Which definition of hate speech does the default behaviour of large language models align with most closely?

A Zero-Shot Probing Study of Two Open-Weight Models

Yuanze Xiong[1]

Supervisors: Pradeep Murukannaiah[1]**, Urja Khurana**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Yuanze Xiong
Final project course: CSE3000 Research Project
Thesis committee: Pradeep Murukannaiah, Urja Khurana, Cynthia Liem

Abstract

What counts as hate speech varies and complicates automated detection systems. Large language models (LLMs) are increasingly used for this task in a zero-shot setting, yet the intrinsic definition of hate speech that such models apply when no definition is supplied remains poorly understood. This paper probes the intrinsic, unguided conception of hate speech that two open-weight instruction-tuned models, Meta Llama 3.1 and Google Gemma 4, apply by default. We combine three complementary measurements: zero-shot binary classification, structured elicitation of Hate Speech Criteria (HSC), and a contamination control that compares both tasks with a set of novel cases, and we add two follow-up analyses: a prompt-paraphrase robustness check and a definition-injection probe on the `dominance` criterion. Both models classify hateful content with high binary accuracy and demonstrate strong target group identification. However, they fail on the `dominance` criterion, defaulting instead to a misinterpretation where almost all hostile speech is labelled as dominating. We conclude that while the default definition these LLMs apply is target-aware, its tendency toward over-inclusive criterion application constrains the reliability of unguided models for fine-grained hate speech characterisation.

1 Introduction

Hate speech is a recognised problem on online platforms, but there is no single, universally agreed definition of it. What is treated as hateful depends on the community, the legal jurisdiction, and the purpose for which a detection system is built [Khurana *et al.*, 2022]. This variation is visible in the many datasets used to train and evaluate detection models: each tends to encode its own "flavour" of hate speech, reflecting the choices its annotators made [Fortuna *et al.*, 2020]. A practitioner who wants to use such a model therefore needs to know which definition it actually applies, not merely that it labels some text as hateful.

Recent work has begun to formalise this question. Khurana *et al.* [2022] propose the Hate Speech Criteria (HSC), a modular framework that decomposes a hate speech definition into separate dimensions, such as the targeted characteristic and whether the speaker incites violence. Building on this, the DefVerify study [Khurana *et al.*, 2025] fine-tunes models on six popular datasets and measures whether the resulting models reflect each dataset's stated definition; it reports that none of the fine-tuned models encode all of the aspects specified in their dataset's definition. In parallel, large language models (LLMs) are increasingly applied to hate speech detection without task-specific fine-tuning: their broad pre-training lets them classify text from a short instruction alone, which is attractive when labelled data or compute for fine-tuning is scarce. In this setting the governing definition is either explicitly modulated within the prompt [Melis *et al.*, 2025] or omitted entirely, forcing the model to rely on its own unsteered, default perspective.

Whereas prior work has largely focused on how fine-tuning or definitional prompting steers model behaviour, we characterise model behaviour in an unguided zero-shot setting, where the model is asked to perform text classification without additional prompting information. Concretely, we investigate the following research question and three sub-questions:

> **RQ:** Which definition of hate speech does the default behaviour of large language models align with most closely?

SQ1 (Baseline consistency). To what extent can the models consistently and correctly classify hate speech instances from the Extended HateCheck dataset without explicit definitional guidance?

SQ2 (Definitional blind spots). How do the models' intrinsic behaviours align with the six modular Hate Speech Criteria, and what specific definitional blind spots does this reveal?

SQ3 (Memorisation). How does the models' behaviour compare between the Extended HateCheck dataset and structurally identical, unseen data?

To answer these questions, we probe two open-weight instruction-tuned models, Llama 3.1 and Gemma 4, using three measurements that require no fine-tuning. We first establish the models' default decision behaviour through zero-shot binary classification on the Extended HateCheck benchmark (SQ1). We then elicit structured judgements for the six HSC dimensions on the hateful subset, which exposes the definitional components that the models do and do not associate with hateful content (SQ2). Finally, because HateCheck predates both models' training cut-offs and may have been seen during pre-training, we repeat both tasks on a set of 50 novel cases written from scratch for this study, to distinguish generalised reasoning from memorisation (SQ3). Two further analyses support these measurements: a robustness check that re-runs them under prompt paraphrasing, and a definition-injection probe that tests whether the dominance failure is definitional.

This study makes three main contributions. First, it characterises the default definitions of hate speech that two widely used open-weight models apply, showing that both exhibit a degree of target awareness but systematically fail to recover the criteria of the HSC framework. Second, it provides a detailed criterion-level error analysis, identifying biases in how different components of the hate speech framework are applied in practice. Third, it presents evidence from a contamination control experiment indicating that these behavioural patterns are not attributable to benchmark memorisation.

2 Background and Related Work

Defining and formalising hate speech. A recurring observation in the literature is that hate speech datasets disagree on what they label, even when they use similar terminology. Fortuna *et al.* [2020] compare several datasets and find that labels such as *toxic*, *hateful*, *offensive*, and *abusive* are applied inconsistently, so that a model trained on one dataset does not simply transfer to the definition of another. Korre *et al.* [2025] make the scale of this variation explicit: decomposing 493 definitions from more than a hundred cultures into semantic components, they show that definitions diverge systematically along the target, the intention, and the act they emphasise. One response is to fix a single definitional position and annotate against it. Röttger *et al.* [2022] formalise this choice as a contrast between *descriptive* annotation, which surveys annotator subjectivity, and *prescriptive* annotation, which trains annotators to apply one stated definition consistently; we argue that a default, unguided model judgement is best understood as an implicit prescriptive stance whose content is unknown until probed. To make such a position explicit and modular, Khurana *et al.* [2022] introduce the Hate Speech Criteria (HSC), a framework that separates a hate speech definition into independent components, which Khurana *et al.* [2025] operationalise into the annotated criteria that this study adopts. Decomposing hate speech into graded components has precedent in the measurement literature: Sachdeva *et al.* [2022] build the Measuring Hate Speech corpus around ten facets, including an explicit *inferior/superior status* dimension that is conceptually related to the HSC notion of dominance.

Subjectivity and annotator disagreement. Whether a definition can be applied consistently at all is itself contested. Mostafazadeh Davani *et al.* [2022] argue that annotator disagreement on subjective tasks encodes meaningful variation rather than noise, and propose modelling annotators rather than collapsing them to a majority vote. This perspectivist view bears directly on our setting: when a model is asked to judge hate speech with no definition supplied, it must resolve exactly the kind of subjective ambiguity that human annotators disagree over, and the stance it defaults to is the object of our study.

Functional evaluation with HateCheck. Aggregating accuracy can mask systematic weaknesses because it summarises performance across heterogeneous inputs. HateCheck [Röttger *et al.*, 2021] addresses this by organising test cases into distinct *functionalities*, each a controlled linguistic pattern (for example, dehumanisation, implicit derogation, threats, or the use of reclaimed slurs). Because each functionality is curated, disaggregating results by functionality reveals *which* kinds of hateful or non-hateful language a model handles well. This control comes at a cost to ecological validity, since template-generated cases are cleaner and more isolated than naturally occurring hate speech; we inherit this trade-off by adopting the benchmark. We use the Extended HateCheck dataset, which augments HateCheck with HSC annotations from the DefVerify study [Khurana *et al.*, 2025], allowing the same cases to be analysed at both binary and criterion level.

Definitions in fine-tuning versus prompting. DefVerify [Khurana *et al.*, 2025] verifies whether models fine-tuned on a dataset reflect that dataset's definition, and reports that they do not encode every specified aspect; by construction, however, it speaks only to the fine-tuned regime and leaves the increasingly common zero-shot setting unexamined. A parallel line of work provides the definition through the prompt instead of through fine-tuning: Melis *et al.* [2025] show that prompting an LLM with different definitional components materially changes its zero-shot classification, and Roy *et al.* [2023] find that LLM hate-speech judgements shift with prompt formulation, for instance, when target-community information is added. What these approaches share is that the model is always given something to work with, a fine-tuned signal, a supplied definition, or an explicit framing. None of them characterises the definition a model applies when given *no* definition at all, yet that unsteered default is both the baseline against which any steering must be measured and what an LLM actually relies on in the common practice of zero-shot moderation from a bare instruction. This study addresses that gap directly, and at the level of the individual HSC criteria rather than the binary label alone.

Benchmark contamination. A known threat when probing LLMs with public benchmarks is contamination: if the benchmark appeared in pre-training, measured performance may reflect recall of seen items rather than generalisation [Golchin and Surdeanu, 2023]. Dong *et al.* [2024] frame this explicitly as a question of generalisation versus memorisation and detect contamination from the peakedness of a model's output distribution. Such detection-based methods yield only indirect evidence of overlap, however, which is why we prefer a constructive control. HateCheck was published in 2021, before the training cut-offs of both models studied here, so this concern applies, and we mitigate it by comparing benchmark performance against a structurally identical but novel test set.

3 Approach

This study uses structured prompting to elicit two kinds of judgement from the models, a binary hateful/non-hateful decision and a criterion-level characterisation against the HSC framework, and analyses both for the intrinsic definition of hate speech they reveal. Figure 1 summarises the method.

