# 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 from a short prompt at negligible 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 labelled at once? We investigate both questions on a new dataset of 277,902 German political TikTok comments (25,974 LLM-labelled, 5,000 human-annotated), comparing seven annotation strategies across four encoders to detect anti-immigrant hostility. A classifier trained on 25,974 GPT-5.2 labels ($43) achieves comparable F1-Macro to one trained on 3,800 human annotations ($316). Active learning offers little advantage over random sampling in our pre-enriched pool and delivers lower F1 than full LLM annotation at the same cost. However, comparable aggregate F1 masks a systematic difference in error structure: LLM-trained classifiers over-predict the positive class relative to the human gold standard. This divergence concentrates in topically ambiguous discussions where the distinction between anti-immigrant hostility and policy critique is most subtle, suggesting that annotation strategy should be guided not by aggregate F1 alone but by the error profile acceptable for the target application.


## 1 Introduction

Annotated data remains a bottleneck for supervised learning in NLP. A recent community survey finds that 94% of practitioners still consider annotation a limiting factor, and that active learning (AL) remains widely used (Romberg et al., 2025). Recent work has begun placing 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 examine how annotation source affects classifier behaviour. When an instruction-tuned LLM can label an entire corpus from a short prompt at a fraction of human annotation cost, the comparison that matters is no longer AL versus human annotation but AL versus LLM annotation at scale. We test whether AL's cost premium survives that comparison.

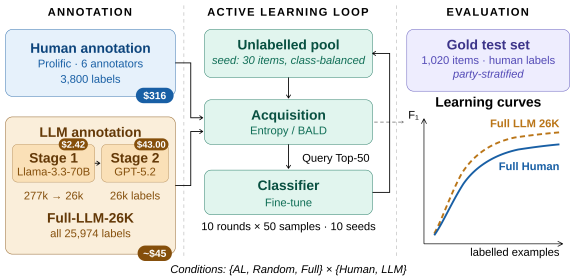


Figure 1: **Experimental framework.** Human and LLM (Llama-3.3-70B prefiltering + GPT-5.2 classification) annotations feed into the same active learning loop, evaluated across seven conditions, four German encoder models, and 10 seeds on a party-stratified test set.

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 is likely to matter. It also represents a common scenario in which a lightweight supervised model must be trained from limited labels for efficient inference at scale.

Our experimental design (Figure 1) compares seven labelling conditions that vary annotation source (human vs. LLM), sampling strategy (active learning vs. random vs. full pool), and label volume (530 to 25,974 instances) across four encoder models and 10 random seeds. The labelling pool comprises 25,974 immigration-relevant comments constructed via a two-stage LLM pipeline, of which 5,000 are independently annotated by six human crowdworkers, yielding 1,200 gold evaluation items and 3,800 training labels. We organise the paper around three research questions:

**RQ1:** Does human-in-the-loop active learning remain cost-competitive with LLM annotation at scale, and does it provide any performance advantage that justifies its cost premium?

**RQ2:** What characteristic errors and confidence patterns emerge in classifiers trained on LLM- versus human-labelled data?

**RQ3:** Where in the discourse space do human and LLM annotations diverge, and what does this reveal about how each source operationalises the annotation boundary?

Our central finding is that LLM annotation at scale achieves comparable F1-Macro to human annotation at one-seventh the cost, but with systematic differences in error structure. We release a new dataset of 277,902 German political TikTok comments (25,974 LLM-labelled and 5,000 human-annotated) with experimental code.

## 2 Related Work

### 2.1 Active Learning and LLMs as Annotators

Active learning (AL) reduces annotation cost by selecting only the most informative instances for human labelling (Settles, 2009), but practical adoption remains limited by unreliable performance gains, hyperparameter sensitivity, poor tooling, and difficulty obtaining reliable uncertainty estimates for acquisition functions (Lowell et al., 2019; Romberg et al., 2025; Schröder and Niekler, 2020).

Instruction-tuned LLMs offer a complementary cost reduction: zero-shot ChatGPT achieves higher annotation agreement than crowd-workers 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, 2024), including German migration discourse (Kostikova et al., 2024). Recent work places LLMs inside the AL loop as annotators (Zhang et al., 2023; He et al., 2024; Xia et al., 2025), though Pangakis et al. (2023) caution that quality is task- and prompt-sensitive and requires validation against human gold labels.

Yet most studies evaluate annotation agreement rather than downstream classifier performance or error profiles (Heseltine and Clemm von Hohenberg, 2024; Fillies et al., 2025); none compare human versus LLM labels across matched AL, random, and full-pool conditions with error-structure analysis. This is the gap we address.

### 2.2 Hate Speech Detection and Annotation

Hate speech detection is well studied (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), but annotation difficulty increases sharply 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. This difficulty affects LLMs as well as humans: Vallejo Vera and Driggers (2025) show that LLMs are sensitive to party cues when annotating political text. German-language resources remain limited: Ross et al. (2017) demonstrate low inter-annotator agreement even with detailed guidelines, and the most related dataset (Fillies et al., 2025) covers far-right ecosystems rather than mainstream party accounts. Annotator disagreement in subjective tasks is increasingly understood as signal rather than noise (Plank, 2022; Röttger et al., 2022; Davani et al., 2022; Wright and Augenstein, 2025), raising the question of whether LLMs can capture such 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, and Die Linke) via the TikTok Research API, covering January 2024 to September 2025, yielding 467,660 comments. 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 such as mentions, URLs, hashtags, and numbers); (3) exact text deduplication within and across party accounts; and (4) language identification via GlotLID (Kargaran et al., 2023) to retain only German-language comments. This yielded a filtered corpus of **277,902** German-language comments (median 14 words per comment); full per-step counts are provided in Table 13 (App. D.1). This corpus forms the input to the LLM-based prefiltering stage described in Section 3.3.

### 3.2 Label Space Definition

We define a binary classification task. A comment is labelled ANTI-IMMIGRANT if it expresses hostility, fear, or opposition toward immigrants, refugees, or asylum seekers as a group, including xenophobic

language, stereotypes, deportation rhetoric, or attribution of social problems to these groups. All other comments are labelled NOT ANTI-IMMIGRANT.

**Two-Question Annotation Framework.** For human annotation, we operationalise this label through two sequential yes/no questions rather than a direct label assignment:

- **Q1:** Does this comment refer to immigrants, refugees, asylum seekers, foreigners, or immigration policy?
- **Q2:** Are these groups portrayed negatively through hostility, stereotypes, deportation rhetoric, or attribution of social or economic problems?

The final label is derived programmatically: ANTI-IMMIGRANT $\Leftrightarrow$ Q1 = YES $\wedge$ Q2 = YES; otherwise NOT ANTI-IMMIGRANT. This decomposition embeds the core decision logic directly into the annotation interface and enables diagnostic tracing of disagreements to the reference step (Q1) or the valence step (Q2).

