

Does LLM Instability Reflect Annotator Disagreement? A Study on Hate Speech Detection

Aradhana Mahati*

Sneha Prashanth*

Maksym Ziemlewski*

Nicolas Loaiza Atehortua*

Urja Khurana

EEMCS, Delft University of Technology

{a.mahati, s.prashanth, m.ziemlewski, n.loaizaatehortua}@student.tudelft.nl
u.khurana@tudelft.nl

Abstract

Offensive Content Warning: This report contains some examples of hateful content strictly for the purposes of enabling this research. Hate speech detection is inherently subjective because annotators routinely disagree on the same item, and collapsing their labels into a majority vote discards information about how contested the item actually is. Large language models exhibit variation by returning different labels for the same input across random seeds. We ask whether this variation empirically lines up with human disagreement. We operationalize annotator disagreement via entropy and the LLM instability via flip rate and model entropy across five seeds at fixed temperature, then measure their rank correlation on the full CREHate dataset using three models from both open- and closed-source families. The alignment with human disagreement is model dependent: Gemini 2.5 Flash and Llama 3.1 8B show significant positive correlations, while Mistral 7B shows only a marginal effect. The relationship is also modest in absolute scale: most items cluster at a flip rate of zero, and the signal is concentrated in a small fraction of high-disagreement items that also produce model instability. Qualitative analysis shows that culturally contested items, where human review is needed, sometimes produce zero model instability, as models classify them in opposite directions across cultural contexts. Our findings do not show evidence that seed-based instability is a reliable proxy for human disagreement for all LLMs.

1 Introduction

Hate speech detection sits at the intersection of two tensions in NLP: the need to moderate harmful content at scale (Gorwa et al., 2020; Gillespie, 2020), and the fact that what counts as harmful is genuinely contested (Fortuna and Nunes, 2018; Sap et al., 2022). Annotation studies make this

tension concrete. When multiple people label the same item, they often disagree, and that disagreement is not just noise. It reflects real differences in interpretation, lived experience, and values (Mostafazadeh Davani et al., 2022). Yet the standard pipeline collapses these per-annotator labels into a single majority-vote “ground truth,” throwing away exactly the information that tells us how ambiguous an item really was (Sachdeva et al., 2022).

Large language models show variation in the sense that when a user prompts the same input under different random seeds, it can return different labels from run to run (Atil et al., 2024), with predictions shifting further still under minor prompt paraphrases (Elazar et al., 2021; Jang et al., 2022). We refer to this run-to-run variation under fixed input as *instability*. We do not assume from the start that an LLM’s variation inherently reflects human disagreement; we test whether this variation empirically lines up with human disagreement.

In this work, we examine whether these two forms of variability are related. We hypothesize that if an item is genuinely contested, a model that has picked up on features correlated with that might also waver across stochastic decoding runs, for reasons that need not resemble human disagreement. If that correspondence holds, LLM instability could work as a cheap, automatic signal for flagging contested items. If it does not, that is informative too: seed variance and human disagreement would be tracking two different things. So we ask: **Does LLM output instability across random seeds correlate with human annotator disagreement on hate speech detection?**

To address this question, we first operationalize annotator disagreement and LLM instability so that they are directly comparable, building on (Elazar et al., 2021; Jang et al., 2022). We then investigate whether LLMs from different families show similar instability patterns on high-disagreement items and whether items that have high human disagreement

*Equal contribution.

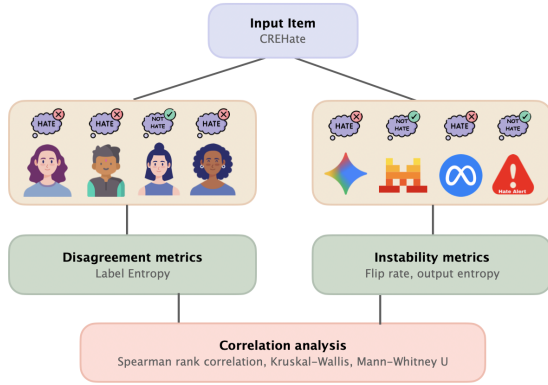


Figure 1: Overview of the three-stage methodology: computing human disagreement scores and tiers from per-annotator labels (Stage 1), measuring per-item LLM instability across random seeds (Stage 2), and analysis of these signals (Stage 3).

are similar to the items where LLMs are unstable on.