Concretely, the study runs two probing experiments and applies two cross-cutting analyses to them. The first experiment is **binary classification**, which establishes the default decision behaviour by presenting text for zero-shot classification (hateful versus non-hateful). The second is **criterion-level elicitation**, which applies the HSC framework [Khurana *et al.*, 2022] by asking the models to map each hateful instance to its six categorical dimensions, exposing the definitional components they associate with hate speech. Both experiments are then read through a **contamination control**, which repeats them on a structurally identical but strictly novel dataset to separate generalisation from memorisation [Golchin and Surdeanu, 2023], and a **disaggregated error analysis**, which breaks results down by linguistic functionality [Röttger *et al.*, 2021] and target identity and inspects the per-class and per-label confusion structure rather than treating the model as a black box. Two further checks support these analyses: a prompt-paraphrase robustness check that re-runs both experiments under meaning-preserving rewordings to confirm the findings are not artefacts of one prompt, and a targeted definition-injection probe that supplies the dominance definition to diagnose whether that criterion's failure is definitional.

All measurements are unguided in the sense relevant to the research question: no definition of hate speech is injected into the binary-classification prompt beyond a minimal task description, and the criterion-level prompt supplies only the label vocabulary, without definitions of how to apply it. The

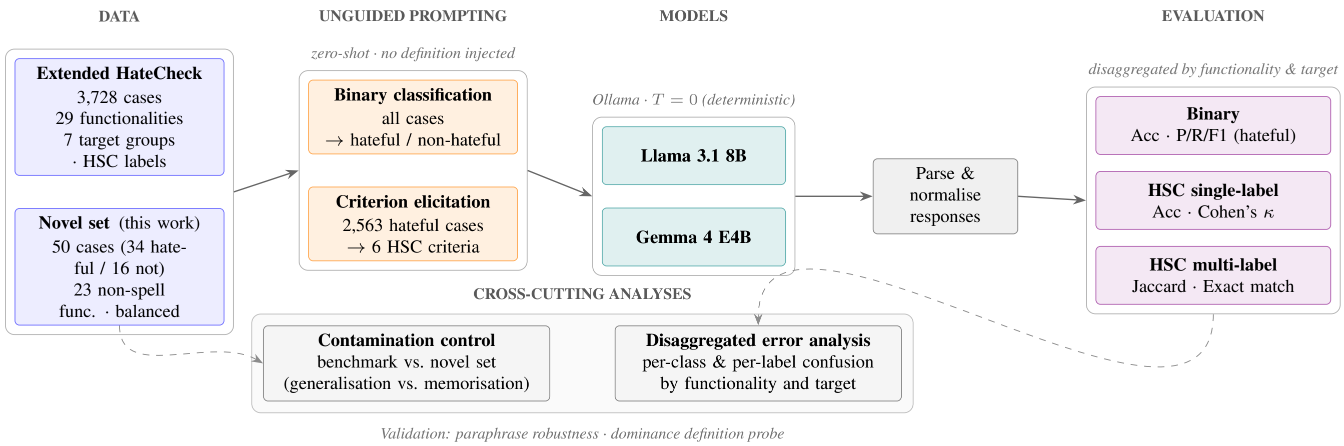

Figure 1: Overview of the method. Both instruction-tuned models are probed with two unguided, zero-shot experiments, binary classification on all cases and HSC criterion elicitation on the hateful subset, evaluated against the Extended HateCheck benchmark and a novel control set. A contamination control and a disaggregated error analysis are then applied across the resulting predictions.

aim is to elicit the model's default interpretation rather than to test how well it can follow an externally supplied definition. The one exception is the dominance definition-injection probe, which deliberately departs from this unguided stance by supplying a definition, in order to test whether that criterion's default failure is definitional; it is a diagnostic layered on top of the unguided measurements rather than part of them.

We considered several alternative designs and rejected each for a reason tied to that aim. Reading the model's label-token log-probabilities would quantify token preference but not the discrete decision a practitioner actually receives, so we elicit explicit labels instead. Training a probe classifier on hidden states would measure what is linearly decodable from internal activations rather than what the model outputs unprompted, altering the very default we set out to characterise. For contamination, detection-based methods such as output-distribution peakedness [Dong *et al.*, 2024] give only indirect evidence of overlap, so we use a constructive control, a hand-built novel set that yields metrics directly comparable to the benchmark. In each case we favoured measuring the model's own unguided output over a more instrumented but less faithful proxy.

4 Experimental Setup

This section details the datasets, models, configuration, implementation, and evaluation metrics needed to reproduce the study.

4.1 Datasets

The study uses two datasets to evaluate model behaviour and control for data contamination in LLMs.

Extended HateCheck Dataset. The primary evaluation corpus is the Extended HateCheck dataset, comprising 3,728 test cases. It merges the functional evaluation structure of the original HateCheck benchmark [Röttger *et al.*, 2021] with structured HSC annotations [Khurana *et al.*, 2025].

- **Linguistic functionalities:** test cases are organised into 29 systematic linguistic functionalities. Five of these involve character-level spelling obfuscation (for example, leetspeak and character deletion), comprising 20.4% of the dataset. Because obfuscation alters surface forms rather than semantic content, four HSC dimensions (`explicit_ref`, `incites`, `group_insult`, `in_group`) are excluded from the evaluation for these test cases; only `target_type` and `dominance` are assessed where ground-truth labels are available.

- **Target identities:** cases are distributed across seven target groups: Muslims, black people, disabled people, gay people, immigrants, trans people, and women. This enables a disaggregated analysis across multiple characteristic dimensions.

Custom novel test set. This set was created from scratch by the authors specifically for this study, to provide data that could not have appeared in either model's pre-training. Because Extended HateCheck is publicly available and predates the training cut-offs of both models, benchmark contamination is a non-trivial concern. To address it, a secondary diagnostic set of 50 novel test cases was constructed manually. Each new instance adheres to the original 19-column schema, ensuring identical parsing and evaluation through the pipeline, and was annotated with additional care for the fields central to this study (binary label, functionality, target identity, and HSC values), using values drawn from the original label sets. The five spelling-obfuscation functionalities were excluded as not central to the research question on the models' intrinsic definitions.

The resulting set comprises 50 items (34 hateful, 16 non-hateful) spanning 23 distinct non-spell functionalities, with all seven target groups represented in the hateful subset. It was also constructed to reduce the distributional imbalance present in the primary benchmark. In the hateful subset of Extended HateCheck, `group_insult` is heavily skewed (90.7% "yes" to 9.3% "no"); the novel set balances this to 64.7% "yes" against 35.3% "no". Having a more balanced distribution helps isolate the criterion alignment of models from statistical baseline effects, an aspect that is examined directly in Section 5.

4.2 Models and Configuration

Both tasks require the model to follow a short natural-language instruction and return an answer in a fixed format, so we evaluate instruction-tuned models, which are trained to adhere to such instructions. The two chosen are Meta Llama 3.1 8B (`llama3.1:latest`), a widely used open-weight baseline, and Google Gemma 4 E4B (`gemma4:latest`), a more recent, edge-optimised model. Contrasting a mature model with a newer one tests whether the observed behaviours are stable across model generations or specific to one; both run on consumer-grade hardware.

4.3 Prompts and Implementation

For the binary task, each case is given a zero-shot prompt consisting of the test text followed by a direct question asking whether it is hateful or non-hateful, with instructions to answer in exactly one word; a system instruction frames the task as academic to reduce safety refusals without revealing labels or injecting a definition, so the model's default decision is elicited. For the criterion task, restricted to the 2,563 hateful instances, the model is shown the label vocabulary for the six criteria (with "none" omitted) and returns, for each criterion, a single label for the four single-label criteria and any applicable labels for the two multi-label ones, again with no definitions, so the judgement stays unguided. Both models are served locally through Ollama at temperature 0, making generation deterministic for fixed weights. The complete conventions for prompt wording, response parsing, refusal handling, checkpointing, and label-normalisation are given in Appendix A.1. The follow-up analyses use two further prompt variants, both documented there: two meaning-preserving paraphrases of each prompt for the robustness check, and a version of the dominance prompt with the dominance definition and operational context appended for the definition-injection probe.

4.4 Evaluation Metrics

Binary classification is evaluated with accuracy and with precision, recall, and F1 for the *hateful* (positive) class, together with the proportion of unparseable responses (`unknown_rate`) and `mean_retries`. Unparseable responses are binarised to *non-hateful* before computing the sklearn metrics and are also tracked separately. For the non-hateful functionalities, which contain no positive (hateful) instances, precision, recall, and F1 are reported as zero by the `zero_division=0` convention; for these rows accuracy is the only informative metric.

