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

Yuanze Xiong - Responsible Professor: Pradeep Murukannaiah - Supervisor: Urja Khurana

Introduction

- Hate speech is a widespread issue. Although what precisely constitutes hate speech remains highly subjective and varies significantly across different countries, communities, and legal jurisdictions.
- Generative Large Language Models (LLMs) offer a flexible alternative to traditional fine-tuning, as they can be guided toward a desired perspective or taxonomy through direct prompting.
- This research establishes a baseline scenario by probing the intrinsic model behaviours without any additional definitional input to learn the default taxonomy and boundaries that models naturally use to classify hate speech.

Objectives

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

SQ1 Baseline consistency

Can the models consistently and correctly classify hate speech with no definitional guidance?

SQ2 Definitional blind spots

How does intrinsic behaviour align with the six HSC criteria, and what blind spots does this reveal?

SQ3 Memorisation

How does behaviour compare between Extended HateCheck and structurally identical, unseen data?

Framework

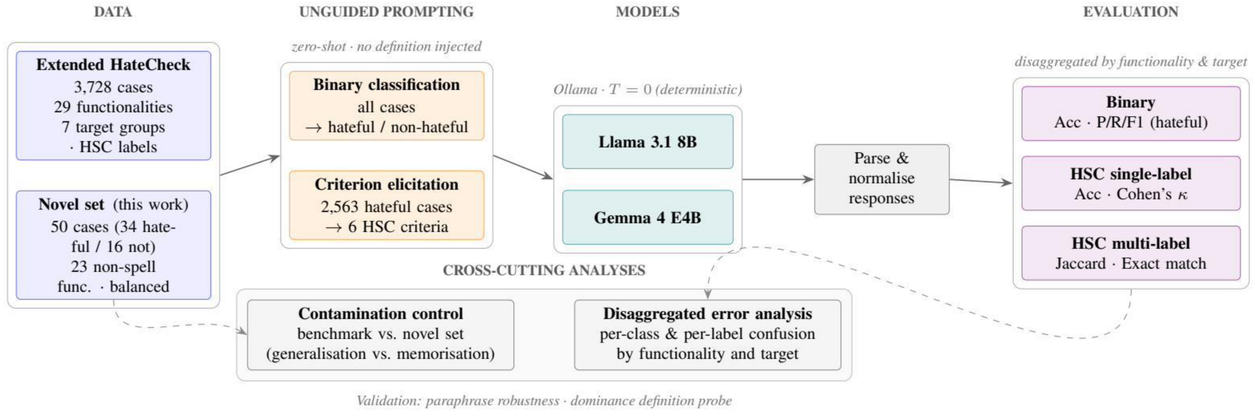
Hate Speech Criteria (HSC)

Khurana et al., 2022 and 2025

A modular framework that decomposes any hate-speech definition into independent components instead of one yes/no verdict to capture the complexity of hate speech.

- **Targeted Type**
- **Dominance (of target)**
- **Explicit Reference**
- **Incitement**
- **Group Insult**
- **In-group (speaker)**

Materials and methods



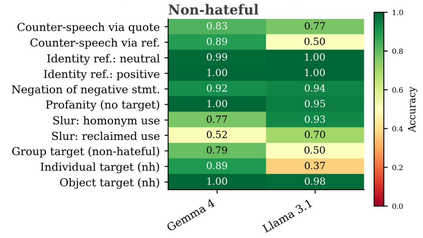
SQ1: Strong on hateful content, weak on non-hateful look-alikes

0.947 Gemma 4 accuracy

0.920 Llama 3.1 accuracy

WHERE THEY FAIL

- Reclaimed slurs are hard for both
- Counter-speech that quotes and bare target references hate trips up Llama 3.1
- Both lean toward labelling group-referential or slur text hateful



SQ2: Three regimes

Reliable `target_type` — Identified with high agreement by both models.

Collapses to the floor `dominance` — definitional blind spot.