We conduct our experiments on the CREHate dataset (Lee et al., 2024), which looks at cross-cultural hate speech detection, using three open- and closed-source model families: Gemini 2.5 Flash, Llama 3.1 8B and Mistral 7B. Our findings indicate that Gemini 2.5 Flash and Llama 3.1 8B show significant positive rank correlations between flip rate and human entropy while Mistral 7B shows only a marginal effect. The correlation in every case is driven by a small subset of items, since the majority of items across all models sit at a flip rate of zero regardless of human disagreement. Additionally, we conduct a quantitative analysis with a content-type categorization of 39 high-disagreement items and find that harmful content accounts for over half of these items, followed by political commentary, with a threshold-based case analysis further showing that culturally contested items produce zero model instability despite maximal human disagreement, since models classify them confidently but in opposite directions across cultural contexts. Thus, our findings do not suggest any evidence for hate speech practitioners to use model instability as a signal for potential human disagreement.

2 Related Work

Human label variation and data perspectivism.

Foundational work on crowd annotation rejected the assumption of a single correct label, arguing that disagreement between annotators reflects the

range of plausible interpretations or perspectives of an item rather than annotation error (Aroyo and Welty, 2015). This has also been referred to as *human label variation*: an informative signal arising from subjectivity, ambiguity, or multiple plausible readings (Plank, 2022) or data perspectivism (Cabrita et al., 2023), where different perspectives of individuals should be included when representing the ground-truth. The different types of disagreement have been investigated by manually labelling the type of disagreement (Sandri et al., 2023). During the data annotating process, annotation paradigms can either prescribe explicit instructions to follow or deliberately elicit subjective judgements (Röttger et al., 2022), and corpora such as Measuring Hate Speech release per-annotator labels by design (Sachdeva et al., 2022). CREHate extends this line cross-culturally, showing that hate speech labels diverge along the national background of annotators from five English-speaking countries (Lee et al., 2024). Preserving raw per-annotator labels, rather than a majority vote, is the methodological prerequisite for our study.

Soft Labels. A survey on learning from disaggregated labels consolidated methods that train directly on the full annotation distribution instead of an aggregated label (Uma et al., 2021). A soft label represents an item by this full distribution over annotator responses instead of a single majority vote (Uma et al., 2020). Models trained on soft labels are optimized to match this distribution, usually with a distance-based loss such as cross-entropy or Jensen-Shannon divergence (Fornaciari et al., 2021). Evaluation compares the model’s output distribution with the human distribution using the same distance metrics, or by correlating the entropy of the soft label with the entropy of the model’s output distribution (Uma et al., 2020), amongst other methods (Rizzi et al., 2024; Khurana et al., 2026). Our flip rate and output entropy metrics follow the same idea for LLMs: instead of eliciting a label distribution directly, we build one from five sampled outputs and compare its entropy to the human distribution.

LLM consistency. Literature examines whether language models respond consistently. Pretrained models give divergent answers to semantically equivalent factual probes (Elazar et al., 2021), benchmark suites formalise multiple consistency types for behavioural evaluation (Jang et al., 2022), and encoder classifiers vary in their predictions

across training runs on subjective tasks (Khurana et al., 2021). Closest to our experiment, repeated LLM calls with identical inputs and identical settings yield non-identical outputs, with accuracy varying across runs even under nominally deterministic configurations (Atil et al., 2024). On the sampling side, temperature controls how much of the probability tail is available during decoding, and higher values increase output diversity at the expense of consistency (Renze, 2024); yet non-determinism persists even at temperature zero, meaning that instability is not purely a function of temperature but a property of the generation process itself (Atil et al., 2024). This body of work treats instability as a property of the model measured against a fixed standard; it does not relate run-to-run variation to the distribution of human judgements on the same items.

Hate speech detection. Surveys of hate speech detection document the progression from feature-based to neural classifiers and the persistent difficulty of defining the target category (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Fine-tuned encoder classifiers such as HateXplain remain a deterministic point of comparison for prompted LLMs (Mathew et al., 2021). The task is subjective in a measurable way: annotator beliefs and identities shift toxicity ratings on the same items (Sap et al., 2022), and several corpora release the underlying raw annotations (Kennedy et al., 2022; Sachdeva et al., 2022; Lee et al., 2024). Hate speech detection therefore offers documented disagreement, per-annotator data, and real deployment of LLMs, making it a suitable testbed for our question.

Model uncertainty vs. human disagreement. The line of prior work closest to ours compares model uncertainty to human judgement distributions. Calibrating predictive probabilities to the majority label is theoretically problematic when humans inherently disagree, motivating instance-level comparison against the full label distribution (Baan et al., 2022). Empirical studies find that LLM-produced label distributions align poorly with annotator distributions on subjective tasks (Pavlovic and Poesio, 2024), and a recent evaluation across tasks and model families reports that LLMs predict majority labels well but fail to capture annotator disagreement, with reasoning-oriented training degrading this ability further (Ni et al., 2026). These studies elicit uncertainty explicitly, through

Entropy	Split	Count	%
0.000	5-vs-0 (unanimous)	888	56.2
0.722	4-vs-1	403	25.5
0.971	3-vs-2	289	18.3
Total		1,580	100.0

Table 1: Distribution of annotator label entropy in CREHate ($n = 1,580$). With five binary annotators, only three entropy values are possible. The majority of items are unanimous across all five countries.

prompted soft labels or token probabilities. No prior work tests whether the instability a model exhibits across random seeds tracks item-level human disagreement. Our study measures exactly this correspondence, treating seed-based instability as a candidate implicit signal of the divergence that perspectivist work locates in human labels.