For the HSC criteria, the four single-label columns (`target_type`, `dominance`, `group_insult`, `in_group`) are evaluated with accuracy and Cohen's κ. Accuracy reports the raw proportion of correct labels but can be inflated by a skewed class distribution, so we pair it with κ, which measures agreement above what chance would give and therefore exposes cases where a high accuracy merely reflects predicting the majority class. When a disaggregated group contains gold labels of only one class, κ is undefined and is reported as missing rather than as a value. The two multi-label columns (`explicit_ref`, `incites`) are evaluated with the mean Jaccard similarity, which credits partial overlap between the predicted and gold label sets, and the exact match ratio, which is stricter and counts only predictions whose label set equals the gold set exactly; reporting both separates partial from complete correctness. All HSC metrics are computed globally and disaggregated by functionality and by target identity. Row-level refusal statistics and per-column missing and unknown rates are reported as diagnostics.

Table 1: Overall zero-shot binary classification on Extended HateCheck ($n = 3{,}728$). Precision, recall, and F1 are computed for the hateful (positive) class.

Model	Accuracy	Precision	Recall	F1
Gemma 4	0.947	0.953	0.971	0.962
Llama 3.1	0.920	0.918	0.970	0.943

5 Results

5.1 SQ1: Baseline Binary Classification

Table 1 reports overall binary-classification performance on Extended HateCheck. Both models classify the full set with high accuracy, Gemma 4 at 0.947 and Llama 3.1 at 0.920, well above the 0.50 chance line, with high recall for hateful content (0.971 and 0.970 respectively). Neither model produced any unparseable binary responses (`unknown_rate` $= 0$) or required refusal retries on this task once the system prompt was supplied (Appendix Figure 4).

Errors concentrate on non-hateful look-alikes. On the hateful and spell-obfuscated functionalities both models sit at or near ceiling and vary only at the second decimal, so we do not dwell on them (exact values in Appendix Table 10); the lowest hateful-row scores are simply the least overt signals, implicit derogation (Gemma 4 0.900) and emotional derogation (Llama 3.1 0.914), and Llama 3.1 is marginally weaker than Gemma 4 on the spell-obfuscated rows. Because that hateful-side variation is uniform and small, the informative differences lie on the non-hateful functionalities (Figure 2).

The weaknesses lie almost entirely in non-hateful cases that superficially resemble hate speech. Reclaimed slurs are the hardest category for both models (Gemma 4 0.519, Llama 3.1 0.704), and counter-speech that quotes or references hate is error-prone, especially for Llama 3.1 (counter-speech via quote 0.769, via reference 0.504). For Llama 3.1, non-hateful statements that merely name a group or an individual as a target are frequently misread as hateful (group target 0.500, individual target 0.369), whereas Gemma 4 handles these better (0.790 and 0.892). Because these functionalities contain no positive instances, their precision and recall are reported as zero by convention and accuracy is the informative metric. The pattern indicates that both models lean toward labelling group-referential or slur-containing text as hateful, which raises recall on genuinely hateful content but lowers specificity on non-hateful look-alikes.

Performance by target group is broadly even. Accuracy is broadly even across the seven target groups (mostly 0.87–0.98 for both models; Appendix Figure 5, Appendix Table 11), with no group near chance, so unlike the criterion level below there is no large target-specific failure at the binary level. The only notable gaps are the untargeted cases, hardest for Llama 3.1

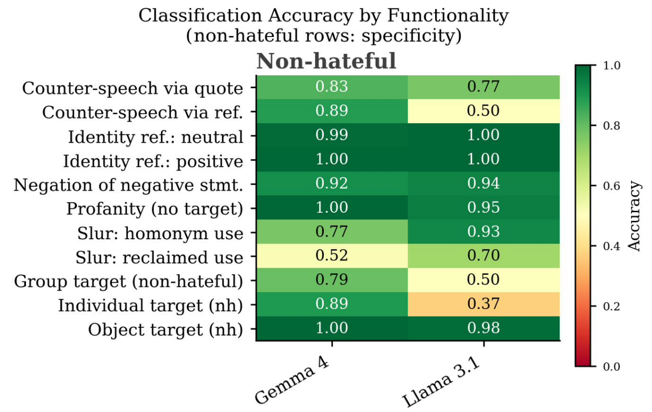

Figure 2: Binary classification accuracy on the non-hateful functionalities (specificity). The hateful and spell-obfuscated functionalities are near-ceiling and are omitted here (exact values in Appendix Table 10). Both models degrade sharply on non-hateful look-alikes, reclaimed slurs, counter-speech, and bare target references, where Llama 3.1 is the weaker of the two.

Table 2: Overall HSC criterion performance on the hateful subset. Single-label columns: accuracy and Cohen's κ. Multi-label columns: mean Jaccard and exact match (EM). $n = 2{,}563$ for `target_type/dominance`; $n = 1{,}803$ otherwise.

Criterion	Metric	Gemma 4	Llama 3.1
`target_type`	Accuracy	**0.981**	0.897
`target_type`	Cohen's κ	**0.977**	0.875
`dominance`	Accuracy	0.036	0.003
`dominance`	Cohen's κ	0.000	0.000
`explicit_ref`	Jaccard	**0.596**	0.179
`explicit_ref`	EM	**0.370**	0.058
`incites`	Jaccard	0.620	**0.672**
`incites`	EM	0.312	**0.529**
`group_insult`	Accuracy	**0.892**	0.498
`group_insult`	Cohen's κ	-0.017	**0.060**
`in_group`	Accuracy	0.784	**0.999**
`in_group`	Cohen's κ	0.000	0.000

(0.733 versus 0.932), and women, where Llama 3.1 leads Gemma 4 (0.929 versus 0.874).

5.2 SQ2: Alignment with the Hate Speech Criteria

The six criteria fall into clearly different regimes (Table 2; per-class F1 in Figure 3; Appendix Figure 9): `target_type` sits near the top for both models, `dominance` collapses to the floor, and the remaining criteria fall in between with model-dependent gaps. Compliance on this harder, structured task remained high: neither model refused any row (row-refusal rate 0), Gemma 4 left a field empty on only 0.67% of `explicit_ref` rows and 0.44% of `incites` rows (zero elsewhere), and the only non-trivial retry load was Llama 3.1's mean of 0.0765 refusal retries per row against Gemma 4's 0 (Appendix Figure 6). Missing predictions are therefore too rare to affect the metrics below.

Target type is identified reliably. Both models identify which characteristic is targeted with high agreement (Gemma 4

Table 3: HSC `target_type` per-class performance for Llama 3.1 on the hateful subset. Its weakness is localised to gender (low recall) and sexual orientation (low precision), the two sides of a single directional confusion. Gemma 4 is uniform across classes (Appendix Table 8).

Class	n	P	R	F1
disability	373	1.000	0.944	0.971
gender	730	0.949	**0.738**	0.831
nationality	357	0.972	0.972	0.972
race	357	0.908	0.992	0.948
religion	373	1.000	0.941	0.970
sexual orientation	373	**0.653**	0.952	0.774

$\kappa = 0.977$, accuracy 0.981; Llama 3.1 $\kappa = 0.875$, accuracy 0.897; Table 2). Gemma 4 is uniformly strong, with every class above 0.96 F1 (Appendix Table 8). Llama 3.1's per-class breakdown (Table 3) shows it matches Gemma 4 on four of the six classes but is weaker on gender and sexual orientation, and the precision/recall split locates the cause: its gender *recall* is low (0.738) while its sexual-orientation *precision* is low (0.653), the signature of a directional confusion in which gender-targeted text is assigned to the sexual-orientation class (confusion matrix in Appendix Table 9). By target group the error falls almost entirely on trans people, where Llama 3.1 reaches only 0.471 accuracy against Gemma 4's 0.975 (Appendix Table 12).

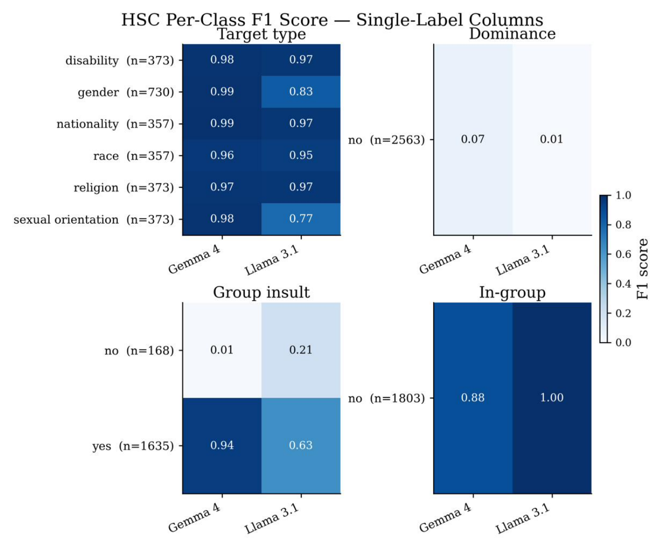

Figure 3: Per-class F1 for the four single-label HSC columns. `target_type` is uniformly high for Gemma 4 but drops for Llama 3.1 on gender (0.83) and sexual orientation (0.77). `dominance` is near zero for the single "no" gold class. `group_insult` shows Gemma 4 scoring 0.01 on the minority "no" class against 0.94 on "yes", the signature of unconditional "yes" prediction. `in_group` is high because the gold class is uniformly "no".