### 3.3 LLM-Based Pool Construction and Annotation

We construct the labelling pool via a two-stage LLM pipeline.

**Stage 1: Topic Prefiltering.** We applied `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024) to all 277,902 comments using a binary filtering prompt asking whether each comment discusses immigration, migrants, refugees, asylum seekers, or related policies (full prompt and configuration: App. A.1). This reduced the corpus to **25,974 comments** (9.3% of the filtered corpus).

**Stage 2: Annotation.** We applied `GPT-5.2` via the OpenAI Batch API to annotate each prefiltered comment as ANTI-IMMIGRANT or NOT ANTI-IMMIGRANT using the label definitions from Section 3.2 (full prompt and configuration: App. A.2. The model was instructed to default to NOT ANTI-IMMIGRANT when uncertain. GPT-5.2 labelled 7,847 comments (30.2%) as ANTI-IMMIGRANT and 18,127 (69.8%) as NOT ANTI-IMMIGRANT. Table 1 shows the distribution across party accounts.

**Annotation Corpus.** From the labelled pool, we sampled 5,000 comments for the human annotation study (Section 3.4) using party-stratified sampling at a target ratio of 40% ANTI-IMMIGRANT to 60% NOT ANTI-IMMIGRANT, a moderate enrichment over the natural 30.2% base rate to ensure sufficient positive instances for active learning. This same set is independently annotated by crowdworkers, enabling direct comparison between LLM and human labels under matched conditions. The LLM receives a single prompt rather than the two-question decomposition, targeting the same label space.

| Party | Total | Anti | % |
|---|---|---|---|
| AfD | 9,379 | 2,691 | 28.7 |
| CDU/CSU | 7,316 | 2,508 | 34.3 |
| SPD | 4,324 | 1,430 | 33.1 |
| Die Linke | 3,311 | 831 | 25.1 |
| Grüne | 1,244 | 246 | 19.8 |
| FDP | 400 | 141 | 35.2 |
| **Total** | **25,974** | **7,847** | **30.2** |

Table 1: GPT-5.2 label distribution after annotation of the Llama-prefiltered pool, by party account.

| 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 2: Annotation study structure. Multiply-annotated items form the gold evaluation set; single-annotated items form the active learning training pool.

### 3.4 Human Annotation Study

We recruited annotators through Prolific, restricting participation to native or highly proficient German speakers. Eight were recruited; six were retained beyond the qualification round. Annotations were collected via Argilla,[1] hosted on Hugging Face Spaces, using the two-question interface described in Section 3.2. The study comprised three phases with decreasing annotation overlap (Table 2), yielding 1,200 multiply-annotated items for the gold evaluation set and 3,800 single-annotated items for the active learning training pool. Details on qualification criteria, attention checks, and item assignment are provided in App. C.

For the derived binary label, inter-annotator agreement is moderate: Krippendorff's $\alpha = 0.49$ on the pilot (6-way, 200 items) and $\alpha = 0.43$ on the triple-annotated main items (3-way, 950 items). Agreement is higher on Q1 (immigra-

[1] https://argilla.io/

| Condition | Source | Sampling | $n_{\text{train}}$ |
|---|---|---|---|
| AL-HUMAN | Human | Entropy / BALD | $\leq$530 |
| AL-LLM | LLM | Entropy / BALD | $\leq$530 |
| RANDOM-HUMAN | Human | Random | $\leq$530 |
| RANDOM-LLM | LLM | Random | $\leq$530 |
| FULL-HUMAN | Human | Full pool | 3,800 |
| FULL-LLM | LLM | Full pool | 3,800 |
| FULL-LLM-26K | LLM | Full pool | 25,974 |

Table 3: Seven experimental conditions. AL conditions use two acquisition functions (Section 3.6) and report maximum $n_{\text{train}}$ at round 10 (30 seed + 500 acquired).

tion reference, $\alpha = 0.57$–$0.59$) than on Q2 (negative portrayal, $\alpha = 0.43$–$0.49$), confirming that the primary source of disagreement is judging valence rather than identifying the topic, consistent with prior work on German hate speech annotation (Ross et al., 2017). Labels are aggregated via majority vote (full pairwise agreement in App. C).

### 3.5 Experimental Conditions

We compare seven conditions spanning three design axes: annotation source (human vs. LLM), sampling strategy (AL vs. random vs. full pool), and label volume (Table 3). Within-pair comparisons isolate label source; AL-vs-Random comparisons isolate acquisition value; and FULL-LLM-26K tests whether scale compensates for label quality. All conditions share the same classifier, training recipe, and evaluation set (Section 3.7); costs are reported in Section 3.8.

### 3.6 Active Learning Design

We implement pool-based AL for AL-HUMAN and AL-LLM. Both conditions start from identical initial states: the same unlabelled pool and a class-balanced, party-stratified seed set of 30 comments (5 per party), following Fairstein et al. (2024). Since labels are collected prior to experiments, we 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 a maximum of 530 labelled examples at round 10 (30 seed + 500 acquired). We evaluate two acquisition strategies: **Entropy**, selecting instances with the highest predictive entropy over the softmax output; and **BALD** (Houlsby et al., 2011), selecting instances maximising mutual information between predictions and model parameters, estimated via MC Dropout (Gal and Ghahramani, 2016) with 10 forward passes. After each round, the classifier is retrained from scratch and evaluated on the held-out test set; learning curves report mean and standard deviation across 10 random seeds.

### 3.7 Model, Training, and Evaluation

**Models.** We evaluate four pre-trained transformer encoders: **german-bert**,[2] **ModernGBERT**,[3] **gbert-base**,[4] and **xlm-r-base**.[5] The latter serves as a multilingual baseline. Each model is fine-tuned with a randomly initialised binary classification head. Comments are tokenised and truncated to 128 tokens.

**Training.** We use AdamW (Loshchilov and Hutter, 2017) with learning rate $2 \times 10^{-5}$, batch size 16, and weight decay $10^{-2}$, with linear warmup (10%) and decay. To address class imbalance we apply focal loss (Lin et al., 2017) with $\gamma = 2.0$ and per-class weights recomputed from class frequencies at each AL round. For full conditions and the AL seed round, training runs for up to 10 epochs with early stopping patience of 5 on development set macro-F1. For AL rounds 1+, patience is reduced to 3 to limit overfitting on small labelled sets.

**Evaluation.** The 1,200-item gold evaluation set is partitioned into a development set (15%, $n \approx 180$, for early stopping) and a test set (85%, $n \approx 1{,}020$, for final reporting), both stratified by class. We report macro-F1 as the primary metric, supplemented by class-specific F1 for ANTI-IMMIGRANT (F1-Anti). Across AL conditions we additionally report the Area under the Learning Curve (ALC), normalised by the number of AL rounds. All experiments are repeated with 10 random seeds;[6] we report mean and standard deviation across seeds.

