on four NVIDIA A100 80GB GPUs using vLLM (Kwon, 2025) with tensor parallelism across all four devices, bfloat16 precision, and a maximum sequence length of 512 tokens. Filtering all 277,902 comments completed in 0.55 wall-clock hours (2.20 GPU-hours total) at a throughput of approximately 140 comments/second. At the Lambda Labs A100 on-demand rate of $2.79/GPU-hour, the cloud-equivalent cost is ≈$6.14 USD. Researchers without GPU access can replicate Stage 1 via the DeepInfra API (`meta-llama/Llama-3.3-70B-Instruct`) at approximately $5.28 for the same job ($0.10/1M input tokens, $0.40/1M output tokens; 41.7M input and 2.8M output tokens estimated).

GPT-5.2 (Stage 2). We submitted all 25,974 prefiltered comments through the OpenAI Batch API. Each request comprised the decomposed annotation prompt (§C.3) plus the comment text, yielding approximately 8.1M input tokens and 2.1M output tokens across the run. The Batch API charges 50% of standard rates and processes within a 24-hour window; the total invoice was **$28 USD** (≈$1.08 per 1,000 comments). All 25,974 comments received valid JSON responses with no content-policy refusals.

Qwen3.5-122B-A10B (Stage 2). We additionally annotated the same 25,974 prefiltered comments with Qwen3.5-122B-A10B, an open-weight 122B-parameter mixture-of-experts model with ∼10B active parameters per token, distributed under Apache-2.0. The model was served locally on four NVIDIA A100 80GB GPUs using vLLM (Kwon, 2025) with tensor parallelism, bfloat16 precision, and the same decomposed prompt as GPT-5.2 (§C.3). Two configurations were run: the main QWEN-DT variant with reasoning enabled (`enable_thinking=True`, max output 16,384 tokens), which exceeded the output token budget on 136 of 25,974 comments (0.5%); and the ablation QWEN-DNT variant with reasoning disabled (max output 4,096 tokens), which exceeded the budget on a comparable share of comments (∼0.5%). The reasoning-enabled run completed in 26 GPU-hours; the no-thinking ablation in 2.5 GPU-hours. At the Lambda Labs A100 on-demand rate of $2.79/GPU-hour, the cloud-equivalent costs are ≈$73 (Qwen-DT) and ≈$7 (Qwen-DNT); on owned hardware, the marginal cost is electricity only.

Human Annotation Cost Breakdown. Annotators were compensated at £9/hour across three phases. Phase-level costs: £18 (test, 20 min per annotator × 6) + £54 (pilot, 60 min × 6) + £162 (main, 3 hours × 6) = £234 base compensation. With Prolific's 33.3% academic platform fee (£77.92), the grand total was £312 ($416 USD). Per-condition costs are computed proportionally: FULL-HUMAN uses 3,800 of 5,000 labels ($3{,}800/5{,}000 \times \$416 = \$316$); AL-HUMAN uses a maximum of 530 ($530/5{,}000 \times \$416 = \$44$).

C.5 LLM Label Distributions

Table 13 reports the per-party distribution of ANTI-IMMIGRANT labels under each of the four LLM annotators. Two observations are worth flagging. First, CDU/CSU comments receive the highest ANTI-IMMIGRANT rate among the four major parties under every LLM variant (24.9% to 37.6%), exceeding AfD (21.6% to 29.7%) in each case. A plausible explanation is that CDU/CSU video topics often center on migration policy debates, drawing migration-themed comment threads from both supportive and critical audiences. Second, the decomposed interface produces roughly 8 to 10 fewer percentage points of ANTI-IMMIGRANT labels than the holistic interface across all six party segments, with the relative ordering of parties preserved: the Q1/Q2 decomposition functions as a roughly uniform correction rather than selectively

Decomposed Annotation

```
System:
You are an annotator answering two questions about a German TikTok comment posted under a video
of a German political party.

You see only the comment text, with no video context.

Return your answer as a single JSON object with these fields:
- "q1": "YES" or "NO"
- "q1_reasoning": one short sentence in German
- "q2": "YES" or "NO"
- "q2_reasoning": one short sentence in German

User:
Annotationsrichtlinie:

# TikTok-Kommentare - Annotationsrichtlinie

## 1. Aufgabenübersicht
Sie werden deutsche TikTok-Kommentare, die unter Videos deutscher politischer Parteien gepostet
wurden, anhand von **zwei Fragen** bewerten.

**Wichtig:**
- Sie sehen nur den Kommentartext, ohne Video-Kontext

## 2. Die zwei Fragen

### Q1: Bezieht sich der Kommentar auf Einwanderer, Flüchtlinge, Asylsuchende, Ausländer oder
Einwanderung/Asylpolitik?
- **YES:** Der Kommentar bezieht sich auf Einwanderer, Flüchtlinge oder Asylsuchende - sei es
direkt, durch Slang/Codewörter oder durch typische Begriffe aus der Migrationsdebatte
- **NO:** Kein Bezug zu Einwanderung oder Einwanderergruppen

### Q2: Werden Einwanderer/Flüchtlinge/Ausländer in diesem Kommentar abwertend dargestellt -
durch Feindseligkeit, Angst, Ablehnung, fremdenfeindliche Sprache, Stereotype, Forderungen nach
strengeren Grenzen, Abschiebungsrhetorik oder Schuldzuweisungen für soziale/wirtschaftliche
Probleme?
- **YES:** Feindseligkeit, Angst, Ablehnung, Stereotype, Abschiebungsrhetorik oder
Schuldzuweisungen für soziale/wirtschaftliche Probleme
- **NO:** Der Kommentar ist sachlich, neutral, positiv gegenüber Einwanderern/Einwanderung, oder
hat keinen Bezug zu Einwanderung.

Kommentar:
"{comment}"

Antworten Sie mit einem JSON-Objekt im folgenden Format:
{
  "q1": "YES" or "NO",
  "q1_reasoning": "...",
  "q2": "YES" or "NO",
  "q2_reasoning": "..."
}
```

Figure 7: Decomposed annotation prompt, identical across the three decomposed LLM variants. The user-message guidelines are verbatim from the Argilla interface used by human annotators (§3.4).

reducing labels in particular discourse contexts.