3 Methods

Our methodology proceeds in three stages, shown in Figure 1. In Stage 1 we compute a human disagreement score for each item from its per-annotator labels and stratify items into tiers. In Stage 2 we query each model across multiple random seeds to produce a per-item instability score. In Stage 3 we analyze the two signals and measure their rank correlation. The rest of this section details each component.

3.1 Dataset

We use CREHate (Lee et al., 2024), a dataset designed to surface cross-cultural disagreement in hate speech annotation. Items were annotated by native speakers from five English-speaking countries (Australia, the United Kingdom, the United States, South Africa, and Singapore). CREHate provides a single aggregated binary label per country rather than individual annotator responses; we treat each country’s label as one vote, giving five votes per item. Disagreement between these five votes therefore reflects cross-national divergence rather than within-population variation. We use the full dataset of 1,580 items; 871 are labelled *not_hate* (55.1%) and 709 are labelled *hate* (44.9%) by majority vote.

3.2 Disagreement Metrics

We quantify human disagreement using **label entropy**:

$$H_i = - \sum_{c \in \mathcal{C}} p_{i,c} \log_2 p_{i,c} \quad (1)$$

where $p_{i,c}$ is the proportion of annotators assigning label c to item i . Entropy is zero when all annotators agree and reaches its maximum when labels are split evenly. For CREHate the three possible entropy values and their distribution are described in Section 3.1. We use entropy as the continuous signal in Spearman rank correlation and as the x -axis in scatter plots (Figure 3 and Figure 4).

Applied to CREHate: Because each item has exactly five binary annotators, only three entropy values are possible. Table 1 shows their distribution. The majority of items (56.2%) are labelled unanimously by all five countries, meaning that cross-cultural consensus is actually the norm in CREHate rather than the exception. A further 25.5% show a 4-vs-1 split, and 18.3% show a genuine 3-vs-2 split. Taken together, 43.8% of items carry some cross-cultural disagreement, meaning our correlation analysis draws on a substantial minority subset rather than a handful of outliers. Because Spearman rank correlation depends only on the relative ordering of items rather than the proportion at each disagreement level, the method remains valid even though unanimous items dominate the dataset numerically.

Stratifying into disagreement tiers: We stratify items into disagreement tiers using entropy tercile boundaries. The 33rd-percentile boundary falls at 0.000 and the 66th-percentile boundary falls at 0.722, producing a low tier of items where all five countries agree ($n = 888$) and a high tier of items where at least one country dissents ($n = 692$). No item falls strictly between 0.000 and 0.722, so there is no medium stratum.

3.3 LLM Instability Metrics

To measure LLM instability we adopt a repeated-sampling design. Each item is presented to each model using the same base prompt (Appendix A) and scored five times under distinct random seeds (seeds 1000 through 1004). We set temperature to $T = 0.7$ to permit variation within standard generation parameters used in content moderation evaluation bounds (Renze, 2024).

Flip rate is our primary instability metric: the proportion of the five runs that produce a label different from the modal label for that item. A flip rate of zero means the model gave the same answer every time; a rate of 0.4 means two of the five runs diverged from the majority.

Output Entropy is the secondary measure we compute over the five labels, mirroring the human entropy computation. Because CREHate uses five binary annotators and we run five binary seeds, both the human and model entropy distributions share the same support, making direct comparison on the same scale possible.

For binary classification with five runs, output entropy and flip rate are monotonically related, since the three possible flip rate values (0.0, 0.2, 0.4) map one-to-one to the three possible entropy values (0.000, 0.722, 0.971). Both metrics are item-level quantities that can be directly compared with human disagreement scores.

3.4 Models

We evaluate three instruction-tuned models spanning open- and closed-source families and a range of access modes.

Gemini 2.5 Flash is a closed-source model from Google, accessed via the Vertex AI API, representative of proprietary models increasingly used for content moderation at scale.

Llama 3.1 8B (Touvron et al., 2023) is an open-weight model from Meta at roughly 8B parameters.

Mistral 7B (v0.3 instruction-tuned; Jiang et al. 2023) is a second open-weight model of comparable size but a different pretraining corpus and architecture.