Dominance fails almost completely for both models. The clearest definitional blind spot is `dominance`. Both models predict "yes" for nearly every hateful instance, with accuracies of only 0.036 (Gemma 4) and 0.003 (Llama 3.1) and

$\kappa = 0.000$, i.e. chance-level; the row-normalised confusion is in Table 4. The failure is uniform across functionalities and target groups (Appendix Figures 7 and 8). In the HSC framework, `dominance` records the historical power position of the *targeted* group: "yes" when the hateful speech is aimed at a socially dominant group (such as men or white people, a category added by DefVerify) and "no" when it is aimed at a marginalised one. It is a property of where the target group stands in the social hierarchy, not a comparison of power between the speaker and the target within the utterance. Because all seven target groups in HateCheck are marginalised, every one of the 2,563 hateful gold labels is "no". A model that simply output "no" would therefore be perfectly correct; both models instead do the opposite. Because the prompt supplies only the bare `yes/no` vocabulary with no definition, the most plausible reading is that the models apply the everyday sense of "dominance" (domination expressed in the text) rather than the framework's demographic sense, so the near-constant "yes" indicates that they are not assessing the target group's social position at all; we develop this interpretation in Section 6. To test this directly, we re-ran the dominance criterion with a definition of dominance inserted into the prompt, together with an operational note specifying that dominance is a structural property of the targeted group rather than of the speaker or the tone of the utterance (full wording in Appendix A.1). Supplying the definition changes the behaviour almost completely: accuracy rises from 0.036 to 0.9996 for Gemma 4 and from 0.003 to 0.970 for Llama 3.1, as both models flip from labelling almost every item "yes" to almost every item "no", the correct answer for this all-marginalised set. Because every gold label here is "no", this measures the models' propensity to give the correct label rather than a genuine ability to tell dominant from marginalised targets, for which the benchmark contains no positive cases; what it does show is that the default failure is largely a definitional gap, and specifically that the disambiguation that fixes it is the one separating the target group's status from the speaker's tone.

Table 4: Row-normalised confusion for the three single-label binary HSC criteria (hateful subset). Each cell is the percentage of the corresponding gold class (N = "no", Y = "yes"). Gold-class sizes: `dominance` 2563 N / 0 Y; `group_insult` 168 N / 1635 Y; `in_group` 1803 N / 0 Y. `dominance` and `in_group` have no gold "yes" class. Both models label almost all `dominance` items "yes"; on `group_insult` Gemma 4 predicts "yes" almost unconditionally while Llama 3.1 under-predicts it; on `in_group` Gemma 4 produces 21.6% false "yes".

		gold N		gold Y	
Criterion	Model	pred N	pred Y	pred N	pred Y
`dominance`	Gemma 4	3.6	96.4	–	–
	Llama 3.1	0.3	99.7	–	–
`group_insult`	Gemma 4	0.6	99.4	1.7	98.3
	Llama 3.1	71.4	28.6	52.4	47.6
`in_group`	Gemma 4	78.4	21.6	–	–
	Llama 3.1	99.9	0.1	–	–

Reference and incitement labels are over-applied. On the multi-label criteria the two models diverge sharply on `explicit_ref` and converge on `incites` (Appendix Figure 9, right panel; per-label confusion rates in Table 5). For `explicit_ref`, Gemma 4 (Jaccard 0.596, exact match 0.370) substantially outperforms Llama 3.1 (0.179, 0.058), and the false-positive rates make the mechanism plain: Llama 3.1 applies *slur* and *stereotype* to almost every instance while missing the most prevalent label, *group_characteristic* (recall ≈ 0.17), whereas Gemma 4 detects *group_characteristic* well (recall ≈ 0.80) and is well-calibrated on *slur* but still over-predicts *stereotype*.

For `incites`, the two models are comparable: Llama 3.1 scores marginally higher on the original prompt (Jaccard 0.672 vs 0.620, exact match 0.529 vs 0.312), but this edge does not survive prompt paraphrasing (Section 5.4) and is best read as a tie. The remaining structure is again a calibration effect: Llama 3.1 applies *hate* to almost all instances, which coincides with the most common gold set {hate} and inflates its exact match, but it under-detects *violence* (recall ≈ 0.63) where Gemma 4 recovers more (≈ 0.85), and Gemma 4 in turn over-applies *discrimination* (Table 5). Llama 3.1's violence misses concentrate on the functionalities that require recovering an implied or inverted meaning, such as negated-positive statements and normalised threats (Appendix Table 14).

Table 5: Per-label outcomes for the two multi-label HSC criteria, expressed as a percentage of all non-spell hateful rows ($n = 1{,}803$). TP/FP/FN are true positives, false positives, and false negatives for each label. Large FP% marks over-applied labels (Llama 3.1: *slur* 75.3%, *stereotype* 72.0%; both models over-apply *discrimination* and *hate*); large FN% marks under-detected labels (Llama 3.1 misses *group_characteristic* on 70.5% of rows).

		Gemma 4			Llama 3.1		
Criterion	Label	TP	FP	FN	TP	FP	FN
`explicit_ref`	group char.	68.4	7.8	17.0	14.9	1.9	70.5
	slur	6.8	3.2	1.2	7.8	75.3	0.2
	stereotype	10.8	54.2	0.8	10.3	72.0	1.3
`incites`	discrim.	12.3	35.1	3.3	5.3	8.4	10.4
	hate	57.2	33.0	1.3	58.3	41.4	0.1
	violence	24.8	2.5	4.3	18.3	1.1	10.8

Group insult and in-group reflect opposite prediction biases. The single-label criteria `group_insult` and `in_group` expose mirror-image biases; the confusion counts for both, together with `dominance`, are collected in Table 4. On `group_insult`, Gemma 4 reaches 0.892 accuracy but $\kappa = -0.017$: it predicts "yes" almost unconditionally (only 1 of 168 true "no" cases recovered), so its accuracy is explained by the 90.7% "yes" class prevalence rather than by discrimination. The F1 for the *yes* and *no* classes separately reflects this directly (Figure 3): 0.94 on "yes" but 0.01 on "no". Llama 3.1 has the opposite bias: it predicts "no" for 857 of the 1,635 "yes" instances, which lowers its accuracy to 0.498 but yields a slightly positive κ (0.060) and a higher "no"-class F1 (0.21 vs Gemma 4's 0.01) because it recovers 120 of the 168 true "no" cases. By target group, Llama 3.1's group-insult accuracy varies strongly (Appendix Table 13).

On `in_group`, where every one of the 1,803 gold labels is "no", Llama 3.1's conservative tendency to predict "no" for binary criteria yields near-perfect accuracy (0.999, a single false positive), whereas Gemma 4 labels 389 instances (21.6%) as in-group speech (Table 4). Gemma 4's false positives concentrate on the targets and functionalities where speaker identity is least explicit (Appendix Tables 13 and 14). κ is 0.000 for both models because the gold labels are single-class, so accuracy is the only usable metric here. Taken together, these two criteria are best read as a calibration difference: Gemma 4 defaults toward "yes" on binary criteria and Llama 3.1 toward "no", and each default happens to align with the gold majority on exactly one of the two criteria.

5.3 SQ3: Memorisation Control

Table 6: Binary classification: Extended HateCheck (HC) versus novel set.

	Gemma 4		Llama 3.1	
Metric	HC	Novel	HC	Novel
Accuracy	0.947	0.920	0.920	0.920
Precision	0.953	1.000	0.918	1.000
Recall	0.971	0.882	0.970	0.882
F1	0.962	0.938	0.943	0.938

On the binary task neither model drops enough to indicate memorisation (Table 6): both score 0.920 accuracy on the novel set, and their few errors fall on the same indirect, opinion-framed, or negated constructions that are hard on the full benchmark, so difficulty tracks linguistic properties rather than item familiarity.

Table 7: HSC criteria: Extended HateCheck (HC) versus novel set ($n = 34$ for all novel columns). EM = exact match. $^{\dagger}\kappa$ undefined (single-class gold).

		Gemma 4		Llama 3.1	
Criterion	Metric	HC	Novel	HC	Novel
`target_type`	Acc	0.981	1.000	0.897	1.000
	κ	0.977	1.000	0.875	1.000
`dominance`	Acc	0.036	0.147	0.003	0.000
	κ	0.000	0.000	0.000	0.000
`explicit_ref`	Jacc	0.596	0.701	0.179	0.353
	EM	0.370	0.500	0.058	0.000
`incites`	Jacc	0.620	0.701	0.672	0.755
	EM	0.312	0.441	0.529	0.529
`group_insult`	Acc	0.892	0.647	0.498	0.559
	κ	−0.017	0.000	0.060	0.147
`in_group`	Acc	0.784	0.647	0.999	1.000
	κ	0.000	0.000	0.000	–†

The criterion task tells the same story (Table 7). `target_type` transfers perfectly (both models 1.000 accuracy, $\kappa = 1.000$), though the 34-item set is too small to contain the gender/sexual-orientation cases behind Llama 3.1's benchmark confusion, so this does not show that confusion is absent on unseen data. The two diagnostic failures persist: `dominance` stays at the floor (both still answer "yes" almost everywhere, as the novel items also target only marginalised groups), and Gemma 4's `group_insult` accuracy drops to 0.647 ($\kappa = 0.000$) once the class balance is corrected, confirming that its 0.892 benchmark figure was an imbalance artefact rather than skill. The `explicit_ref` and `incites` scores rise on the novel set, but this most likely reflects its label distribution happening to favour each model's existing prediction bias rather than a genuine gain in capability; the novel set is released so that this can be inspected directly.