D Models and Acquisition Functions

D.1 Encoder Model Details

Table 14 summarizes the four pre-trained encoders evaluated in our supervised pipeline. All four are encoder-only BERT-style transformers fine-tuned with a randomly initialized binary classification head; parameter counts cover the full pre-trained encoder and exclude the two-class output head ($2 \times 768 + 2 = 1{,}538$ parameters for the base-sized models). Three of the four models are German-specific, with xlm-r-base serving as a multilingual baseline. All four encoders are fine-tuned end-to-end with identical training hyperparameters (§3.7): all encoder parameters are updated jointly with the binary classification head, with no frozen layers.

All downstream classifier training was run on NVIDIA RTX A6000 48GB GPUs (one GPU per training run; up to 8 A6000s dispatched in parallel). The complete experimental sweep comprises 1,360 training runs across 10 random seeds and 4 encoders for each of: FULL-HUMAN, FULL-LLM (3.8K) under 4 LLM variants, FULL-

Party	n_{PARTY}	GPT-D		GPT-H		Qwen-DT		Qwen-DNT	
		n	%	n	%	n	%	n	%
AfD	9,379	2,022	21.6	2,691	28.7	2,786	29.7	2,780	29.6
CDU/CSU	7,316	1,821	24.9	2,508	34.3	2,557	35.0	2,753	37.6
SPD	4,324	981	22.7	1,430	33.1	1,384	32.0	1,426	33.0
Linke	3,311	570	17.2	831	25.1	841	25.4	928	28.0
Grüne	1,244	144	11.6	246	19.8	260	20.9	274	22.0
FDP	400	113	28.2	141	35.2	133	33.2	154	38.5
Total	25,974	5,651	21.8	7,847	30.2	7,961	30.6	8,315	32.0

Table 13: Per-party distribution of ANTI-IMMIGRANT labels across the four LLM annotation variants on the 25,974-comment Llama-prefiltered pool (complementing Table 1 in the main paper). n_{PARTY} is the number of comments from each party's TikTok account; n and % report the count and share of those comments labeled ANTI-IMMIGRANT under each LLM. Counts and percentages are computed over the full 25,974-comment pool; Qwen-DT and Qwen-DNT each excluded 136 items (0.5%) where the output exceeded the token budget (§C.4). GPT-D and GPT-H denote the decomposed and holistic GPT-5.2 prompts; Qwen-DT and Qwen-DNT denote Qwen3.5-122B-A10B with and without enabled thinking.

Model	Params	Pretraining language(s)
german-bert	110M	German
ModernGBERT	134M	German
gbert-base	110M	German
xlm-r-base	278M	100 languages incl. German

Table 14: Pre-trained encoder models. All four are fine-tuned end-to-end with identical training hyperparameters (§3.7). Parameter counts exclude the binary classification head. Model sources: german-bert, ModernGBERT, gbert-base, xlm-r-base.

LLM-26K under 4 LLM variants, RANDOM-HUMAN, RANDOM-LLM under 4 LLM variants, AL-HUMAN under 4 acquisition strategies, and AL-LLM under 4 acquisitions $\times$ 4 LLM variants. Total downstream compute is **244.6 A6000 GPU-hours**, with AL-LLM (109.2 GPU-hours; 45%) and FULL-LLM-26K (82.9 GPU-hours; 34%) accounting for the bulk of the footprint. Per-backbone totals are 53.8 (`german_bert`), 74.1 (`ModernGBERT`), 57.0 (`gbert-base`), and 59.7 (`xlm-r-base`) GPU-hours. Stage 1 (Llama prefiltering, 4×A100 80GB) and Stage 2 (Qwen3.5 annotation, 4×A100 80GB) add a further ∼31 A100 GPU-hours on a separate machine; GPT-5.2 annotation runs on OpenAI's hosted infrastructure and is not included in the local GPU-hour total.

D.2 Active Learning Acquisition Functions

We evaluate four pool-based acquisition strategies, two uncertainty-based (Entropy, BALD) and two diversity-based (Core-Set, BADGE), covering the four most widely used acquisition families in the AL literature. At round t, all four operate on the current classifier f_{θ_t}, the labeled set $\mathcal{L}_t$, and the unlabeled pool $\mathcal{U}_t$, and select the top-k instances (with $k = 50$ in our setup) to be added to $\mathcal{L}_{t+1}$.

Entropy. The predictive entropy of the classifier's softmax output over the two classes:

$$\text{Score}_{\text{Ent}}(x) = -\sum_{c \in \{0,1\}} p_\theta(c \mid x) \log p_\theta(c \mid x).$$

The 50 instances with the highest score are selected. Entropy is the standard uncertainty baseline and reduces to selecting instances closest to the decision boundary in our binary setting.

BALD (Houlsby et al., 2011). Bayesian Active Learning by Disagreement selects instances on which an ensemble of stochastic predictions disagrees most:

$$\text{Score}_{\text{BALD}}(x) = H[\bar{p}(\cdot \mid x)] - \mathbb{E}_{\theta'}\big[H[p_{\theta'}(\cdot \mid x)]\big],$$

where $\bar{p}$ is the average prediction across stochastic forward passes and $H[\cdot]$ is the entropy. We estimate the ensemble via $T = 10$ MC Dropout (Gal and Ghahramani, 2016) passes through the classifier with dropout enabled at inference time; the inner expectation is approximated as the mean entropy across the T passes. BALD prioritizes instances where individual stochastic predictions disagree (high $H[\bar{p}]$) but each prediction is itself sharp (low expected entropy), capturing epistemic uncertainty rather than aleatoric uncertainty.