3.5 Statistical Analysis

We conduct three layers of statistical analysis. First, to test whether LLM instability is monotonically related to human disagreement at the item level, we compute **Spearman’s rank correlation** (ρ) between flip rate and human entropy for each model. Spearman correlation is appropriate here because both distributions are heavily right-skewed: the majority of items receive a flip rate of zero, with a small tail of more unstable items.

Second, to test whether the two disagreement tiers produce reliably different instability distributions, we apply the **Kruskal-Wallis test**, a non-parametric one-way ANOVA on ranks. Third, we follow up with pairwise **Mann-Whitney U tests** between tiers to locate where the difference lies. We report significance at $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***); all tests are two-sided.

Model	Acc.	Prec.	Rec.	F1
Gemini 2.5 Flash	80.2±0.6	80.4±0.6	79.5±0.7	79.7±0.6
Llama 3.1 8B	74.6±0.6	74.8±0.6	75.0±0.6	74.6±0.6
Mistral 7B	65.0±0.5	66.6±0.3	62.5±0.6	61.4±0.8

Table 2: Classification performance (% , macro-averaged) against the CREHate majority-vote gold label ($n \approx 1,580$ per seed), mean $\pm$ std across the 5 seeds.

Model	Low	High
Gemini 2.5 Flash	0.016	0.054
Llama 3.1 8B	0.037	0.071
Mistral 7B	0.005	0.009

Table 3: Mean flip rate by model and disagreement tier on CREHate ($n = 1,580$).

4 Results

4.1 Classification Performance

For completeness, we report standard classification performance against the CREHate majority-vote gold label (Table 2), computed independently for each of the five seeds and averaged. Gemini 2.5 Flash is the strongest classifier (macro-F1 = 0.797 ± 0.006), followed by Llama 3.1 8B (0.746 ± 0.006) and Mistral 7B (0.614 ± 0.008). This ordering lines up with the Spearman correlations in Table 4: the model that classifies best is also the model whose instability aligns most closely with human disagreement, while Mistral’s weaker classifier is also the one whose flip rate carries the least signal.

4.2 Flip Rate Patterns Across Tiers

Table 3 and Figure 2 report mean flip rates for each model across the two CREHate disagreement tiers. The directional pattern is consistent across all generative models: high-disagreement items produce higher flip rates than low-disagreement items. The magnitude of that difference varies substantially.

Gemini 2.5 Flash moves from 0.016 at the low tier to 0.054 at the high tier, a 3.4-fold increase. Llama 3.1 8B has the highest absolute flip rates throughout (0.037 at low, 0.071 at high), though its relative increase is smaller. Mistral 7B stays near zero across both tiers (0.005 at low, 0.009 at high), a difference too small to carry any practical signal.

4.3 Spearman Rank Correlation

Table 4 reports item-level Spearman correlations between flip rate and human entropy. Gemini 2.5

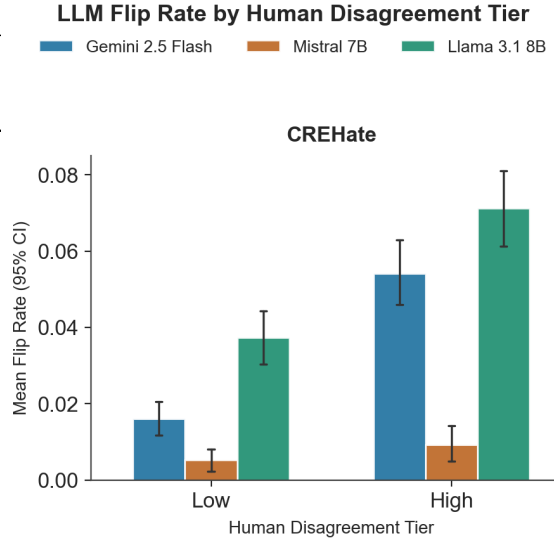


Figure 2: Mean LLM flip rate (95% CI) per human disagreement tier for Gemini 2.5 Flash, Mistral 7B, and Llama 3.1 8B on CREHate. Both Gemini and Llama show a significant increase in instability at the high-disagreement tier ($p < 0.001$, Kruskal-Wallis).

Model	ρ	p	n
Gemini 2.5 Flash	0.217	<0.001***	1,580
Llama 3.1 8B	0.159	<0.001***	1,580
Mistral 7B	0.053	0.035*	1,579 ¹

Table 4: Spearman rank correlation between LLM flip rate and human label entropy on CREHate. Bold indicates $p < 0.001$.

Flash shows the strongest alignment ($\rho = 0.217$, $p < 0.001$, $n = 1,580$), and Llama 3.1 8B also reaches significance ($\rho = 0.159$, $p < 0.001$, $n = 1,580$). Mistral 7B yields a positive but small coefficient ($\rho = 0.053$, $p = 0.035$, $n = 1,579$) that amounts to a marginal effect.