### 3.8 Cost-Effectiveness Analysis

We compare annotation costs across conditions (Table 4). All costs are reported in USD (£1 = $1.334). Human annotators were paid at Prolific's "good rate" of £9/hour across three phases with 6 annotators, totalling $416 ($0.083 per label across 5,000 annotations).[7] The full LLM pipeline (Llama pre-filtering + GPT-5.2 annotation) cost $45.72 ($0.002 per label). Full token estimates and infrastructure details are provided in App. B. Cost-performance tradeoffs are analysed in Section 4.3.

[2] https://hf.co/dbmdz/bert-base-german-uncased
[3] https://hf.co/LSX-UniWue/ModernGBERT_134M
[4] https://hf.co/deepset/gbert-base
[5] https://hf.co/FacebookAI/xlm-roberta-base
[6] Seeds: 42, 123, 1997, 2026, 1337, 7, 183, 56, 919, 777.
[7] Phase breakdown: £18 (test, 20 min) + £54 (pilot, 60 min) + £162 (main, 3 hours) = £234, plus £77.92 platform fee.

| Condition | $n_{\text{train}}$ | Cost | \$/label |
|---|---|---|---|
| FULL-HUMAN | 3,800 | \$316 | \$0.083 |
| FULL-LLM | 3,800 | \$6.34 | \$0.002 |
| FULL-LLM-26K | 25,974 | \$43.30 | \$0.002 |
| AL-HUMAN | 530 | \$44 | \$0.083 |
| AL-LLM | 530 | \$0.88 | \$0.002 |
| Full human study | 5,000 | \$416 | \$0.083 |

Table 4: Annotation costs per condition (£1 = \$1.334). Human costs are computed proportionally from the \$416 total; LLM annotation costs from the \$43.30 GPT-5.2 total. All LLM conditions additionally depend on a one-time Llama prefiltering cost of \$2.42 (not included per condition). RANDOM conditions share the cost structure of their AL counterparts and are omitted.

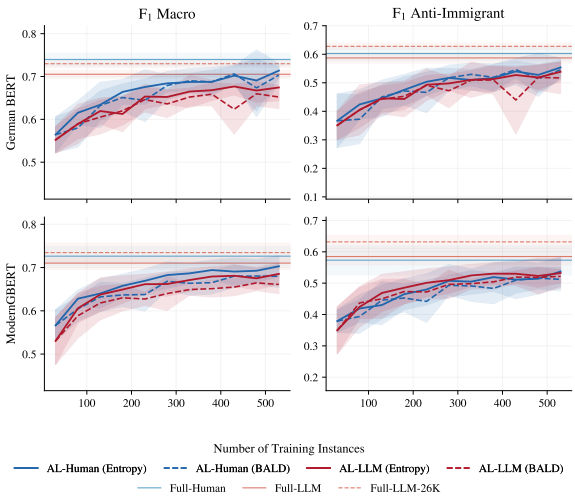


Figure 2: Learning curves for german_bert (top) and ModernGBERT (bottom) on F1-Macro (left) and F1-Anti (right). Shaded regions indicate ±1 SD across 10 seeds. Horizontal lines mark full-pool baselines. Results for gbert-base and xlm-r-base are in App. G.1.

## 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 stable across all four architectures. German-specific models consistently outperform the multilingual baseline (xlm-r-base) by up to 0.035 F1-Macro points, a gap comparable in magnitude to the annotation strategy effects reported below. We focus on german_bert and ModernGBERT in the main text: german_bert achieves the highest absolute performance, while ModernGBERT shows a different pattern on the scale–quality tradeoff where FULL-LLM-26K exceeds FULL-HUMAN. Results for all four models are provided in App. G.1.

### 4.1 Learning Curves and Annotation Source (RQ1)

**Key finding.** Human labels yield a small F1-Macro advantage at matched volume. At scale (25,974 vs. 3,800 labels), FULL-LLM-26K achieves comparable F1-Macro and higher F1-Anti than FULL-HUMAN across models.

AL-HUMAN (Entropy) outperforms all AL conditions on both metrics. AL-LLM (Entropy) tracks within 0.018–0.040 F1-Macro points at round 10. The gap is most pronounced in early rounds ($n_{\text{train}} < 200$) and narrows as data accumulates, indicating that volume only partially compensates for label quality within the AL budget. FULL-HUMAN (0.725–0.740 F1-Macro across models) sets the performance ceiling at matched volume. FULL-LLM falls 0.016–0.035 points below, isolating the label quality effect. FULL-LLM-26K recovers this gap with approximately 7× more training examples (0.730–0.735 vs. 0.725–0.740). On F1-Anti, FULL-LLM-26K achieves the highest scores across all four models (0.627–0.632), exceeding FULL-HUMAN (0.573–0.603) by 0.025–0.059 points, reflecting higher recall at the cost of more false positives (Section 4.4).

### 4.2 Acquisition Functions and Sampling Strategy

**Key finding.** Entropy-based acquisition outperforms BALD, but neither provides advantage over random sampling in a pre-enriched pool.

AL-HUMAN (Entropy) > AL-HUMAN (BALD) and AL-LLM (Entropy) > AL-LLM (BALD) across all four models (Table 6). The gap is larger within LLM-labelled conditions (up to 0.035 on xlm-r-base), suggesting that BALD's epistemic uncertainty estimates are more sensitive to systematic divergence between training labels and the evaluation standard. Entropy, operating on mean softmax outputs, is more robust to this effect.

However, the AL–Random gap is more consequential than the Entropy–BALD gap. The Δ rows in Table 6 show that AL (Entropy) provides near-zero advantage over random sampling for german_bert, gbert-base, and ModernGBERT. For xlm-r-base, AL performs worse than random in both regimes ($\Delta = -0.007$ and $-0.009$), likely reflecting poor uncertainty calibration. This pattern is consistent with our pre-enriched pool construction (Section 3.3), which reduces the marginal benefit of uncertainty-based selection.

| Condition | germ. | Mod. | gbert | xlm-r |
|---|---|---|---|---|
| | *F1-Macro* | | | |
| FULL-HUMAN | **0.740** | 0.726 | **0.735** | 0.725 |
| FULL-LLM | 0.705 | 0.710 | 0.706 | 0.703 |
| FULL-LLM-26K | 0.730 | **0.735** | 0.732 | **0.730** |
| AL-HUMAN (Ent.) | 0.714 | 0.703 | 0.706 | 0.704 |
| AL-LLM (Ent.) | 0.674 | 0.685 | 0.677 | 0.677 |
| RAND.-HUMAN | 0.695 | 0.677 | 0.690 | 0.702 |
| RAND.-LLM | 0.670 | 0.668 | 0.664 | 0.677 |
| | *F1-Anti* | | | |
| FULL-HUMAN | 0.603 | 0.573 | 0.599 | 0.594 |
| FULL-LLM | 0.587 | 0.585 | 0.587 | 0.586 |
| FULL-LLM-26K | **0.628** | **0.632** | **0.627** | **0.627** |
| AL-HUMAN (Ent.) | 0.554 | 0.537 | 0.545 | 0.562 |
| AL-LLM (Ent.) | 0.539 | 0.532 | 0.533 | 0.528 |
| RAND.-HUMAN | 0.527 | 0.492 | 0.529 | 0.548 |
| RAND.-LLM | 0.517 | 0.531 | 0.520 | 0.531 |