Core-Set (Sener and Savarese, 2018). A pure diversity-based acquisition that selects a subset of $\mathcal{U}_t$ providing the best coverage of the unlabeled-pool embedding space. We use the k-Center-Greedy algorithm: initialized with the current labeled set $\mathcal{L}_t$, it iteratively selects the unlabeled

instance with maximum Euclidean distance to its nearest already-selected neighbour, repeating until k instances are picked:

$$\text{Score}_{\text{CS}}(x) = \min_{x' \in \mathcal{L}_t \cup \mathcal{S}} \|\phi(x) - \phi(x')\|_2,$$

where $\phi(x)$ is the [CLS] embedding from the classifier's encoder and $\mathcal{S}$ is the set of instances selected so far in the current round. Core-Set does not use the classifier's predictions at all; it picks based on geometry of the representation space.

BADGE (Ash et al., 2020). Batch Active learning by Diverse Gradient Embeddings combines uncertainty and diversity. For each unlabeled instance, the gradient of the cross-entropy loss with respect to the final classifier layer is computed, assuming the model's most likely predicted label as the pseudo-target:

$$g(x) = \nabla_{w_{\text{cls}}} \mathcal{L}\big(f_\theta(x), \hat{y}\big), \quad \hat{y} = \arg\max_c p_\theta(c \mid x).$$

The gradient magnitude captures uncertainty (high for instances the model would update most strongly on), while the gradient direction captures the type of update. BADGE applies k-means++ initialization (Arthur and Vassilvitskii, 2007) on the gradient embeddings $\{g(x) : x \in \mathcal{U}_t\}$ and selects the k chosen centroid instances. This procedure tends to pick instances that are both uncertain and dissimilar from each other.

Implementation details. All acquisitions are computed once per round on the current trained classifier. We use stratified mini-batch processing of the pool to avoid GPU memory issues, with the same batch size as during training (16). Embeddings, entropies, and gradient embeddings are recomputed at every round on the latest classifier. Ties are broken by comment ID order to ensure deterministic acquisition across seeds. For BADGE, we use the scikit-learn `kmeans++` implementation; for Core-Set, our own implementation of k-Center-Greedy.

E Annotation Ablations

This appendix reports full downstream-classifier results under the two annotation ablations: the holistic interface (GPT-H) and the reasoning-disabled Qwen variant (Qwen-DNT). Both share the data splits, training recipe, and evaluation protocol of the main-paper experiments; the only changes are at the LLM-annotation stage.

E.1 Interface Ablation: GPT-H

The holistic (H) interface queries GPT-5.2 with a single prompt asking for the binary label directly, rather than the two-question (Q1 reference, Q2 valence) decomposition used by GPT-D. The H variant labels 7,847 of 25,974 prefiltered comments as ANTI-IMMIGRANT (30.2%), compared with 21.8% under GPT-D (§3.3).

Aggregate performance. Table 15 reports F1-Macro and F1-Anti under the H interface. FULL-LLM-26K (GPT-H) reaches 0.730 to 0.735 F1-Macro across the four backbones, slightly below FULL-HUMAN (0.725 to 0.740) and substantially below FULL-LLM-26K (GPT-D) at 0.787 to 0.790. The performance gap between the two interfaces at full scale is approximately 0.06 F1-Macro per backbone, attributable solely to the prompt design.

Condition	germ.	Mod.	gbert	xlm-r
	F1-Macro			
FULL-LLM (3.8K)	.705	.710	.706	.703
FULL-LLM-26K	.730	.735	.732	.730
AL-LLM (Ent.)	.674	.685	.677	.677
RANDOM-LLM	.670	.668	.664	.677
	F1-Anti			
FULL-LLM (3.8K)	.587	.585	.587	.586
FULL-LLM-26K	.628	.632	.627	.627
AL-LLM (Ent.)	.539	.532	.533	.528
RANDOM-LLM	.517	.531	.520	.531

Table 15: F1-Macro and F1-Anti under the holistic (H) interface; all rows use GPT-H. Mean across 10 seeds.

Error structure. Under the H interface, FULL-LLM-26K produces a FP:FN ratio of 7.54 to 9.89 across backbones, an order of magnitude above FULL-HUMAN (1.11 to 2.43) and approximately 3× FULL-LLM-26K (GPT-D) at 2.66 to 2.74. The H classifiers also under-produce false negatives (21.7 to 27.1 FNs per backbone, compared with 43.8 to 44.7 under GPT-D and 62.9 to 90.3 under Full-Human), confirming that the H interface drives over-prediction of the ANTI-IMMIGRANT class. Counterintuitively, mean confidence on errors is *lower* under the H interface than under GPT-D: 0.752 to 0.767 vs. 0.847 to 0.896. The H interface produces more errors but the model holds them less confidently. Combined with the per-topic gap reductions in §4.6, this identifies the H interface as the primary source of the over-detection patterns.

E.2 Reasoning-Mode Ablation: Qwen-DNT

The Qwen-DNT variant queries Qwen3.5-122B-A10B via the same decomposed prompt as the main-paper Qwen-DT variant, but with the `enable_thinking` flag set to `False`, eliminating the intermediate reasoning trace before the JSON response. The DNT variant labels 8,315 of 25,974 prefiltered comments as ANTI-IMMIGRANT (32.0%), comparable to DT at 30.6% (§3.3).

Aggregate performance. Table 16 reports F1-Macro and F1-Anti under DNT. FULL-LLM-26K (Qwen-DNT) reaches 0.739 to 0.746 F1-Macro across the four backbones, slightly below Qwen-DT at 0.751 to 0.758. Reasoning improves performance by approximately 0.01 F1-Macro on average at full scale, a small effect compared with the GPT-H to GPT-D interface gap.