Figure 3 shows the underlying item-level distributions. In every panel, the majority of items sit at flip rate zero, forming a dense cluster along the x-axis. In the Gemini panel, the sparse non-zero points tend to fall at higher entropy values, producing a positive regression slope. In the Llama panel, high-flip-rate points are spread more evenly across entropy levels, which explains the weaker coefficient despite Llama’s higher absolute flip rates overall.

¹One item, crehate-SBIC-476, was excluded for Mistral: the model returned the non-binary label "needs_review" on all five seeds rather than hate/not_hate, so no valid prediction could be parsed.

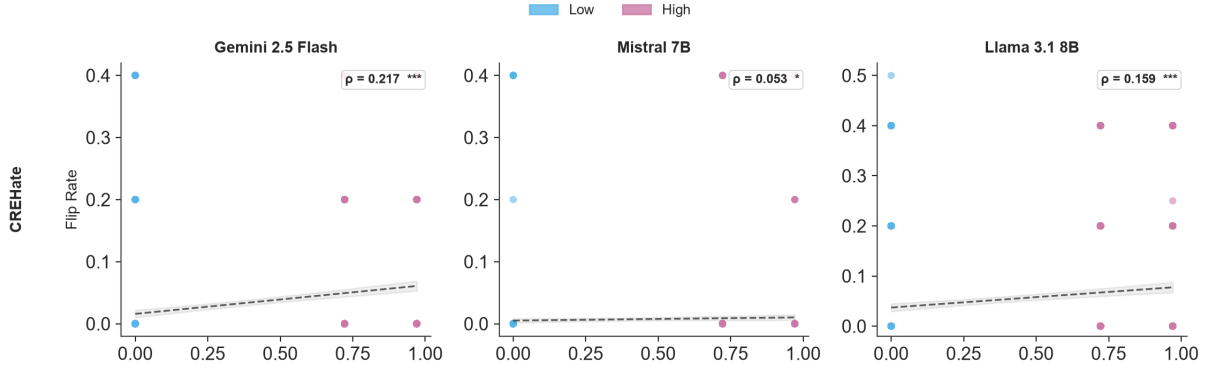


Figure 3: **Flip rate vs. human entropy** scatter plots on CREHate ($n = 1,580$), arranged by model. Points are coloured by disagreement tier (blue = low, pink = high). Dashed lines show OLS regression fits with 95% confidence bands. The vast majority of items have flip rate zero, producing a dense cluster at the base of each panel; non-zero points drive the regression fits almost entirely. Only the Gemini panel shows a clearly positive slope, consistent with the strongest Spearman result.

4.4 Model Entropy vs Human Entropy

Figure 4 plots model output entropy directly against human label entropy, placing both on the same scale. The dotted diagonal marks where the two quantities are equal; points below the diagonal mean the model is more confident than the human annotators on that item.

For all three models, the overwhelming majority of points sit well below the diagonal. Even on high-disagreement items, where human entropy reaches its maximum of 0.971, most models return entropy of 0. This confirms that the significant Spearman correlations reported in Table 4 are driven by a small subset of items rather than a broad correspondence across the distribution. The few points near the diagonal are concentrated in the high-disagreement tier, and are most visible in the Gemini panel, consistent with its stronger correlation.

As noted in Section 3, output entropy and flip rate share the same rank order for five binary runs, so the Spearman values here are identical to those in Table 4. The entropy framing adds a different reading: it quantifies how much more confident models are than humans, not just whether that confidence varies with disagreement.

4.5 Kruskal-Wallis and Pairwise Tier Tests

Table 5 reports Kruskal-Wallis statistics and epsilon-squared (ϵ^2) effect sizes, testing whether flip-rate distributions differ between the low and high tiers. Gemini 2.5 Flash ($H = 72.12$, $p < 0.001$) and Llama 3.1 8B ($H = 38.90$, $p < 0.001$) both show a highly significant tier difference. Mis-

Model	H	p	ϵ^2
Gemini 2.5 Flash	72.12	< .001***	0.045
Llama 3.1 8B	38.90	< .001***	0.024
Mistral 7B	3.22	.073	0.001

Table 5: Kruskal-Wallis H statistics and epsilon-squared (ϵ^2) effect sizes comparing flip rate across disagreement tiers on CREHate ($n = 1,580$). *** $p < .001$.

Model	Comparison	p
Gemini 2.5 Flash	Low vs. High	<0.001***
Llama 3.1 8B	Low vs. High	<0.001***
Mistral 7B	Low vs. High	0.073

Table 6: Mann-Whitney U pairwise tier comparisons on CREHate ($n = 1,580$). Mistral 7B does not reach significance.

tral 7B does not reach significance ($H = 3.22$, $p = 0.073$), consistent with its near-zero flip rates across both tiers.