5.4 Robustness to Prompt Paraphrasing

To test whether the criterion-level and binary findings depend on prompt wording, we re-ran both experiments under the original prompt and two meaning-preserving paraphrases on a stratified-by-functionality subset (full protocol in Appendix A.1), treating a metric as wording-sensitive when its range across the three phrasings exceeds 0.05 (full figures in Appendix Tables 15 and 16).

The qualitative findings of SQ1 and SQ2 are stable. On the binary task, accuracy varies by less than 0.05 for both models; the only wording-sensitive metric is precision, which falls under both paraphrases while recall rises, consistent with the over-inclusive default described above rather than in tension with it. On the criterion task, `target_type` is near-ceiling under every phrasing (range ≤ 0.014) and the `dominance` failure is entirely insensitive to wording (Gemma 4 range 0.017; Llama 3.1 fixed at 0.000), so the study's central result is not a prompt artefact. Gemma 4's `explicit_ref` advantage over Llama 3.1 also holds throughout.

Two cells are genuinely wording-sensitive, both for Llama 3.1. Its `incites` Jaccard falls from 0.656 on the original prompt to about 0.44 on both paraphrases, so the slight `incites` edge over Gemma 4 reported above does not survive paraphrasing and should be read as "comparable". Its `group_insult` accuracy moves in the opposite direction, from 0.546 to about 0.74, so the magnitude of Llama 3.1's `group_insult` deficit is partly a wording effect, although the mirror-image calibration pattern (Gemma 4 over-predicting "yes", Llama 3.1 under-predicting it) still holds on the original prompt. In both cases the original prompt sits at one end of the observed range, which is why we report these two cells with their across-phrasing spread in mind.

6 Discussion

Reading the three experiments together shows which facets of hate speech the default conception prioritises and which it ignores. Both models prioritise surface-level markers. They reliably identify the targeted group, which is a capability that the contamination control (SQ3) confirms is generalised rather than memorised, but they systematically over-apply labels based on explicit cues like slurs and hostile expressions. This reliance on surface signals links the two findings, causing the misclassification of non-hateful look-alikes (such as counter-speech and reclaimed slurs) in the binary setting, and the collapse of fine-grained incitement and reference distinctions in the criterion setting, particularly for Llama 3.1.

Conversely, the default conception ignores structural context, most prominently the `dominance` criterion. The framework defines this as the target group's position in the social hierarchy; because the benchmark targets only marginalised groups, every gold label is "no". Instead, both models predict "yes" almost universally. This indicates that they are not making random errors but are misinterpreting the term in its colloquial sense, reading domination as a theme of the hateful text rather than applying the framework's technical definition. A definition-injection probe confirms this reading: supplying a definition that ties dominance to the target group's social position rather than the speaker's tone raises accuracy from near-zero to near-perfect (Section 5). Because every gold label on this criterion is "no", that jump reflects the removal of a definitional gap rather than proof that the models can identify a dominant target, but it does locate the default failure in missing definitional context. This near-total failure perfectly illustrates why decomposed probing is essential: while binary accuracies of 0.92–0.95 suggest a robust understanding of hate speech, criterion-level analysis reveals that this unguided grasp fundamentally misses critical definitional components.

These findings connect to previous work in a natural way. DefVerify [Khurana *et al.*, 2025] reports that models fine-tuned on hate speech datasets do not encode every aspect of their dataset's definition; our results show an analogous gap for the *unguided* zero-shot setting, with specific dimensions (dominance, and the finer reference distinctions) being the ones it fails to apply. Melis *et al.* [2025] show that supplying definitional components through the prompt changes zero-shot classification outcomes; the blind spots we identify suggest concrete targets for such definitional prompting, since the failures concentrate on dimensions where a brief contrastive definition could plausibly realign the model. The over-inclusive labelling we observe is also coherent with the broader observation that hate speech datasets and models encode different "flavours" of the phenomenon [Fortuna *et al.*, 2020]: a model that defaults to surface markers will align with definitions that emphasise those markers and diverge from definitions that require relational judgements.

The two models differ mainly in calibration rather than in overall understanding: Gemma 4 follows the structured format more faithfully and defaults toward "yes" on the binary criteria (inflating `group_insult` under class imbalance and the in-group false-positive rate), whereas Llama 3.1 is more conservative (helping `in_group`, hurting `group_insult`) and carries the gender/sexual-orientation confusion that mislabels anti-trans content; the size of the two models' gap on `incites` and `group_insult` is itself sensitive to prompt wording (Section 5.4), so we do not over-read it. That both nonetheless share the dominance failure and the over-inclusive reference and incitement biases suggests these are not idiosyncratic to one model.

6.1 Responsible Research

This research analyses hate speech but does not generate new hateful content: all test cases come from the established HateCheck benchmark and a small manually written extension, and the models are asked only to classify or characterise existing text. The work is intended to make the limitations of unguided LLM judgements visible, which supports more cautious deployment; the results in fact argue *against* treating zero-shot LLM labels as reliable for fine-grained hate speech characterisation without further safeguards, particularly given the systematic mislabelling of anti-trans content by one model, which could cause real harm if used unexamined in a moderation pipeline. We report the dominance and group-insult findings as definitional mismatches rather than as the models being "broken", and we note that the gold standard reflects a particular definition rather than a ground truth that holds across all communities. Because the gold labels and the manually constructed novel set embody one annotation scheme and the authors' own judgements, every characterisation we report is relative to that stance; we make no claim that it is the only defensible labelling, and we release the data so that others can re-annotate against a different definition. Documenting which constructions evade detection also carries a dual-use risk, since the same failure modes could in principle inform attempts to evade automated moderation. We judge this risk low: every functionality we analyse is already public in HateCheck, the findings concern two specific small models rather than any deployed system, and the practical message is precisely that such models should not be relied upon unexamined.

For reproducibility, the study fixes the model weights by serving both models locally, uses deterministic generation (temperature 0) for both reported experiments, records every model output to per-run checkpoints, and re-reads gold labels only at analysis time so that evaluation logic is applied in one place. The datasets, prompts, generation parameters, parsing rules, and metric definitions are documented in Section 4, which together with the released code should allow the experiments to be repeated. Two caveats remain: the exact build of each model behind the `:latest` Ollama tag should be pinned to a fixed version for an exact replication, and the manually constructed novel set reflects the authors' annotation judgements and would benefit from independent re-annotation.

We close by assessing the main threats to the study's validity along three standard axes.

Internal validity. The pipeline relies on automated prompting and parsing. One error was found and fixed during analysis: a label-string mismatch for the *sexual orientation* class initially produced zero recall, and after correction the residual gender/sexual-orientation confusion was confirmed to be a model property; the deterministic parsers and the separation of gold labels from collection limit silent errors, but others cannot be ruled out. Each criterion was elicited with a single fixed prompt at temperature 0, so the figures characterise behaviour under this elicitation rather than an upper bound, and stating in the HSC prompt that the text is hateful, though necessary to keep the task on characterisation, may itself contribute to the over-inclusive labelling.

External validity. We evaluate two open-weight models in the 4–8B range on one benchmark family (HateCheck and its HSC extension) in English, so the findings characterise these models on this benchmark rather than LLMs in general; larger models, other benchmarks, and other languages may differ. The novel set, though unseen, has only 50 items and serves as a contamination check rather than a precise second measurement.

Construct validity. Several measurement choices limit what the metrics show. For `dominance` and `in_group` the gold is single-class, so κ is undefined or trivially zero and only accuracy is interpretable; the dominance result is meaningful precisely because the gold is uniform and the predictions uniformly wrong. Since the prompt supplies the label but no definition, that result should be read as a property of the *unguided* setting, the dimension most likely to move under definitional prompting, rather than a ceiling. Exact match is strict and is therefore reported alongside Jaccard. Finally, the HSC gold reflects one annotation scheme and one definitional stance, so a prediction judged wrong on `dominance` is wrong relative to that scheme.

7 Conclusion

This paper asked which definition of hate speech the default behaviour of large language models aligns with most closely, and probed two open-weight models, Llama 3.1 and Gemma 4, without supplying any definition. The answer can be stated as a definition. By default, these models treat text as hateful when it directs hostile surface markers (slurs, group references, expressions of hate) at an identifiable group, and which group is targeted is the one definitional component they represent most reliably. What the default omits is equally specific: it is indifferent to the target group's position in the social hierarchy (`dominance`), it does not track the speaker's identity (`in_group`) or whether the group as a whole is insulted (`group_insult`), answering these from a model-level yes/no prediction prior rather than from the relevant cue, and it collapses the finer distinctions between reference mechanisms and kinds of incitement toward the most salient labels. The default definition is therefore target-aware but surface-oriented, and over-inclusive at its boundary.