Condition	germ.	Mod.	gbert	xlm-r
	F1-Macro			
FULL-LLM (3.8K)	.716	.722	.715	.688
FULL-LLM-26K	.744	.746	.744	.739
AL-LLM (Ent.)	.650	.688	.682	.687
RANDOM-LLM	.669	.667	.669	.659
	F1-Anti			
FULL-LLM (3.8K)	.594	.593	.585	.573
FULL-LLM-26K	.639	.642	.640	.633
AL-LLM (Ent.)	.498	.548	.540	.536
RANDOM-LLM	.530	.523	.522	.520

Table 16: F1-Macro and F1-Anti under Qwen-DNT; all rows use Qwen-DNT. Mean across 10 seeds.

Error structure. FULL-LLM-26K (Qwen-DNT) produces a FP:FN ratio of 6.51 to 6.91 across backbones, comparable to but slightly below Qwen-DT at 7.05 to 8.14, and far above the human-supervised baseline (1.11 to 2.43). Mean confidence on errors is also notably lower under DNT (0.743 to 0.834) than under DT (0.841 to 0.936); the share of high-confidence errors falls from 40 to 78% under DT to 19 to 47% under DNT. Disabling reasoning produces classifiers with similar FP:FN structure but substantially better calibration on their errors. This confirms the §4.5 observation that the confidence-inflation effect is specific to the reasoning-enabled Qwen variant rather than to Qwen3.5 as a model family.

F Soft-Label Robustness

This appendix reports the methodology and full results for the soft-label evaluation discussed in §4.4. The hard-label majority-vote ground truth used in the main results obscures annotator disagreement on subjective items; the soft-label evaluation rescores each classifier against two Bayesian posteriors over the per-item labels, weighting items by inter-annotator agreement rather than collapsing them to a single hard label.

F.1 Posterior Construction

Both Bayesian posteriors are computed from raw annotator votes on the 1,200 evaluation items (3 to 6 annotators per item).

Beta-Binomial (BB). Applies a Laplace prior to the ANTI-IMMIGRANT vote count: $q_i = (k_i{+}1)/(N_i{+}2)$, where k_i is the number of positive votes and N_i the number of annotators on item i. This treats annotators as exchangeable Bernoulli voters.

Dawid-Skene (DS). Jointly estimates per-annotator confusion matrices π_j and per-item posteriors μ_i via EM with a Dirichlet prior ($\alpha = 1$) on each π_j, treating each (phase, annotator) pair as a distinct worker. EM converges within approximately 50 iterations on our data.

Annotator exclusions. Posterior estimation uses the same exclusions applied to the hard-label gold standard (§B.3): Annotator_6 and Annotator_7. The BB and DS posteriors are therefore computed over the same vote pool that defines the majority vote.

Soft-label metrics. Expected confusion-matrix entries are computed as $\mathbb{E}[\text{TP}] = \sum_i q_i \cdot \mathbf{1}[\hat{y}_i = 1]$, $\mathbb{E}[\text{FP}] = \sum_i (1 - q_i) \cdot \mathbf{1}[\hat{y}_i = 1]$, $\mathbb{E}[\text{FN}] = \sum_i q_i \cdot \mathbf{1}[\hat{y}_i = 0]$, and $\mathbb{E}[\text{TN}] = \sum_i (1 - q_i) \cdot \mathbf{1}[\hat{y}_i = 0]$, where q_i is the soft label under the relevant posterior. Precision, recall, and F1 are computed from these expected counts in the usual way; macro-F1 averages over the two classes.

F.2 Per-Backbone Soft-Label Evaluation

Table 17 reports the full per-backbone soft-label evaluation discussed in §4.4, confirming that the pattern (Full-LLM-26K under both LLM annotators outperforms Full-Human under both Bayesian posteriors) holds across all four encoders.

G Error Structure

This appendix reports the full error-structure analysis for the three full-pool conditions discussed in §4.5: FP:FN ratios and confidence on errors (Table 18 and Figure 8). Equivalent error-

Condition	Hard F1-Macro				BB exp-F1				DS exp-F1			
	germ.	Mod.	gbert	xlm-r	germ.	Mod.	gbert	xlm-r	germ.	Mod.	gbert	xlm-r
FULL-HUMAN	.740	.726	.735	.725	.613	.600	.614	.617	.735	.714	.732	.728
FULL-LLM, GPT-D	.736	.760	.736	.724	.620	.622	.620	.619	.742	.753	.740	.731
FULL-LLM-26K, GPT-D	**.789**	**.787**	**.790**	**.788**	**.643**	**.642**	**.644**	**.644**	**.793**	**.790**	**.794**	**.793**
FULL-LLM, Qwen-DT	.708	.724	.716	.692	.616	.620	.619	.608	.724	.733	.732	.708
FULL-LLM-26K, Qwen-DT	.752	.758	.755	.751	.641	.642	.641	.640	.776	.780	.778	.775

Table 17: Soft-label evaluation under Beta-Binomial (BB) and Dawid-Skene (DS) posteriors of the eval-set labels, all four encoder backbones. Hard F1 included for reference. Mean across 10 seeds. **Bold**/underline = best/second-best per backbone per metric.

structure analyses for the interface ablation (GPT-H) and reasoning-mode ablation (Qwen-DNT) are in App. E.1 and App. E.2.

G.1 FP:FN Ratios and Confidence Summary

The three full-pool conditions separate cleanly across all four encoders. FULL-HUMAN maintains a near-balanced FP:FN ratio (1.11 to 2.43) and moderate confidence on errors (0.691 to 0.782). FULL-LLM-26K (GPT-D) preserves a similarly low ratio (2.66 to 2.74) but pushes mean confidence on errors substantially higher (0.847 to 0.896), with over half of all errors made at confidence above 0.90 on three of four encoders. FULL-LLM-26K (Qwen-DT) departs most sharply on the FP:FN dimension: ratios climb to 7–8, indicating heavily FP-dominant errors, while confidence on errors reaches 0.84–0.94 with up to 78% of errors above 0.90 confidence on ModernGBERT.