Table 6 reports pairwise Mann-Whitney comparisons, testing whether one tier’s flip rates are stochastically larger than the other’s. For Gemini and Llama the difference is significant at $p < 0.001$; for Mistral it does not reach the conventional threshold.

4.6 Qualitative Characterisation of High-Disagreement Items

We manually reviewed the 39 items from the full CREHate dataset (1,580 items) that exhibited both high human disagreement and high cross-model instability, to identify the types of content that most frequently produced annotation instability. An item

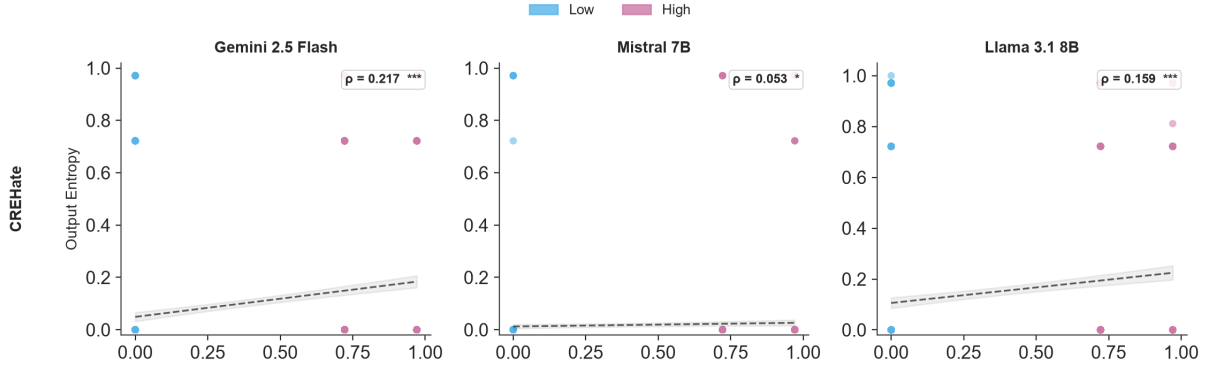


Figure 4: **Model output entropy vs. human label entropy** on CREHate ($n = 1,580$), arranged by model. The dotted diagonal marks where model entropy equals human entropy; points below it indicate the model is more confident than the human annotator panel. Dashed lines show OLS regression fits with 95% confidence bands. Both axes share the same three discrete values (0.000, 0.722, 0.971), producing visible point clusters rather than a continuous cloud.

Content Type	Count	%
Offensive	22	56.4%
Political	13	33.3%
Explicit	4	10.3%
Total	39	100%

Table 7: Content type distribution among the 39 high-disagreement, high-instability CREHate items.

was flagged as high human disagreement if its annotator entropy fell in the top tercile of the dataset (tier = *high*) and at least one generative model (Gemini, Mistral, or Llama) flipped its predicted label across five repeated runs. An item was additionally flagged as high cross-model instability if two or more of the three generative models flipped label across their five runs.

The manual review was conducted jointly by two of the authors, who discussed each item together and reached category assignments by consensus. Three recurring content patterns emerged inductively from this review: content targeting a specific community or group (*offensive*), content centred on political or nationalistic discourse (*political*), and content involving explicit or sexual language (*explicit*). Each item was assigned to the category reflecting its dominant theme.

The dominance of offensive content reflects CREHate’s cross-cultural design, where interpretations of hostility and derogatory language vary across annotator communities; within this category, the reviewers noted that a majority of items were *dark jokes*, whose humorous framing plausibly softens perceived hostility for some readers

while others react to the literal target, destabilising both human and model judgements. The political cluster reflects persistent disagreement over where legitimate commentary ends and discriminatory rhetoric begins. Across categories, disagreement stems from a shared tension: whether surface hostility, vulgarity, or humour alone constitutes hate speech, or whether discriminatory intent toward a protected group must also be present.

4.7 Qualitative Threshold Matrix Analysis

To explain the mechanical patterns in our statistical results, we select four items at the extremes of human entropy (H) and flip rate (FR), forming a 2×2 matrix of human disagreement and model instability.

Profile 1: True signal (High H / High FR): crehate-SBIC-582 (“*Smack these h*es on they ass like we birthed these b*tches*”) split annotators nearly evenly ($H = 0.971$). Based on the pattern of the split alone, we interpret this as a possible division between readers who took the phrasing as rap vernacular and readers who read it as misogynistic hate speech; the dataset does not record annotators’ own stated rationale, so we offer this as our speculative interpretation rather than a confirmed account. Gemini 2.5 Flash produces variable output on the same item ($FR = 0.4$): both readings are plausible, so seed differences tip it either way. This is the item in our qualitative sample where model instability and human disagreement most plausibly track the same underlying ambiguity.