Table 5: F1-Macro and F1-Anti (mean across 10 seeds) for all seven conditions. AL and Random conditions report round 10 values. **Bold** and underline indicate best and second-best per model.

| Condition | Acq. | germ. | Mod. | gbert | xlm-r |
|---|---|---|---|---|---|
| AL-HUM. | Entropy | .668 | .668 | .659 | .638 |
| AL-HUM. | BALD | .658 | .649 | .652 | .616 |
| RAND.-HUM. | — | .658 | .642 | .643 | .645 |
| Δ | | +.010 | +.026 | +.015 | −.007 |
| AL-LLM | Entropy | .642 | .653 | .643 | .611 |
| AL-LLM | BALD | .629 | .632 | .623 | .576 |
| RAND.-LLM | — | .641 | .640 | .628 | .621 |
| Δ | | +.001 | +.013 | +.015 | −.009 |

Table 6: Normalised ALC (mean, 10 seeds). Δ = AL (Entropy) minus Random; positive values favour AL.

### 4.3 Cost-Effectiveness and the Scale–Quality Tradeoff (RQ1)

**Key finding.** FULL-LLM-26K matches FULL-HUMAN on F1-Macro at one-seventh the cost, but with a different error profile (Section 4.4).

Figure 3 visualises the cost–performance frontier. At matched volume (3,800 comments each, \$6 vs. \$316), FULL-LLM falls 0.016–0.035 F1-Macro points below FULL-HUMAN, isolating the pure label quality effect. FULL-LLM-26K recovers this gap by providing 7× more training examples at \$43 (0.730–0.735 vs. 0.725–0.740 across models). Label quality can be compensated with volume when LLM annotation is cheap enough to scale. The cost comparison between AL and LLM annotation is striking. Active learning acquires a maximum of 530 human labels at a proportional cost of $530/5{,}000 \times \$416 = \$44$, virtually identical to FULL-LLM-26K (\$43). At equal absolute cost, FULL-LLM-26K provides 49× more training labels and higher final F1-Macro (Table 5). Active learning reduces the cost of human annotation relative to a full human study; it does not reduce annotation cost below LLM annotation at scale.

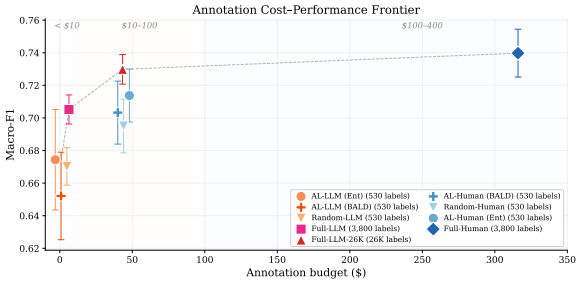


Figure 3: Annotation cost–performance frontier for german_bert (mean ± 1 SD across 10 seeds). Budget zones mark three regimes: under \$10 (LLM-only), \$10–100 (LLM at scale), and \$100–400 (human annotation).

### 4.4 Error Profiles and Confidence Analysis (RQ2)

**Key finding.** Comparable aggregate F1 masks a systematic difference in error structure: LLM-trained classifiers over-predict the positive class relative to the human gold standard with higher confidence than human-trained models (Fig. 4).

**FP/FN asymmetry.** Figure 4 (left) shows that FULL-HUMAN produces approximately balanced FP and FN counts across seeds for german_bert (124 FPs, 87 FNs). FULL-LLM-26K produces 243 FPs but only 21 FNs, a nearly 12:1 ratio. Whether this reflects genuine misclassification or a broader but defensible interpretation of anti-immigrant content is examined in Section 4.5.

This asymmetry traces to how each annotation source operationalises the policy-versus-people distinction. The human two-question framework explicitly separates immigration reference (Q1) from negative group portrayal (Q2), requiring annotators to judge each step independently. GPT-5.2's single holistic prompt evaluates both dimensions simultaneously, and appears to draw a broader boundary for what constitutes group-directed negativity in the presence of policy-adjacent framing. Training on 25,974 examples with this consistently broader boundary embeds the LLM's decision rule into the classifier, producing high recall but low precision for the ANTI-IMMIGRANT class when evaluated against human labels. FULL-LLM at matched volume shows an intermediate FP skew, confirming the effect scales with LLM-labelled data volume.

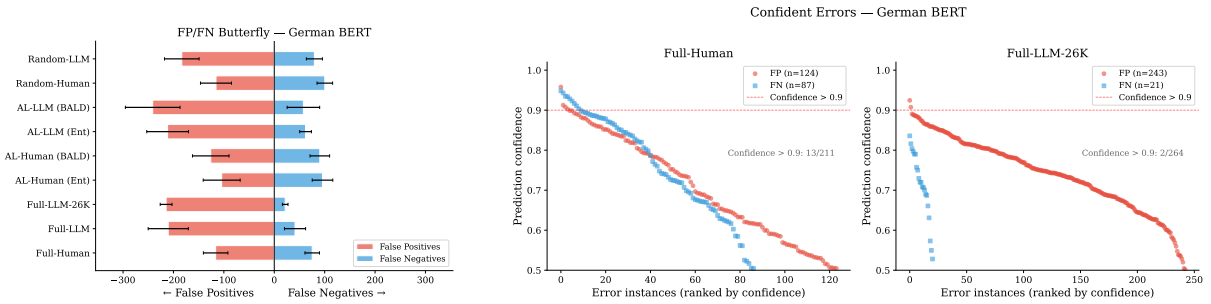


Figure 4: *Left*: FP/FN counts per condition for german_bert (mean ± 1 SD across 10 seeds; dots = individual seeds). *Right*: Errors ranked by prediction confidence for FULL-HUMAN and FULL-LLM-26K.

**Confidence on errors.** Figure 4 (right) shows that FULL-LLM-26K's FP confidence curve is shifted higher relative to FULL-HUMAN. FULL-HUMAN-trained models show lower confidence on incorrect predictions, making them more amenable to threshold tuning and confidence-based quality assurance. In deployment scenarios where prediction confidence is used for human review triage, LLM-trained models would route fewer errors to review because the errors appear confident, a potentially costly failure mode for content moderation.