G.2 Confidence-Ranked Errors

Figure 8 visualizes the same pattern as confidence-ranked errors on ModernGBERT. Each curve plots the predicted-class confidence of all misclassified instances in descending order; the horizontal line at 0.90 marks the high-confidence threshold, and the intersection of each curve with this line gives the %>.9 value from Table 18. The LLM-supervised curves remain well above 0.90 for most error instances, while FULL-HUMAN crosses 0.90 early and tapers toward 0.50 for the long tail.

G.3 BERTopic Cluster Structure

Figure 9 shows the t-SNE projection of all 25,974 immigration-relevant comments, clustered into 22 named discourse topics via BERTopic (Grootendorst, 2022). Each cluster is labeled by its dominant keyword pattern. Topics span explicit hostility (C10: Criminal Violence & Public Safety; C14: Islam & Islamization), coded political language (C05: Merkel/CDU & 2015 Legacy; C16: Female Politicians), policy-adjacent discourse (C02: Democracy & Governance; C06: Migration Policy & Terminology), and international framing (C11: Racism & Fascism Discourse; C18: Israel, Palestine & Middle East; C19: War & Militarism).

G.4 Per-Cluster Human-LLM Agreement

To localize where the holistic-interface (H) overflagging in §5.1 concentrates, Table 19 reports per-cluster human–GPT-H agreement across all 22 BERTopic discourse topics on the overlapping human-annotated subset. The five lowest-agreement clusters (C13 Border Control, C14 Islam & Islamization, C12 Refugees & Asylum, C10 Criminal Violence, C3 Welfare State) are precisely the policy-adjacent themes where the co-occurrence shortcut from §5.1 fires: immigration-adjacent vocabulary combined with negative valence triggers an ANTI-IMMIGRANT label under the holistic prompt even when the target is policy or institutions rather than people. The Q1/Q2 decomposition under GPT-D resolves these cases by forcing a separate reference judgment (see Pathway 1 in §4.6).

Border Control & Grenzschutz (45% agreement, 18 disagreements). Comments fall into two groups. The majority (10/18) are rhetorical questions directed at politicians over the 2015 border opening: "Wer hat die Grenzen geöffnet?" ('Who opened the borders?'), "Wer war denn 2015 der Auslöser?" ('Who caused this in 2015?'), "merkel 2015... wir vergessen nicht" ('Merkel 2015... we don't forget'). The remainder (6/18) demand border controls without referencing any group: "wann kommen die Grenzkontrollen?" ('When are the border controls coming?'), "Diese Grenze war vorher auch geschützt" ('This border was protected before too'). In all 18 cases, the negativity targets political actors or policy decisions rather than immigrant

Condition	german_bert			ModernGBERT			gbert-base			xlm-r-base		
	ratio	$\bar{c}$	%>.9	ratio	$\bar{c}$	%>.9	ratio	$\bar{c}$	%>.9	ratio	$\bar{c}$	%>.9
FULL-HUMAN	1.54	.782	37.5	1.11	.749	26.3	1.78	.719	20.0	2.43	.691	10.3
FULL-LLM-26K (GPT-D)	2.73	.896	68.5	2.68	.847	52.2	2.66	.887	63.4	2.74	.874	56.9
FULL-LLM-26K (Qwen-DT)	7.85	.882	62.6	7.05	.936	78.3	7.09	.884	64.5	8.14	.841	40.1

Table 18: Error structure for the three full-pool conditions, mean across 10 seeds. **ratio** = FP:FN ratio; $\bar{c}$ = mean predicted-class confidence on misclassified instances; **%>.9** = share of errors with confidence > 0.90.

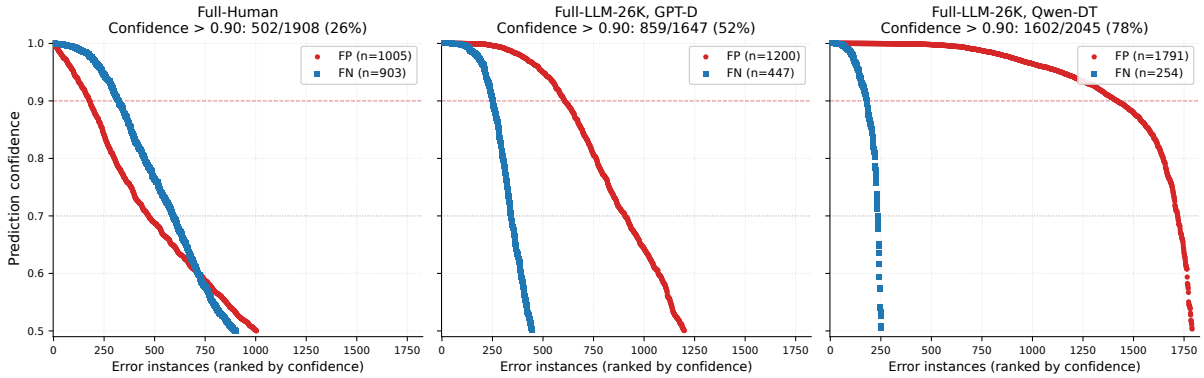


Figure 8: Errors ranked by predicted-class confidence on ModernGBERT for FULL-HUMAN, FULL-LLM-26K (GPT-D), and FULL-LLM-26K (Qwen-DT). The horizontal red line marks confidence $= 0.90$; intersection with each curve gives the share of errors that fall above that threshold (cf. the %>.9 column in Table 18).