The three sub-questions establish this answer step by step. SQ1 showed a stable, consistently applied default exists at all, with errors concentrated on non-hateful look-alikes rather than random noise. SQ2 located that default's content across the six HSC criteria: reliable target identification (with Llama 3.1's gender/sexual-orientation exception), near-total failure on `dominance` that a definition-injection probe traces to a definitional rather than a deeper gap, and over-application of reference and incitement labels. SQ3 confirmed this default is generalised rather than memorised, replicating the same skills and the same blind spots on unseen data. Read together, the three experiments show that the default prioritises the target and surface markers of hate, and ignores the target group's social position and, for Llama 3.1 specifically, the finer reference and incitement distinctions.

Taken together, these results indicate that, at least for the two open-weight models studied here, an unguided zero-shot LLM should not be assumed to apply a neutral or complete definition of hate speech: each applies a particular, surface-oriented one with measurable gaps, and that characterisation is itself relative to the HSC framework used to measure it. Whether the same holds for larger or differently trained models is an open question that this study cannot settle. This has a practical consequence for the larger project of which this study is the first part. Because the default conception diverges from the HSC framework along specific, identifiable dimensions, those dimensions are natural targets for definitional prompting, and the figures reported here provide the baseline against which any such steering should be measured.

7.1 Future Work

Several directions follow. Because a structural definition already closes the dominance gap almost entirely, the open question is whether a single shared definition can repair the weaker criteria at once without harming the ones that already work, and whether the gender/sexual-orientation confusion is similarly addressable through prompting. Chain-of-thought or decomposed prompting, in which each criterion is elicited separately and the relevant evidence is named before a label is assigned, may reduce the over-inclusive labelling on the multi-label criteria. Extending the evaluation to larger and more diverse models would show whether the shared failures observed here are general properties of instruction-tuned LLMs or particular to the 4–8B range, and extending it to other languages and annotation schemes would test how far the target-centred, dominance-blind default generalises. A further direction is a larger novel control set annotated by multiple independent raters with reported inter-annotator agreement, which the present 50-item set, constructed by the authors alone under the project's time and resource constraints, does not provide.

References

[Dong *et al.*, 2024] Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050. Association for Computational Linguistics, 2024.

[Fortuna *et al.*, 2020] Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France, 2020. European Language Resources Association.

[Golchin and Surdeanu, 2023] Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023.

[Khurana *et al.*, 2022] Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid), 2022. Association for Computational Linguistics.

[Khurana *et al.*, 2025] Urja Khurana, Eric Nalisnick, and Antske Fokkens. DefVerify: Do hate speech models reflect their dataset's definition? In *Proceedings of the 31st*

International Conference on Computational Linguistics, pages 4341–4358, Abu Dhabi, UAE, 2025. Association for Computational Linguistics.

[Korre *et al.*, 2025] Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198. Association for Computational Linguistics, 2025.

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

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

[Röttger *et al.*, 2021] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online, 2021. Association for Computational Linguistics.

[Röttger *et al.*, 2022] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 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. Association for Computational Linguistics, 2022.

[Roy *et al.*, 2023] Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. Probing LLMs for hate speech detection: Strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128. Association for Computational Linguistics, 2023.

[Sachdeva *et al.*, 2022] Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. European Language Resources Association, 2022.

A Appendices

A.1 Implementation and Prompt Details

Tooling and checkpointing. The pipeline is implemented in Python as a modular command-line application, using `pandas` for data manipulation and `scikit-learn` for metric computation. Each model output is written immediately to a per-run, per-model checkpoint file, so an interrupted run resumes without redundant model calls, and run identifiers separate experimental variations. Gold labels are stored in the checkpoints for manual inspection but are re-read from the source dataset and joined on the row index only at analysis time, so gold-label normalisation is applied in one place, independently of data collection. The maximum number of generated tokens was capped at 4096; as binary prompts expect one word and HSC prompts six numbered lines, typical responses are much shorter.

Label normalisation. The dataset uses blank cells for several HSC columns following Khurana *et al.* [2025]: for `target_type`, `dominance`, `explicit_ref`, `incites`, and `group_insult` a blank denotes "none", while for `in_group` a blank denotes "no". This is applied at analysis time.

Binary-classification prompt. Each case is presented individually with a zero-shot prompt consisting of the test text followed by the question "Is this text hateful or non-hateful? Reply with exactly one word: hateful or non-hateful." The response is parsed with a deterministic regular expression (a `Label: X` pattern) and, failing that, substring matching in which "non-hateful" is checked before "hateful" to avoid prefix collision; if neither matches, the response is recorded as `unknown`.

HSC-elicitation prompt. The criterion task is restricted to the 2,563 hateful instances because the HSC criteria characterise *how* speech is hateful rather than *whether* it is. Each case is presented with a single structured prompt stating that the text is hateful and asking the model to select, for each of the six criteria, from a supplied label vocabulary, returning six numbered lines. "none" is omitted from every option; if a model nevertheless outputs it, the prediction is scored incorrect. Supplying the vocabulary but no definitions avoids conflating conceptual understanding with knowledge of the label terms while keeping the judgement unguided.

Refusal handling. Each response is checked for refusal markers before parsing; on a refusal the request is re-issued with a fallback prompt up to five times, and the retry count is reported as `mean_retries`. Markers that also occur in valid analyses of hateful content (for example, "hate speech" or "promote violence") are excluded to avoid false positives. The academic-framing system prompt is used because a control showed it materially reduces refusals: without it, Llama 3.1 produced an unknown-response rate of 0.0019 and a mean of 0.19 refusal-triggered retries per item, both of which fell to zero with it (Appendix Figure 4), without injecting a definition or revealing labels.

Dominance definition-injection probe. For the diagnostic probe reported in Section 5, the following definition and operational context were appended to the dominance prompt:

> **Definition of dominance:** We define a dominant (cultural) group as a group whose members are (possibly without them being aware) positively privileged, unstigmatized, and generally favored by societal institutions.
>
> **Operational context:** Dominance is a structural, societal characteristic of the targeted identity group itself. Dominance is NOT a description of the speaker's tone, aggression, or behaviour within the text. Even if a speaker

sounds highly hostile, aggressive, or dominating, the target group itself may still be socially non-dominant (marginalized).

Prompt-paraphrase robustness. The robustness check reported in Section 5.4 re-runs each task under two meaning-preserving paraphrases of the instruction text on a stratified-by-functionality subset, holding the label vocabulary, system framing, and output format fixed so that only the wording varies. The binary subset samples evenly across all 29 functionalities (target 20 items per functionality, capped at the available count); the HSC subset samples from the hateful subset evenly across functionalities, targeting approximately 300 rows. The same sampled item IDs, fixed by a random seed, are reused for the original prompt and both paraphrases, so that all differences between phrasings are attributable to wording rather than to sampling.

A.2 Figures

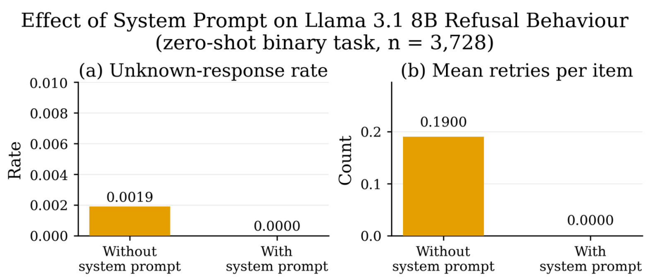

Figure 4: Effect of the academic-framing system prompt on Llama 3.1 refusal behaviour in the zero-shot binary task ($n = 3{,}728$). The unknown-response rate and the mean number of refusal-triggered retries per item both fall to zero once the system prompt is supplied.

A.3 Tables

Table 8: HSC `target_type` per-class performance for Gemma 4 on the hateful subset. Gemma 4 is uniformly strong, with every class above 0.96 F1. The Llama 3.1 counterpart is in the main text (Table 3).

Class	n	P	R	F1
disability	373	1.000	0.965	0.982
gender	730	0.990	0.985	0.988
nationality	357	0.992	0.986	0.989
race	357	0.932	1.000	0.965
religion	373	0.992	0.957	0.974
sexual orientation	373	0.974	0.989	0.981

Table 9: Llama 3.1 `target_type` confusion matrix (gold rows, predicted columns; hateful subset, $n = 2{,}563$). The off-diagonal mass is concentrated in one cell: 185 of the 730 gender-targeted items are predicted *sexual orientation*. Abbreviations: dis. disability, gen. gender, nat. nationality, rel. religion, s.o. sexual orientation. Gemma 4's matrix is near-diagonal (largest off-diagonal cell 13) and is omitted.