Profile 2: Generation noise (Low H / High FR): crehate-SBIC-143 (“...*I just want to state my hate for them /s might be needed*”) was unanimous among annotators ($H = 0.0$); the sarcasm tag does not cancel explicit racial animosity. Nevertheless, Llama 3.1 8B still splits 3-vs-2 across seeds ($FR = 0.4$). Rather than reflecting genuine semantic uncertainty, we interpret this as an instance where the model’s stochastic sampling appears to be influenced by surface formatting tokens like punctuation and slang markers, producing model instability, while humans agree.

Profile 3: Cultural knowledge gap (High H / Low FR): crehate-AU-Reddit-21 (“*Typical abos...*”) produced near-maximal human disagreement ($H = 0.971$) along cultural lines: Australian, UK, and South African annotators labelled it hate speech; US and Singaporean annotators did not. The word “abos” is a severe slur in Australia but unknown elsewhere. All models returned $FR = 0.0$, confidently but split: Gemini 2.5 Flash called it hate across all seeds, while Llama 3.1 8B and Mistral 7B both returned not hate. We interpret this to mean that when disagreement stems from local knowledge gaps, models do not waver; instead, output instability is absent as they lean on fixed regional biases.

Profile 4: Clean baseline (Low H / Low FR): crehate-SBIC-23 (“*From 1480 to 1913 European Queens waged far more wars than European Kings*”) was unanimously classified as not hate speech by all five country annotators ($H = 0.0$). All three generative models returned *not_hate* on every seed ($FR = 0.0$). The item is a neutral historical observation with no hate target and no ambiguous phrasing, making this the clearest possible case: human agreement and model instability coincide because neither humans nor models find anything to contest.

5 Discussion

Our guiding question was whether across-seed LLM instability lines up human annotator disagreement closely enough to proxy for label disagreement. On CREHate, the answer is partially yes, but the effect is model-dependent and limited in scope.

We first operationalized disagreement and instability for direct comparison. We reduce both to item-level variability scores, and directly compute Spearman correlation between flip rate and entropy.

The significant correlations we find for Gemini and Llama suggest the two measures share some signal. Shared form does not imply shared cause, though. Flip rate captures a single stochastic process wavering across runs; human entropy captures a population of annotators diverging over meaning. The two overlap where a model sits on the same semantic fence that divides annotators, which our results suggest is more common for capable models, but far from universal even for them.

We then investigated whether models of different sizes and families show similar instability patterns. They do not, and the differences are not simply scalar. Gemini achieves the highest rank correlation ($\rho = 0.217$) at comparatively low absolute flip rates, meaning its instability is concentrated on genuinely contested items. Llama 3.1 8B has higher absolute flip rates but a weaker correlation ($\rho = 0.159$), because part of its variance comes from surface sensitivity rather than semantic variance, as Profile 2 illustrates with the sarcasm tag. Mistral 7B is so conservative that it cannot encode contention at all: with flip rates barely above zero, there is no variation left to correlate. Instability magnitude and informativeness are therefore decoupled across model families.

Furthermore, the items that drive human disagreement and those that drive model instability overlap, but only partially. Both Gemini and Llama produce significantly higher flip rates on high-disagreement items (Kruskal-Wallis $H = 72.12$ and $H = 38.90$ respectively, both $p < 0.001$), and the scatter in Figure 3 shows a positive trend for both. At the same time, the majority of high-disagreement items still receive a flip rate of zero, so the signal is sparse. Profile 3 captures the hardest failure case: the slur “abos” divides annotators sharply along cultural lines, but all models respond with complete confidence. Disagreement rooted in localized cultural knowledge produces no instability at all, because the model is not uncertain about what it reads. It is wrong in different directions depending on its pretraining.

The correlations remain modest, and the practical utility of seed-based instability as a moderation signal is limited. For every high-disagreement item where instability fires, many others go undetected. A content-moderation system that keys escalation to flip rate would stay silent on a large fraction of the most contested content.

6 Conclusion

We tested whether LLM output instability across random seeds correlates with human annotator disagreement in hate speech detection, using the full CREHate dataset of 1,580 items annotated by speakers from five countries. Gemini 2.5 Flash and Llama 3.1 8B both show significant positive correlations with cross-cultural human disagreement, while Mistral 7B shows only a marginal effect. The direction is consistent: high-disagreement items produce more instability than low-disagreement ones across all generative models. The magnitude, however, is modest. Most items receive a flip rate of zero regardless of how contested they are among humans, and the correlation is driven by the subset where model uncertainty and human disagreement happen to coincide.