**Annotation boundaries.** Cross-seed analysis reveals instances that are systematically misclassified across all seeds and conditions. Many 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 🤡')

The first clause paraphrases a well-known quote by Konrad Adenauer ("Nehmen Sie die Menschen, wie sie sind, andere gibt's nicht" — 'Take people as they are, there are no others'), juxtaposed with *Leitkultur*, a politically charged concept implying cultural assimilation, and a clown emoji signalling sarcasm. The comment is genuinely ambiguous: if the emoji targets *Leitkultur*, the author is mocking the demand for cultural assimilation (NOT ANTI-IMMIGRANT); if it targets the acceptance statement, the author is mocking tolerance toward immigrants as naive (ANTI-IMMIGRANT). An LLM with knowledge of German political rhetoric may recognise the Adenauer reference and interpret the comment as a political citation being ironically reframed, while crowdworkers without this background are more likely to rely on surface-level cues. This illustrates how annotation source can affect not only the rate but also the *type* of labelling decisions on culturally embedded content.

### 4.5 Topical Clustering and Annotation Disagreement (RQ3)

**Key finding.** Human–LLM annotation disagreement concentrates in themes where the distinction between anti-*immigrant* hostility and anti-*immigration* policy critique is most ambiguous.

We embed all 25,974 prefiltered comments using `paraphrase-multilingual-MiniLM-L12-v2` and apply BERTopic (Grootendorst, 2022), yielding 22 named topics after merging four semantically overlapping pairs (clustering details and t-SNE visualization in App. E).

Agreement ranges from 45% (Border Control & Grenzschutz) to 93% (Pro-AfD Mobilization; Table 7), with the lowest agreement in topics where policy-adjacent negativity is most prevalent. Islam & Islamization shows moderate agreement (71%) but the largest label-rate gap: GPT-5.2 labels 82% as ANTI-IMMIGRANT versus 9% by humans. The LLM ANTI-IMMIGRANT rate exceeds the human rate for all 22 topics without exception, confirming systematic directional divergence. This propagates into classifier error rates: FP increases under LLM annotation are largest in Border Control & Grenzschutz, Criminal Violence & Public Safety, and Welfare State & Economic Competition, topics where negative framing targets policies or institutions rather than immigrant groups directly (per-topic error rates in App. G.3).

**Two mechanisms of disagreement.** Manual inspection of all disagreement cases in the two most informative clusters reveals two distinct mechanisms (full inventory in App. F).

In **Border Control & Grenzschutz** (45% agreement, 18 disagreements), all 18 cases target political actors or policy decisions rather than immi-

| Topic | n | Agr. | H% | L% |
|---|---|---|---|---|
| Border Control & Grenzschutz | 225 | 45% | 6 | 69 |
| Welfare & Econ. Competition | 1,476 | 59% | 5 | 52 |
| Migration Policy & Terminology | 1,050 | 69% | 8 | 54 |
| Islam & Islamization | 195 | 71% | 9 | 82 |
| Pro-AfD Mobilization | 2,688 | 93% | 1 | 7 |
| Racism & Fascism Discourse | 309 | 93% | 1 | 11 |

Table 7: Selected discourse topics with lowest and highest human–LLM agreement. Agr. = label agreement rate; H% and L% = human and LLM ANTI-IMMIGRANT rates. Islam & Islamization has moderate agreement but the largest label-rate gap (+0.72). Full table in App. 14.

grant groups e.g. "Wer hat die Grenzen geöffnet?" ('Who opened the borders?') and "wann kommen die Grenzkontrollen?" ('When are the border controls coming?'). Human annotators classify these as NOT ANTI-IMMIGRANT because Q2 (negative portrayal of immigrants) is not satisfied; GPT-5.2 treats immigration-adjacent vocabulary combined with negative framing as sufficient evidence for ANTI-IMMIGRANT, conflating policy critique with group-directed hostility. In **Islam & Islamization** (71% agreement, largest gap at +0.72, 7 disagreements), a different pattern emerges. Disagreement comments target Muslims as a religious group e.g. "der Islam gehört nicht zu Deutschland" ('Islam does not belong in Germany') without explicit immigration framing. The distinguishing factor is the Q1 gateway: comments framing hostility in purely religious terms may receive Q1 = NO, producing an automatic NOT ANTI-IMMIGRANT regardless of the comment's hostility. GPT-5.2 treats anti-Muslim and anti-immigrant hostility as overlapping categories. Whether this broader scope is a labelling error or a defensible interpretation depends on whether the framework intends to capture hostility toward immigration-associated groups or only toward groups explicitly framed as immigrants.

## 5 Discussion

**The case for LLM annotation at scale.** For research groups operating under tight budgets, common in computational social science and low-resource NLP, large-scale LLM annotation can substitute for human annotation on aggregate performance metrics without requiring crowdsourcing infrastructure or extended recruitment timelines. LLM annotation also enables broader coverage of the target distribution, which may contribute to the strong F1-Anti performance of FULL-LLM-26K (Table 5). Beyond cost, our human annotation pipeline required several weeks across setup, recruitment, qualification, pilot, and main phases; GPT-5.2 annotated all 25,974 comments within hours. This speed difference enables rapid iteration over label definitions and prompt designs before committing to a full annotation run.

**When human annotation remains essential.** LLM-trained classifiers systematically over-predict the positive class with higher confidence (Section 4.4). We do not claim human labels are uniformly more accurate. The Adenauer example and the Islam & Islamization cluster suggest cases where the LLM may capture hostility that the two-question framework excludes. However, where false positives carry real costs, the FP-skewed error profile may be unacceptable regardless of aggregate F1. Which source better captures the intended category depends on the research question and the definition of harm being operationalised. LLM annotation does carry an additional ethical advantage: it eliminates human exposure to harmful content, a non-trivial consideration given documented psychological costs (Steiger et al., 2021).

**Active learning in pre-enriched pools.** The near-zero AL advantage over random sampling reflects a successful pool construction strategy rather than a failure of AL. Practitioners should treat pool construction as a first-class design decision: investment here yields more reliable returns than investment in acquisition function sophistication.

## 6 Conclusion

We compared human and LLM annotation across seven conditions, four encoders, and 10 seeds for detecting anti-immigrant hostility in German political TikTok comments. LLM annotation at scale achieves comparable F1-Macro to human annotation at approximately one-seventh the cost, while AL with human labels delivers lower F1 at essentially the same price and provides little advantage over random sampling in our pre-enriched pool. However, comparable aggregate performance masks systematic differences in error structure that concentrate in topically ambiguous discourse, where the distinction between group-directed hostility and policy critique is most subtle. The choice between human and LLM annotation should be guided not by aggregate F1 alone but by the error profile acceptable for the target application.