ID	Topic	n	H%	GPT-H		GPT-D	
				Agr.	L%	Agr.	L%
C0	Germany in Decline & National Identity	5,489	3	71%	32	84%	21
C1	Pro-AfD Mobilization	2,688	1	93%	7	96%	4
C2	Democracy & Governance Critique	1,915	1	87%	16	91%	10
C3	Welfare State & Economic Competition	1,476	5	59%	52	75%	30
C4	Directed Personal Hostility	1,188	2	74%	25	87%	12
C5	Merkel/CDU & 2015 Legacy	1,072	2	70%	33	78%	28
C6	Migration Policy & Terminology	1,050	8	69%	54	73%	50
C7	Ukraine & Ukrainians	876	10	76%	50	74%	49
C8	Afghanistan, Syria & Turkey	356	8	76%	51	75%	50
C9	Left-Wing Party Critique (Grüne/Linke)	345	3	85%	28	83%	22
C10	Criminal Violence & Public Safety	333	8	70%	57	65%	44
C11	Racism & Fascism Discourse	309	1	93%	11	96%	6
C12	Refugees & Asylum Policy	308	9	75%	65	68%	63
C13	Border Control & Grenzschutz	225	6	45%	69	51%	70
C14	Islam & Islamization	195	9	71%	82	50%	66
C15	EU & European Politics	172	3	73%	22	80%	15
C16	Female Politicians (Faeser/Storch)	162	2	62%	25	71%	19
C17	Poll Numbers & Electoral Math	156	0	87%	12	95%	4
C18	Israel, Palestine & Middle East	126	5	80%	20	72%	20
C19	War & Militarism	123	7	81%	43	86%	37
C20	Rejection Rhetoric	112	1	70%	27	91%	4
C21	Brandmauer & Wall Metaphors	100	2	77%	20	93%	20

Table 19: Per-cluster human–LLM label agreement across all 22 BERTopic discourse topics on the human-annotated subset, under both the holistic (GPT-H) and decomposed (GPT-D) prompts. *Agr.* = label agreement rate; *H%* and *L%* = human and GPT ANTI-IMMIGRANT label rates respectively (H% is shared across prompts; L% is prompt-specific); n is cluster size. Decomposition systematically reduces the GPT ANTI-IMMIGRANT rate toward the human rate (e.g. Welfare 52→30, Directed Hostility 25→12, Germany in Decline 32→21), removing the co-occurrence shortcut described in §5.1. Two clusters resist this pattern: Border Control & Grenzschutz (agreement 45→51%, L% 69→70) and Islam & Islamization (agreement 71→50%, L% 82→66), supporting the topic-intrinsic disagreement pathway discussed in §4.6.

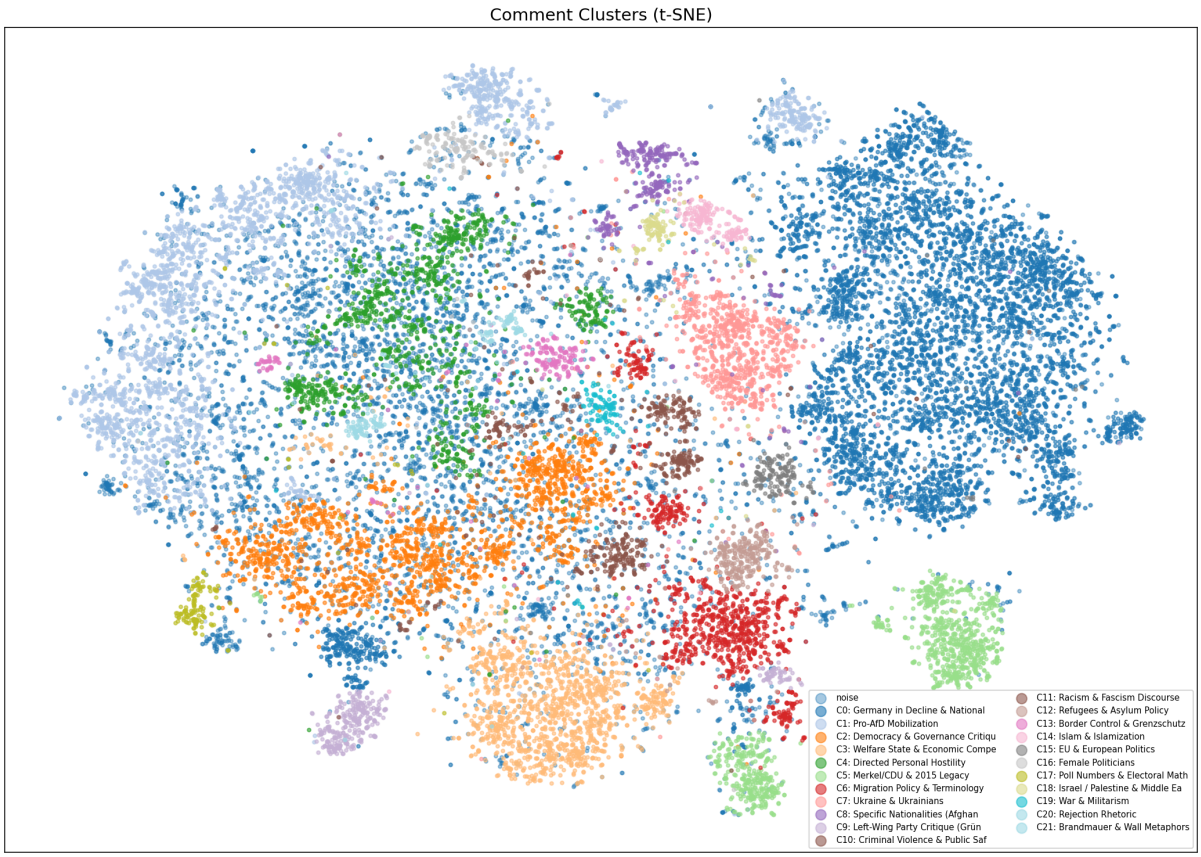


Figure 9: t-SNE projection of all 25,974 immigration-relevant comments, clustered into 22 named discourse topics via BERTopic. Colours denote topic membership; grey points indicate the noise category. The clusters showing the strongest FP-dominant bias under LLM conditions are discussed in §4.6.

groups.