Profile 3 shows the most important failure case: a culturally specific slur that splits an international annotator panel produces no instability at all, because no model is uncertain about it. All three generative models return zero flip rate on this item regardless of architecture or provider, suggesting the blind spot is not specific to any one model’s sampling behaviour.

Our qualitative review of high-disagreement items finds that offensive content (56.4%), explicit language (10.3%), and political commentary (33.3%) account for the bulk of human disagreement. The threshold matrix analysis adds nuance: when instability does appear, it sometimes tracks genuine semantic variance (Profile 1, Gemini) and sometimes reflects surface-level generation noise (Profile 2, Llama). The line between the two is not visible from flip rate alone. Future work could probe richer uncertainty signals such as token-level logprobs, extend the analysis to other languages and subjective tasks, and increase the number of seeds to better characterise the instability distribution.

Takeaways: Our findings suggest that there is not yet sufficient evidence for hate speech practitioners to use LLM instability as a signal for an input that warrants different judgments. It aligns only for some models, only weakly, and only on a small subset of items. Most contested items produce no instability at all, so in the cases where human review is most needed. Practitioners should not read a stable model as a signal that an item is uncontested.

7 Limitations and Ethical Considerations

Our quantitative analysis relies on a single English-language dataset (CREHate, $n = 1,580$), which limits generalisability to other languages, annotation paradigms, and cultural contexts beyond those represented by the five countries in CREHate. CREHate aggregates labels at the country level only, so we cannot separate cross-national disagreement from other demographic splits (e.g. gender, age) that may drive different disagreement patterns (Sap et al., 2022). We adopt this design because cross-national divergence is what CREHate was built for (Lee et al., 2024) and matches our qualitative finding that models fail specifically on culturally-bound content (Profile 3, §4.7). With only five seeds per item, the maximum observable flip rate is 0.4, which compresses the instability scale and may hide finer variation. Additionally, decoding variance depends on temperature ($T = 0.7$); while non-determinism persists even at $T = 0.0$ (Atil et al., 2024), testing different temperatures remains an important direction for future work to fully map temperature sensitivity against human label variation. We evaluate three generative models across open- and closed-source families; testing additional architectures was infeasible due to the team size and API costs. Furthermore, cross-model variance reflects differences in pretraining corpora and safety alignment rather than human cultural divergence, and model architectural differences should not be read as an exact proxy for human demographic variation. The qualitative case studies are drawn from a small set of purposively selected items and are illustrative rather than exhaustive; interpretations of model behaviour rely on output alone, since internal probability distributions are inaccessible for closed-source systems.

Our results argue against using LLM instability as a standalone content-moderation signal. If adopted despite this, platforms could treat low-instability items as settled when they are in fact deeply contested, as Profile 3 illustrates for culturally specific slurs. A stronger future correlation should likewise not be read as licence to replace human annotator panels: the instability signal is mechanistically different from human disagreement, and substituting one for the other would erase the perspectival diversity that disagreement encodes.

Acknowledgements

We thank the anonymous reviewers for their comments that helped improve this paper. All remaining errors are our own. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research.

References

- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 36(1):15–24.
- Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. [LLM stability: A detailed analysis with some surprises](#). *arXiv preprint arXiv:2408.04667*.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. [Stop measuring calibration when humans disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhishava Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. [Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.
- Paula Fortuna and Sérgio Nunes. 2018. [A survey on automatic detection of hate speech in text](#). *ACM Comput. Surv.*, 51(4).
- Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.
- Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.
- Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. [BECEL: Benchmark for consistency evaluation of language models](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Alejandro Lara, Gabriel Cárdenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, and 3 others. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.
- Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2021. [How emotionally stable is ALBERT? testing robustness with stochastic weight averaging on a sentiment analysis task](#). In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Urja Khurana, Michiel van der Meer, Enrico Liscio, Antske Fokkens, and Pradeep K Murukannaiah. 2026. Not all subjectivity is the same! defining desiderata for the evaluation of subjectivity in nlp. *arXiv preprint arXiv:2603.28351*.
- 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.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.
- 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.
- Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Miserlis Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2026. Can reasoning help large language models capture human annotator disagreement? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–54, Rabat, Morocco. Association for Computational Linguistics.
- Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.
- 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*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.
- Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.
- 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*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging Rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don’t you do it right? analysing annotators’ disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. 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*, pages 5884–5906, Seattle, United States. 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*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

A Prompt Template

All generative models receive the same system prompt for every item:

You are annotating whether a social media post constitutes hate speech. Hate speech is content that attacks, threatens, demeans, or dehumanizes a person or group based on protected attributes such as race, ethnicity, national origin, religion, sex, gender identity, sexual orientation, disability, or serious disease. Respond with ONLY a JSON object: {"label": "hate"} or {"label": "not_hate"}. No explanation.

The user message contains only the raw post text with no additional framing.