In between, model-dependent `explicit_ref` · `incites` `group_insult` · `in_group` — model's prediction bias

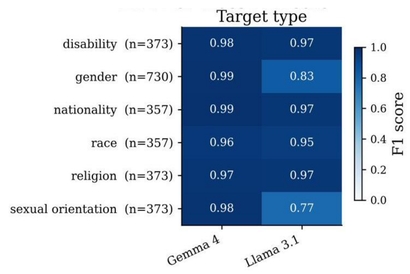
SQ2: Target type is reliably identified

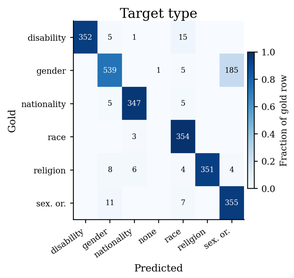
κ = 0.977 Gemma 4

κ = 0.875 Llama 3.1

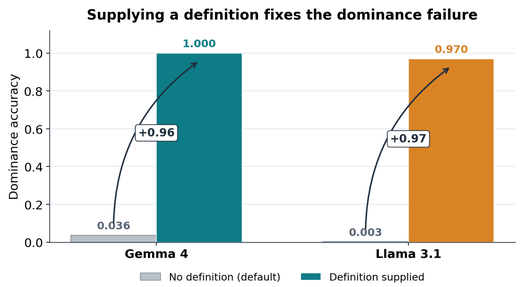
ONE LOCALISED EXCEPTION

Llama 3.1 carries a directional confusion: gender-targeted text is assigned to sexual orientation





SQ2: Dominance misinterpreted with no definition



HSC framework:

Dominance = social position of the targeted group.

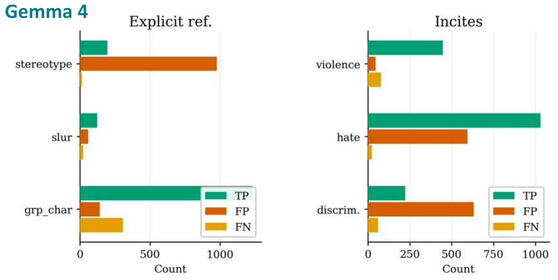
Every HateCheck group is marginalised, so every gold label is "no".

Interpretation: the default failure is a definitional gap, not a deeper inability.

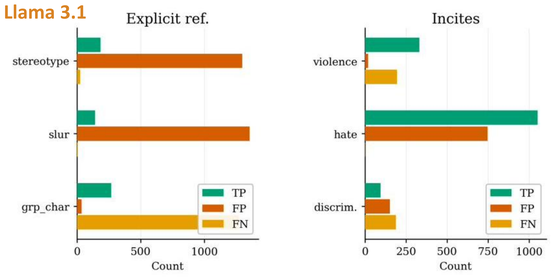
Caveat: All gold label is "no", the jump shows the gap is removed, not that the models can tell dominant from marginalised targets.

SQ2: Over applied labels for multi-label criteria

Gemma 4



Llama 3.1



SQ3: Skills and blind spots persist to unseen data

Rebalancing exposes opposite biases:

On the benchmark *group_insult* is 91% "yes"; the novel set cuts that to 65%, so a model can no longer coast on the majority class.

Gemma 4 eagerly says "yes": group_insult drops 0.892 --> 0.647 once balanced the benchmark figure was an imbalance artefact, not skill.

Llama 3.1 eagerly says "no: rewarded on in_group, where every label is "no" (0.999), but penalised on the mostly-"yes" group_insult.

Conclusion

Text is hateful when hostile surface markers (slurs, group references, expressions of hate) are aimed at an identifiable group, **and which group is targeted is the one component the models represent reliably.**

What the default omits

- the target group's social position (dominance)
- the speaker's identity (in_group)
- whether the whole group is insulted (group_insult)
- finer reference and incitement distinctions, collapsed toward salient labels

Target-aware, but surface-oriented and over-inclusive at its labelling and prompting with no definition is not recommended for nuanced hate speech detection.