Gold / Pred.	dis.	gen.	nat.	none	race	rel.	s.o.
disability	352	5	1	0	15	0	0
gender	0	539	0	1	5	0	**185**
nationality	0	5	347	0	5	0	0
race	0	0	3	0	354	0	0
religion	0	8	6	0	4	351	4
s.o.	0	11	0	0	7	0	355

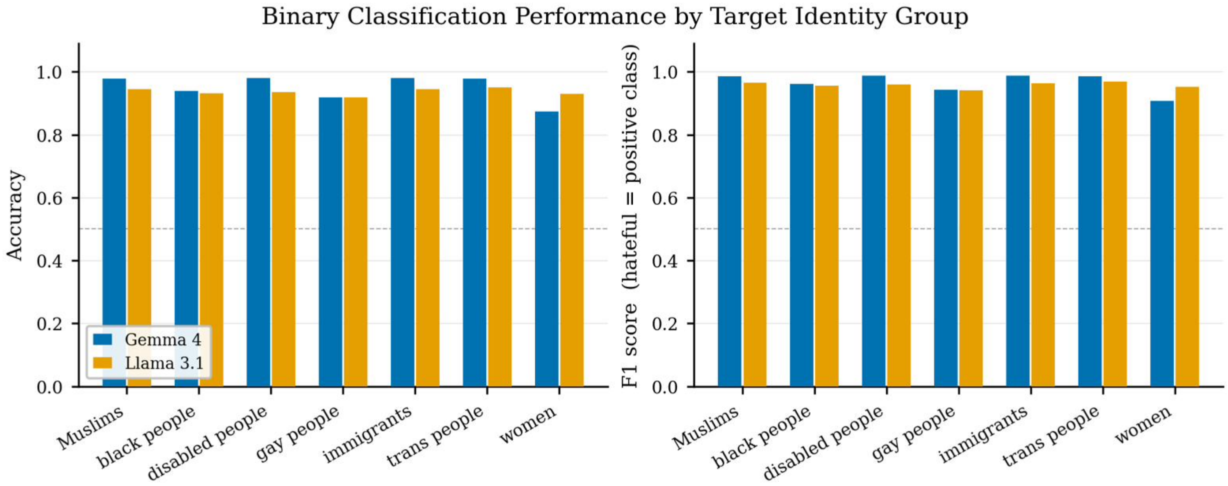

Figure 5: Binary classification accuracy (left) and macro F1 (right) by target identity group. Performance is broadly even across groups; the largest model difference is on women, where Llama 3.1 is more sensitive than Gemma 4. F1 here is macro-averaged over the hateful and non-hateful classes within each group, unlike the overall binary table (Table 1), which reports precision, recall, and F1 for the hateful class only; the two are not directly comparable.

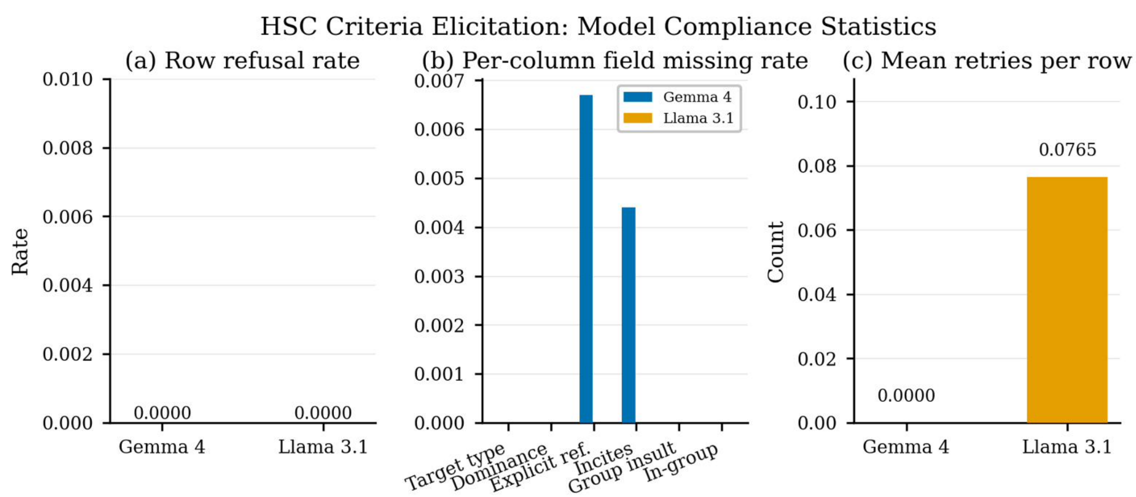

Figure 6: Compliance on the HSC elicitation task. Neither model refused any row; field-missing rates are below 0.7% and confined to the two multi-label columns for Gemma 4; the only non-trivial retry load is Llama 3.1's 0.0765 mean refusal retries per row.

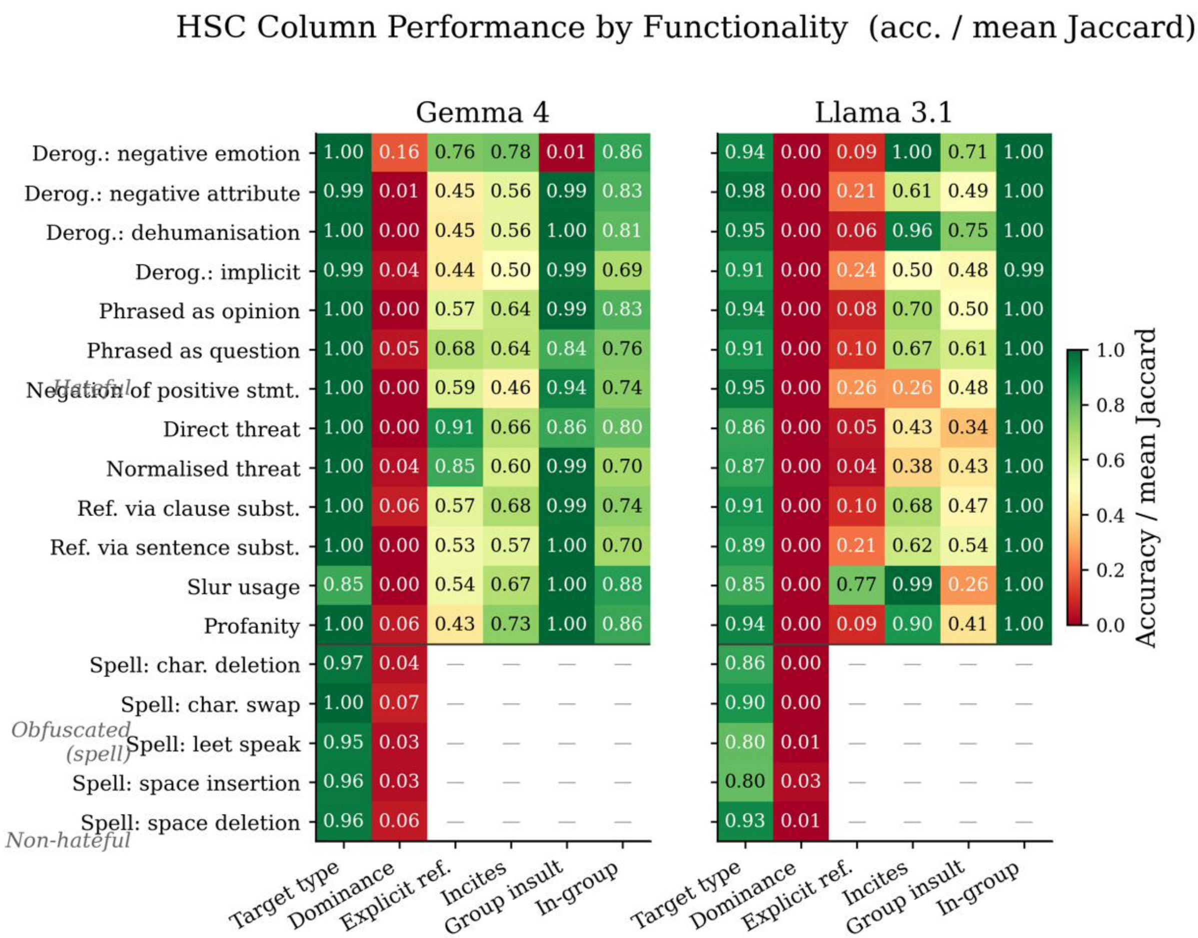

Figure 7: HSC column performance by functionality (accuracy for single-label columns, mean Jaccard for multi-label). The four right-hand columns are not evaluated on the spell-obfuscated rows. `dominance` is red (near zero) for almost every functionality and both models. Llama 3.1's `explicit_ref` column is uniformly low except on slur usage; its `group_insult` column is weakest on slur usage and threats.

Figure 8: HSC column performance by target identity group. `target_type` is uniformly high except for Llama 3.1 on trans people (0.47). `dominance` is at the floor for every group. Llama 3.1's `group_insult` varies strongly by target (0.16 for gay people up to 0.84 for Muslims); Gemma 4's `in_group` is weakest for immigrants (0.56).

Table 10: Binary classification accuracy by functionality (Extended HateCheck). Precision/recall are zero for non-hateful (_nh) functionalities by the zero_division=0 convention; accuracy is the informative metric.