## 7 Limitations

**Pre-enriched annotation pool.** Our annotation pool was constructed via LLM prefiltering, which concentrates the pool around immigration-relevant content and reduces the marginal benefit of uncertainty-based acquisition. Our finding that AL provides little advantage over random sampling (Section 4.2) should be interpreted in this context; results may differ with an unfiltered corpus.

**Crowdsourced rather than expert annotation.** Human annotations were collected from Prolific crowdworkers, not domain experts in hate speech research. Crowdworker judgements may differ systematically from those of trained experts, particularly for borderline cases involving implicit hostility or culturally specific references. Expert annotations might shift the decision boundary and alter the relative performance of human versus LLM conditions.

**Evaluation against human gold labels.** All conditions are evaluated against the human gold standard, which systematically disadvantages LLM-trained models whose annotation boundary diverges from the human one. Instances where GPT-5.2 and the downstream classifier agree on ANTI-IMMIGRANT but the human majority votes NOT ANTI-IMMIGRANT are counted as false positives regardless of which label is more appropriate. An expert adjudication study on contested instances would help disentangle genuine model errors from evaluation bias.

**Single task and domain.** All experiments are conducted on a single binary classification task in a single language (German) on a single platform (TikTok). Findings may not generalise to other subjective annotation tasks, languages, or platforms.

**Single LLM annotator and prompt design.** LLM labels are produced by a single model (GPT-5.2) with a single classification prompt. We tested multiple prompt formulations during development, varying label definitions and instruction specificity, before selecting the current design based on agreement with a pilot subset of human labels. Nevertheless, the systematic annotation boundary divergence we observe (Section 4.5) may be specific to this model and prompt design rather than a general property of LLM annotation. Moreover, our human annotators use a two-question decomposition whereas the LLM receives a single holistic prompt. This asymmetry reflects a practical constraint: pilot testing showed that annotators who received a holistic label definition did not consistently follow the guidelines. Part of the observed FP asymmetry may therefore stem from this difference in annotation interface rather than from a fundamental difference between human and LLM judgment.

**Simulated active learning.** We simulate the AL loop by pre-collecting both human and LLM labels and revealing the appropriate label when an instance is queried, rather than collecting annotations on-demand. In a true online setting, annotators may calibrate their understanding over rounds, which could affect the human–LLM gap in ways our simulation does not capture.

## 8 Ethical Considerations

**Annotator exposure and compensation.** Our annotation task required crowdworkers to read and classify comments containing xenophobic language, stereotypes, and dehumanising 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 Section 3.8 and App. B). 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 under 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 it appeared under. Usernames and user IDs are stripped from the released dataset. The annotation study was conducted with informed consent from all annotators.

**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. *Journal of Computational Social Science*, 8(17).

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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(4).

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

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv: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 (Volume 6: Industry Track)*. Association for Computational Linguistics.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1).

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

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*, pages 6155–6218.

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 *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 397–412. Springer.

Aida Kostikova, Benjamin Paassen, Dominik Beese, 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*, pages 5884–5907, Miami, Florida, USA. Association for Computational Linguistics.

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

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 (ICCV)*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of EMNLP-IJCNLP*.

Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2025. Can large language models capture human annotator disagreements? *SSRN preprint 5319936*.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, disagreement, and model evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Julia Romberg, Christopher Schröder, Julius Gonsior, Katrin Tomanek, and Fredrik Olsson. 2025. Have LLMs made active learning obsolete? Surveying the NLP community. *arXiv preprint arXiv:2503.09701*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kuber, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC)*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet 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)*, pages 175–190. Association for Computational Linguistics.

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*, pages 5477–5490.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. pages 1–10.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Miriah Steiger, Timir 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Petter Törnberg. 2024. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Dennis 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, et al. 2025. From selection to generation: A survey of LLM-based active learning. *arXiv preprint arXiv:2502.11767*.

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*, pages 13088–13103. Association for Computational Linguistics.

# A LLM Annotation Prompts

## A.1 Stage 1: Topic Prefiltering Prompt

**Prompt A.1: 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: LLM prompts. **A.1**: Prefiltering prompt applied to all 277,902 comments (temperature = 0, JSON output).