Islam & Islamization (71% agreement, largest gap at +0.72, 7 disagreements). Disagreement comments target Muslims as a religious group: "der Islam gehört nicht zu Deutschland" ('Islam does not belong in Germany'), "muslim free bitte" ('Muslim-free please'), "Die CDU hat Millionen Moslems reingeholt" ('The CDU brought in millions of Muslims'). By contrast, the 17 cases where both sources agree on ANTI-IMMIGRANT typically combine religious references with explicit immigration vocabulary ("muslimische Flüchtlinge", "Islamistenwanderung") or with clearly dehumanizing language.

G.5 Per-Cluster Agreement under the Decomposed Interface

Table 19 reports the equivalent per-cluster agreement under the decomposed prompt (GPT-D). Relative to GPT-H, agreement rises on 20 of 22 clusters (mean 75% to 80%), with the largest gains on the policy-adjacent clusters that drove the holistic interface's over-detection: Welfare & Economic Competition (Agr. 59% to 75%, GPT ANTI-IMMIGRANT rate 52% to 30%), Germany in Decline (71% to 84%, 32% to 21%), Directed Personal Hostility (74% to 87%, 25% to 12%), and Criminal Violence (70% to 65%; rate 57% to 44%). Two clusters resist this pattern: Border Control & Grenzschutz (rate 69% to 70%, agreement 45% to 51%) and Islam & Islamization (rate 82% to 66%, agreement 71% to 50%). These are clusters where surface immigration vocabulary triggers the model regardless of prompt structure, supporting the topic-intrinsic disagreement pathway identified in §4.6.

H Additional Results

This appendix reports the full learning-curve sweep across all four LLM annotation sources and all four encoders (§H.1), complementing Figure 2 in §4.1, and the full normalized ALC table for all acquisition strategies, LLM annotators, and encoder backbones (§H.2).

H.1 Per-Source Learning Curves

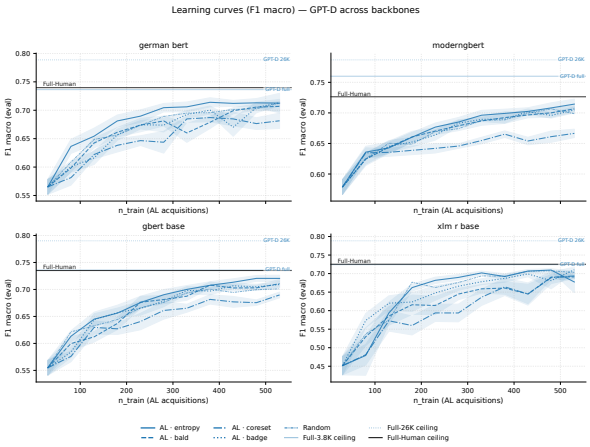
For completeness we report the full learning-curve sweep across all four encoders. Each panel shows the four active-learning acquisition strategies (entropy, BALD, Core-Set, BADGE), the random-sampling baseline, and the source's own full-pool ceilings at 3.8K and 26K labels. The FULL-HUMAN full-pool ceiling is drawn as a solid black reference line.

H.2 Full ALC Results

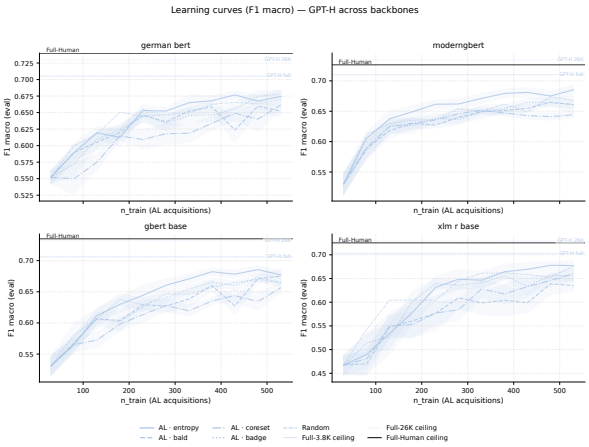
Table 20 reports the normalized ALC for all acquisition strategies, LLM annotators, and encoder backbones discussed in §4.3.

Condition	germ.	Mod.	gbert	xlm-r
Human				
AL-HUMAN Entropy	.668	.668	.659	.638
AL-HUMAN BALD	.658	.649	.652	.616
AL-HUMAN Core-Set	.648	.640	.642	.611
AL-HUMAN BADGE	.663	.645	.655	**.656**
RANDOM-HUMAN	.658	.642	.643	.645
LLM: GPT-D				
AL-LLM, Entropy	**.685**	**.675**	**.676**	.648
AL-LLM, BALD	.664	<u>.670</u>	.664	.622
AL-LLM, Core-Set	.649	.645	.646	.601
AL-LLM, BADGE	.663	.669	<u>.667</u>	.646
RANDOM-LLM	<u>.671</u>	.669	.666	<u>.650</u>
LLM: Qwen-DT				
AL-LLM, Entropy	.633	.660	.637	.611
AL-LLM, BALD	.625	.630	.613	.563
AL-LLM, Core-Set	.619	.620	.611	.560
AL-LLM, BADGE	.629	.644	.632	.602
RANDOM-LLM	.633	.639	.629	.611

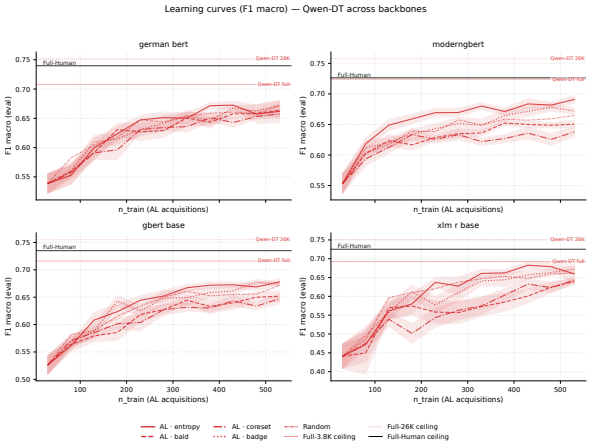
Table 20: Normalized ALC (area under the F1-Macro learning curve, divided by the budget range), mean across 10 seeds. **Bold**/<u>underline</u> = best/second-best per encoder. Results under GPT-H and Qwen-DNT are in App. E.1 and App. E.2.