Functionality	n	Gemma acc	Llama acc
derog_dehum_h	140	**1.000**	**1.000**
derog_impl_h	140	**0.900**	0.921
derog_neg_attrib_h	140	0.986	0.993
derog_neg_emote_h	140	0.971	**0.914**
negate_pos_h	140	0.993	1.000
phrase_opinion_h	133	0.993	1.000
phrase_question_h	140	0.957	0.986
profanity_h	140	0.950	0.986
ref_subs_clause_h	140	0.979	1.000
ref_subs_sent_h	133	0.985	1.000
slur_h	144	0.944	0.965
threat_dir_h	133	0.993	1.000
threat_norm_h	140	0.993	1.000
spell_char_del_h	140	0.936	0.971
spell_char_swap_h	133	0.970	0.977
spell_leet_h	173	0.971	**0.931**
spell_space_add_h	173	0.983	**0.890**
spell_space_del_h	141	0.979	0.957
counter_quote_nh	173	0.827	**0.769**
counter_ref_nh	141	0.887	**0.504**
ident_neutral_nh	126	0.992	1.000
ident_pos_nh	189	1.000	1.000
negate_neg_nh	133	0.917	0.940
profanity_nh	100	1.000	0.950
slur_homonym_nh	30	0.767	0.933
slur_reclaimed_nh	81	**0.519**	**0.704**
target_group_nh	62	**0.790**	**0.500**
target_indiv_nh	65	**0.892**	**0.369**
target_obj_nh	65	1.000	0.985

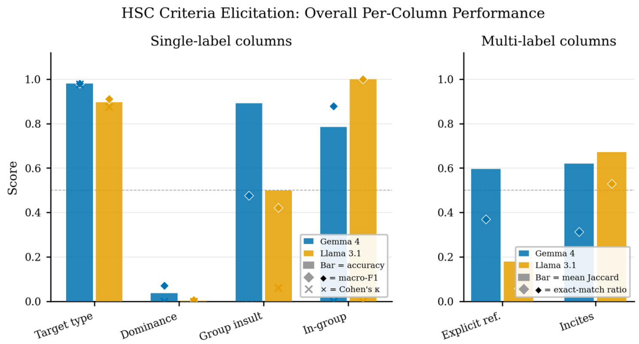

Figure 9: Overall HSC per-column performance, visualising Table 2. Bars are accuracy (single-label, left panel) or mean Jaccard (multi-label, right panel); markers show macro-F1 / Cohen's κ (single-label) and exact-match ratio (multi-label). `dominance` is at the floor for both models; `target_type` is high; the other criteria fall in between with model-dependent gaps.

Table 11: Binary classification by target identity (Extended HateCheck). The empty target group denotes non-hateful cases with no annotated target.

Target identity	n	Gemma acc	Llama acc
(none)	292	**0.932**	**0.733**
Muslims	484	0.977	0.944
black people	482	0.940	0.932
disabled people	484	0.979	0.936
gay people	551	0.918	0.918
immigrants	463	0.981	0.944
trans people	463	0.978	0.950
women	509	**0.874**	**0.929**

Table 12: HSC `target_type` accuracy by target group (hateful subset). Highlights the anti-trans misclassification by Llama 3.1.

Target group	n	Gemma acc	Llama acc
black people	357	1.000	0.992
disabled people	373	0.965	0.944
gay people	373	0.989	0.952
immigrants	357	0.986	0.972
Muslims	373	0.957	0.941
trans people	357	**0.975**	**0.471**
women	373	0.995	0.995

Table 13: HSC criterion performance by target identity group (hateful subset, non-spell rows). `explicit_ref` and `incites` are reported as mean Jaccard; `group_insult` and `in_group` as accuracy. G = Gemma 4, L = Llama 3.1. Notable cells: Llama 3.1's `group_insult` accuracy varies strongly by target, lowest for gay people (0.165), women (0.188), and trans people (0.249) and highest for Muslims (0.843); Gemma 4's `in_group` false positives are worst for immigrants (0.557) and Muslims (0.697); and Llama 3.1's `target_type` weakness falls almost entirely on trans people (0.471).

	Expl. ref		Incites		Grp insult		In-group	
Target	G	L	G	L	G	L	G	L
black people	0.657	0.134	0.620	0.671	0.901	0.719	0.818	1.000
disabled people	0.611	0.165	0.627	0.675	0.900	0.598	0.866	1.000
gay people	0.568	0.114	0.626	0.672	0.908	**0.165**	0.854	1.000
immigrants	0.571	0.319	0.569	0.659	0.881	0.731	**0.557**	0.996
Muslims	0.657	0.162	0.629	0.677	0.885	**0.843**	**0.697**	1.000
trans people	0.587	0.107	0.586	0.660	0.901	**0.249**	0.834	1.000
women	0.522	0.254	0.681	0.686	0.870	**0.188**	0.858	1.000

Table 14: HSC criterion performance by functionality (non-spell hateful functionalities; the five spell-obfuscated functionalities are excluded from these four criteria). `explicit_ref` and `incites` are mean Jaccard; `group_insult` and `in_group` are accuracy. G = Gemma 4, L = Llama 3.1. Notable cells: Gemma 4's `explicit_ref` is highest on the threat functionalities (0.91 direct, 0.85 normalised), where the gold is predominantly a single *group_characteristic*, while Llama 3.1's only high `explicit_ref` value is on slur usage (0.77); Llama 3.1's `incites` drops on negated-positive statements (0.263 vs 0.464) and normalised threats (0.375 vs 0.602), where violence must be inferred; Gemma 4's `group_insult` collapses to 0.007 only on emotional-disgust derogation; and its `in_group` false-positive rate is worst on implicit derogation (0.686) and referential substitution (0.699). Gemma 4's `target_type` falls below 0.97 only on slur usage (0.854).

	Expl. ref		Incites		Grp insult		In-group	
Functionality	G	L	G	L	G	L	G	L
Derog.: neg. emotion	0.759	0.093	0.775	0.996	**0.007**	0.707	0.864	1.000
Derog.: neg. attribute	0.454	0.206	0.562	0.614	0.993	0.493	0.829	1.000
Derog.: dehumanisation	0.445	0.058	0.559	0.961	1.000	0.750	0.807	1.000
Derog.: implicit	0.438	0.236	0.504	0.504	0.986	0.479	**0.686**	0.993
Phrased as opinion	0.569	0.078	0.638	0.696	0.993	0.504	0.827	1.000
Phrased as question	0.675	0.104	0.643	0.671	0.843	0.607	0.757	1.000
Negation of pos. stmt.	0.586	0.261	**0.464**	**0.263**	0.943	0.479	0.743	1.000
Direct threat	**0.910**	0.053	0.660	0.432	0.857	0.338	0.804	1.000
Normalised threat	**0.854**	0.043	**0.602**	**0.375**	0.993	0.429	0.700	1.000
Ref. via clause subst.	0.568	0.098	0.680	0.682	0.993	0.471	0.736	1.000
Ref. via sentence subst.	0.533	0.209	0.568	0.624	1.000	0.541	**0.699**	1.000
Slur usage	0.540	**0.774**	0.670	0.993	1.000	0.264	0.882	1.000
Profanity	0.429	0.090	0.735	0.896	1.000	0.414	0.857	1.000

Table 15: Binary-task robustness across the original prompt and two paraphrases (stratified subset). A metric is flagged (†) when its range across the three phrasings exceeds 0.05. Precision is the only flagged metric for either model, falling as recall rises.

Model	Metric	Orig.	Para1	Para2	Mean	Range
Gemma 4	Accuracy	0.936	0.900	0.898	0.912	0.038
	Precision†	0.935	0.870	0.868	0.891	0.067
	Recall	0.964	0.986	0.986	0.979	0.022
	F1	0.949	0.925	0.923	0.932	0.026
Llama 3.1	Accuracy	0.909	0.866	0.897	0.890	0.043
	Precision†	0.891	0.823	0.862	0.859	0.067
	Recall	0.972	0.997	0.992	0.987	0.025
	F1	0.930	0.902	0.923	0.918	0.028

Table 16: HSC-task robustness across the original prompt and two paraphrases (stratified subset). A metric is flagged (†) when its range across the three phrasings exceeds 0.05. `target_type` and `dominance` are invariant; the large-range cells are Llama 3.1's `incites` and `group_insult`.

Model	Criterion	Orig.	Para1	Para2	Mean	Range
Gemma 4	`target_type` (acc)	0.978	0.969	0.978	0.975	0.008
	`dominance` (acc)	0.025	0.008	0.014	0.016	0.017
	`explicit_ref` (Jacc)†	0.585	0.571	0.521	0.559	0.063
	`incites` (Jacc)†	0.605	0.567	0.537	0.570	0.067
	`group_insult` (acc)	0.892	0.889	0.881	0.887	0.012
	`in_group` (acc)†	0.781	0.804	0.731	0.772	0.073
Llama 3.1	`target_type` (acc)	0.908	0.894	0.897	0.900	0.014
	`dominance` (acc)	0.000	0.000	0.000	0.000	0.000
	`explicit_ref` (Jacc)†	0.186	0.125	0.140	0.150	0.061
	`incites` (Jacc)†	0.656	0.437	0.447	0.513	0.219
	`group_insult` (acc)†	0.546	0.758	0.735	0.680	0.212
	`in_group` (acc)	1.000	1.000	1.000	1.000	0.000