## A.2 Stage 2: Anti-Immigrant Annotation Prompt

**Prompt A.2: GPT-5.2 Anti-Immigrant Annotation**

```
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: **A.2**: Classification prompt applied to 25,974 prefiltered comments via GPT-5.2 Batch API (temperature = 0, structured JSON output).

# B LLM 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 1,982 seconds (0.55 wall-clock hours; 2.20 GPU-hours total) at a throughput of approximately 140 comments/second. At the representative Lambda Labs A100 spot rate of $1.10 per GPU-hour, the estimated infrastructure cost is $2.42 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).** Each request comprised the fixed classification prompt (≈240 tokens), comment text (≈20 tokens on average), and a JSON schema definition (≈50 tokens), yielding approximately 310 input tokens per call. Each response consisted of a structured JSON object with a binary label and brief explanation (≈80 output tokens). Across 25,974 comments this amounts to approximately 8.1M input tokens and 2.1M output tokens. At GPT-5.2 Batch API rates of $1.75 per 1M input tokens and $14.00 per 1M output tokens: $14.18 (input) + $29.12 (output) = **$43.30 USD** ($1.67 per 1,000 comments).

**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 Annotation Study Details

## C.1 Annotator Demographics

The six annotators who completed all phases ranged in age from 19 to 60 (mean 40), with four identifying as female and two as male. All six were residents in Germany at the time of the study. Four

were native German speakers; the remaining two were highly proficient non-native speakers (one Japanese national, one Portuguese national), both listing German among their fluent languages on Prolific.

### C.2 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.

### C.3 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.

### C.4 Inter-Annotator Agreement

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

| Metric | Q1 | Q2 | Label |
|---|---|---|---|
| Krippendorff's $\alpha$ | 0.574 | 0.493 | 0.494 |
| Cohen's $\kappa$ (avg) | 0.580 | 0.488 | 0.490 |
| Cohen's $\kappa$ 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 $\alpha$ | 0.591 | 0.432 | 0.433 |
| Cohen's $\kappa$ (avg) | 0.599 | 0.437 | 0.439 |
| Cohen's $\kappa$ 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 $\alpha$ is used as Fleiss' $\kappa$ 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 $\kappa$ 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 $\kappa$ 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.

### C.5 Per-Annotator Statistics

Table 12 reports per-annotator average $\kappa$ (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 $\kappa$ (on the derived binary label, averaged over all pairwise comparisons) and ANTI-IMMIGRANT label rate, by study phase.

### C.6 Class-Conditional Agreement

Agreement is higher for the majority class across both phases. On the pilot (200 items, majority vote), items labelled NOT ANTI-IMMIGRANT (n=156) show 90.2% average agreement with 63% unanimity, while items labelled 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.