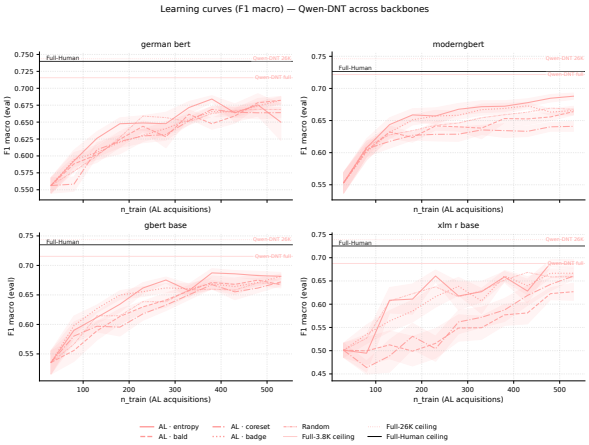
(a) GPT-D



(b) GPT-H

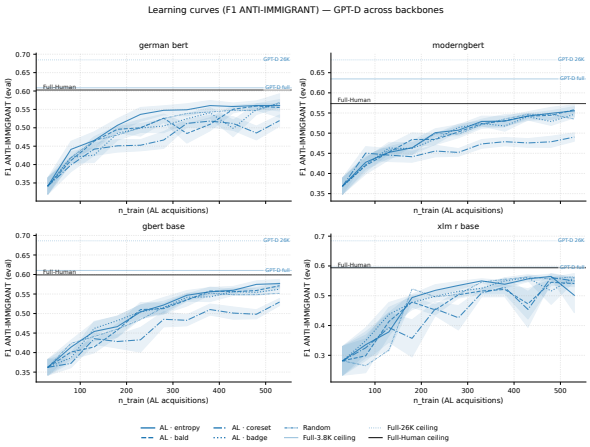


(c) Qwen-DT

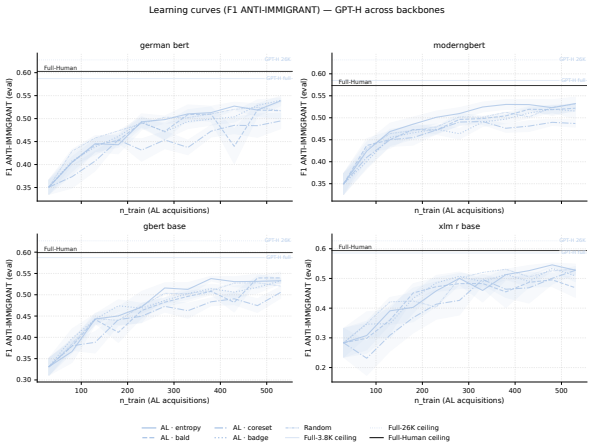


(d) Qwen-DNT

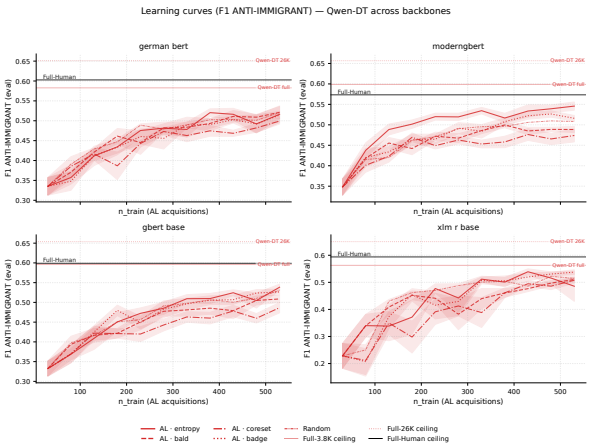
Figure 10: F_1-Macro learning curves per annotation source, across all four encoders. Each panel shows the four active-learning acquisition strategies, the random baseline, the source's own full-pool ceilings at 3.8K and 26K labels, and the FULL-HUMAN full-pool ceiling (solid black) as reference.



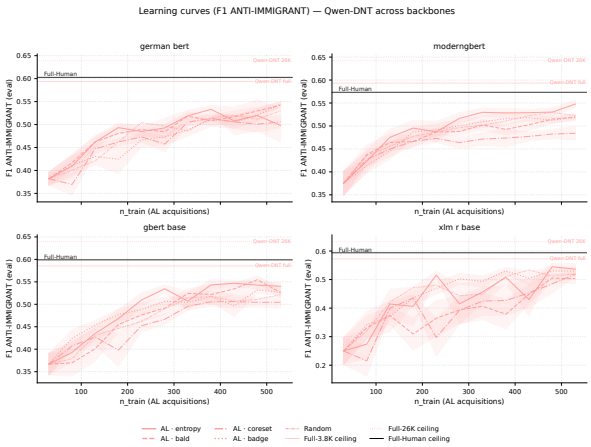
(a) GPT-D



(b) GPT-H



(c) Qwen-DT



(d) Qwen-DNT

Figure 11: F-$_1$ ANTI-IMMIGRANT learning curves per annotation source, across all four encoders. Each panel shows the four active-learning acquisition strategies, the random baseline, the source's own full-pool ceilings at 3.8K and 26K labels, and the FULL-HUMAN full-pool ceiling (solid black) as reference.