## D Dataset Statistics

### D.1 Preprocessing Pipeline

| 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 13: Per-step preprocessing counts applied to 467,660 raw TikTok comments collected via the TikTok Research API. GlotLID (Kargaran et al., 2023) is used for language identification to retain only German-language comments.

## E Discourse Topic Structure

Figure 7 and Table 14 provide the full topic structure underlying the error analysis in Section 4.5. Each cluster is labelled by its dominant keyword pattern; human and LLM label rates (H% and L%) are reported for the annotated subset.

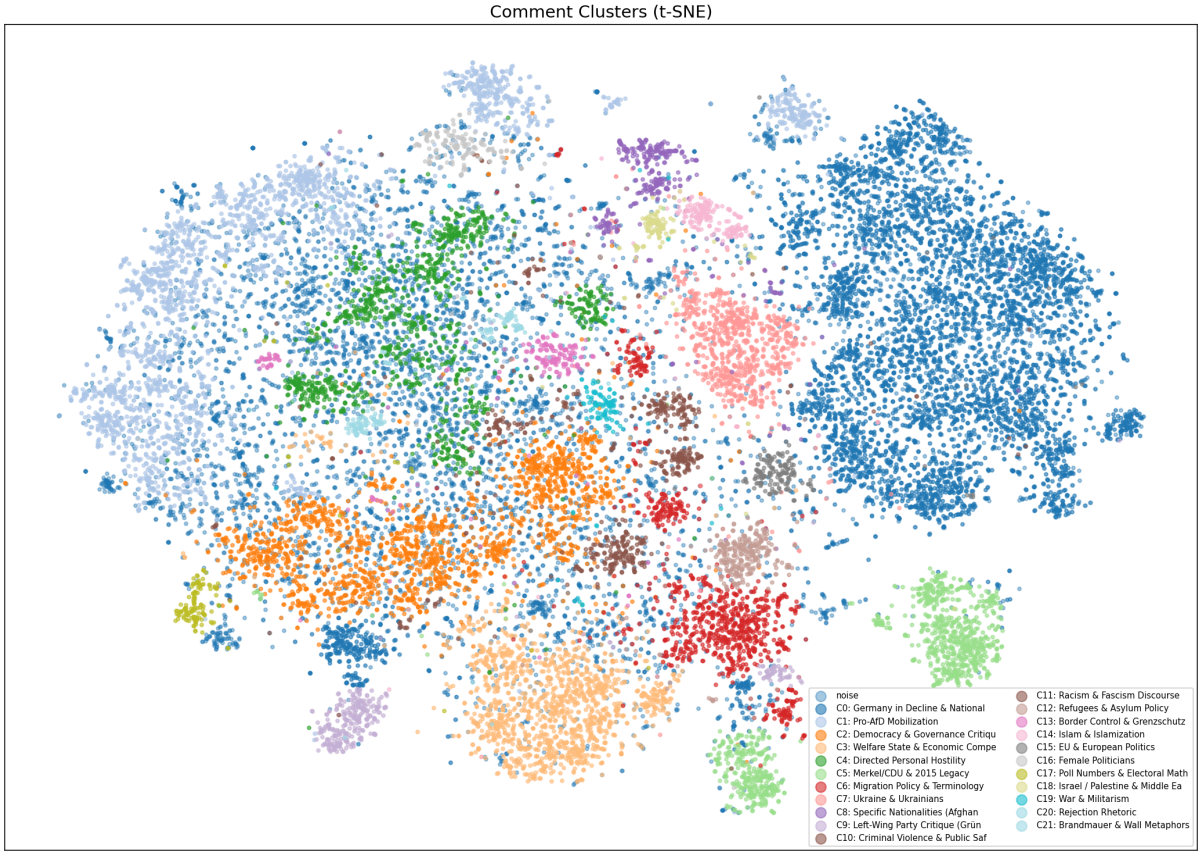


Figure 7: t-SNE projection of all 25,974 immigration-relevant comments, clustered into 22 named discourse topics via BERTopic (Grootendorst, 2022). Colours denote topic membership; grey points indicate the noise category. Topics span explicit hostility (C10: Criminal Violence & Public Safety; C14: Islam & 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). The clusters showing the strongest FP-dominant bias under LLM conditions are discussed in Section 4.5.

## F Disagreement Examples by Cluster

This section provides the full inventory of human–LLM disagreement cases summarised in Section 4.5.

**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 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 dehumanising language.

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

Table 14: Human–LLM label agreement across all 22 BERTopic discourse topics. Agr. = label agreement rate on the overlapping human-annotated subset; H% and L% = human and LLM ANTI-IMMIGRANT label rates. Topics sorted by cluster ID.

# G Additional Results

## G.1 All-Model Learning Curves

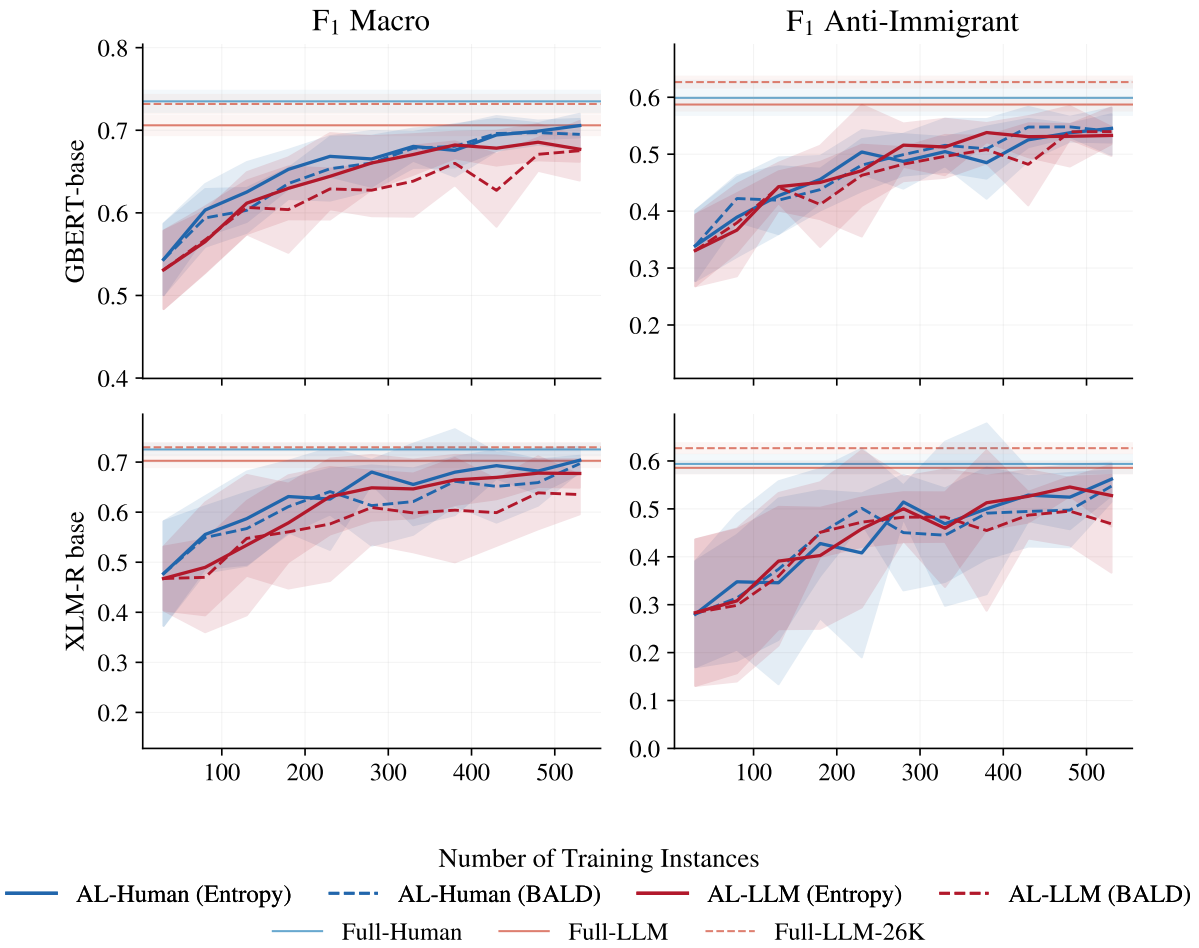


Figure 8: Learning curves for gbert-base (top) and xlm-r-base (bottom) on F1-Macro (left) and F1-Anti (right). Shaded regions indicate $\pm 1$ SD across 10 seeds. The relative ordering of annotation strategies is consistent with german_bert and ModernGBERT (Figure 2). xlm-r-base shows higher variance and lower absolute performance, confirming that German-specific models are more robust on this task.

## G.2 Confidence Distributions

Figure 9 shows the predicted P(ANTI-IMMIGRANT) distributions split by error type for german_bert. FP distributions for LLM-trained conditions peak substantially higher than for human-trained conditions, confirming that LLM-trained classifiers assign high confidence to their false positives. The pattern is consistent across all four encoder models.

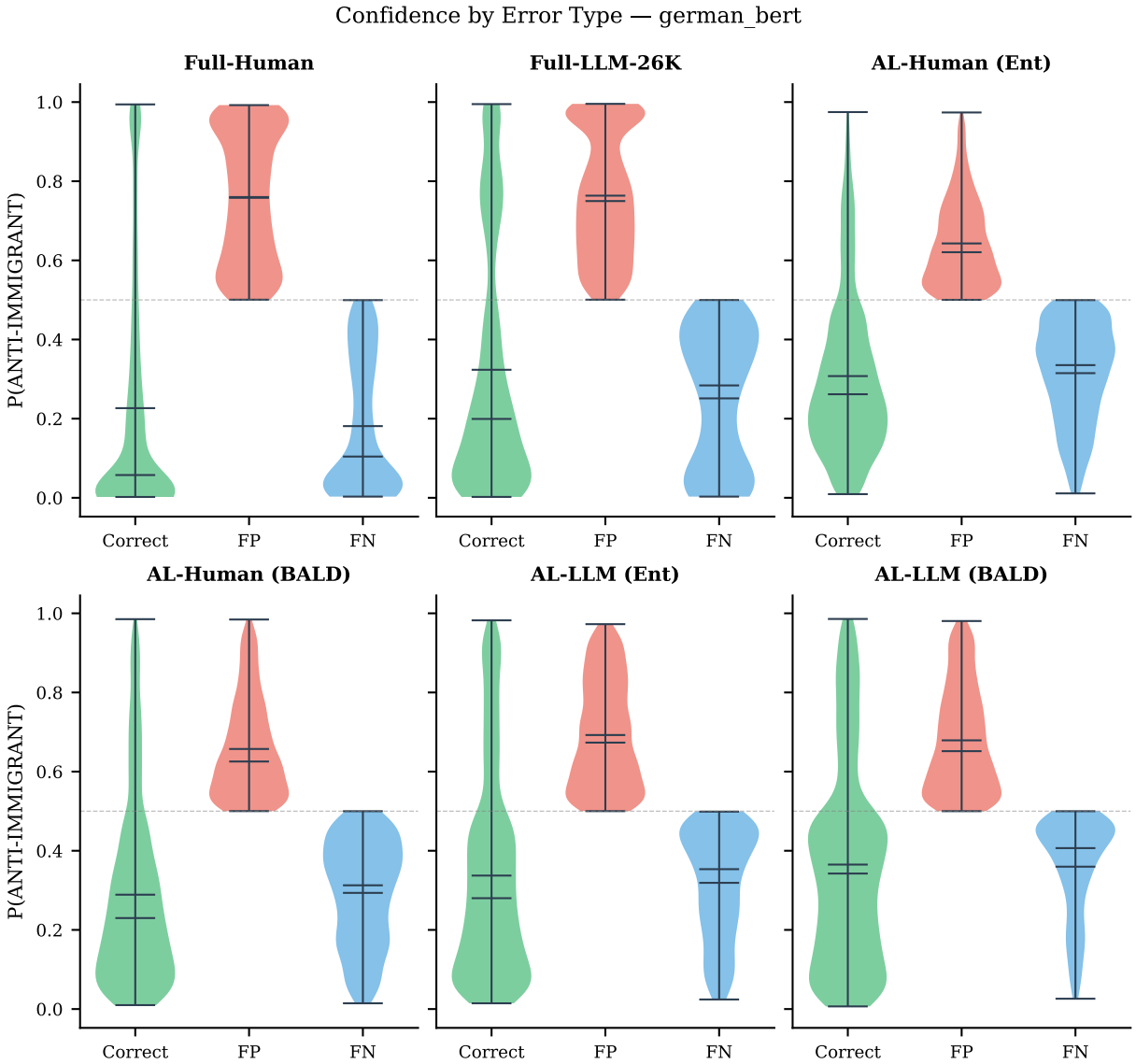


Figure 9: Confidence distributions for `german_bert`.

## G.3 Topic Error Rates

Figure 10 shows per-topic FP–FN imbalance across all nine conditions for german_bert. The FP-dominant pattern is consistent across annotation strategies. Notable exceptions are AL Human (Ent), which is the coldest column overall, and EU & European Politics, which is consistently FN-dominant under all human-trained conditions.

Figure 10: FP–FN imbalance by topic and condition for `german_bert`. Positive values indicate FP-dominant conditions; negative values indicate FN-dominant conditions. AL-Human (Ent) is consistently the least FP-skewed condition across topics.