The Alignment of Large Language Models' Responses to Subjective Variations in Hate Speech

Comparing Alignment to Real-Life-Inspired Definitions in Zero-Shot Hate Speech Classification

TU Delft Delft University of Technology

Viktoria Bunovska

Email: V.E.Bunovska@student.tudelft.nl

Responsible Professor: Pradeep Murukannaiah, Supervisor: Urja Khurana

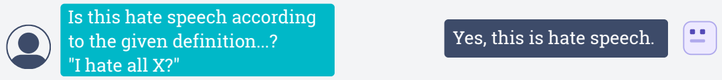
1 Introduction

Background / Motivation

- Hate speech is commonly encountered on social media
- Linked to increased **PTSD** symptom severity
- Manual moderation has become infeasible at scale → **automated** detection systems are essential

Recent proposal: Large Language Models (LLMs)

- Classification through **zero-shot prompting** based on **provided definition**



Central challenge: hate speech is inherently subjective and context-dependent

Many different definitions of hate speech exist and LLMs must **align** their responses with them → NLP research definitions are often mialigned with real-life use cases → Real-world-inspired definitions may better reflect **real-world** moderation practices:

	Definition type	Key Characteristics
Real-world used	Laws	Jurisdictional terms and tone
	Social media policies	Understandable and unambiguous language
Recent proposition based on real-world observations	Framework-based definitions	Created with frameworks for definition design Contain common components encountered in real-world hate speech definitions Based on law and social science

2 Research Question

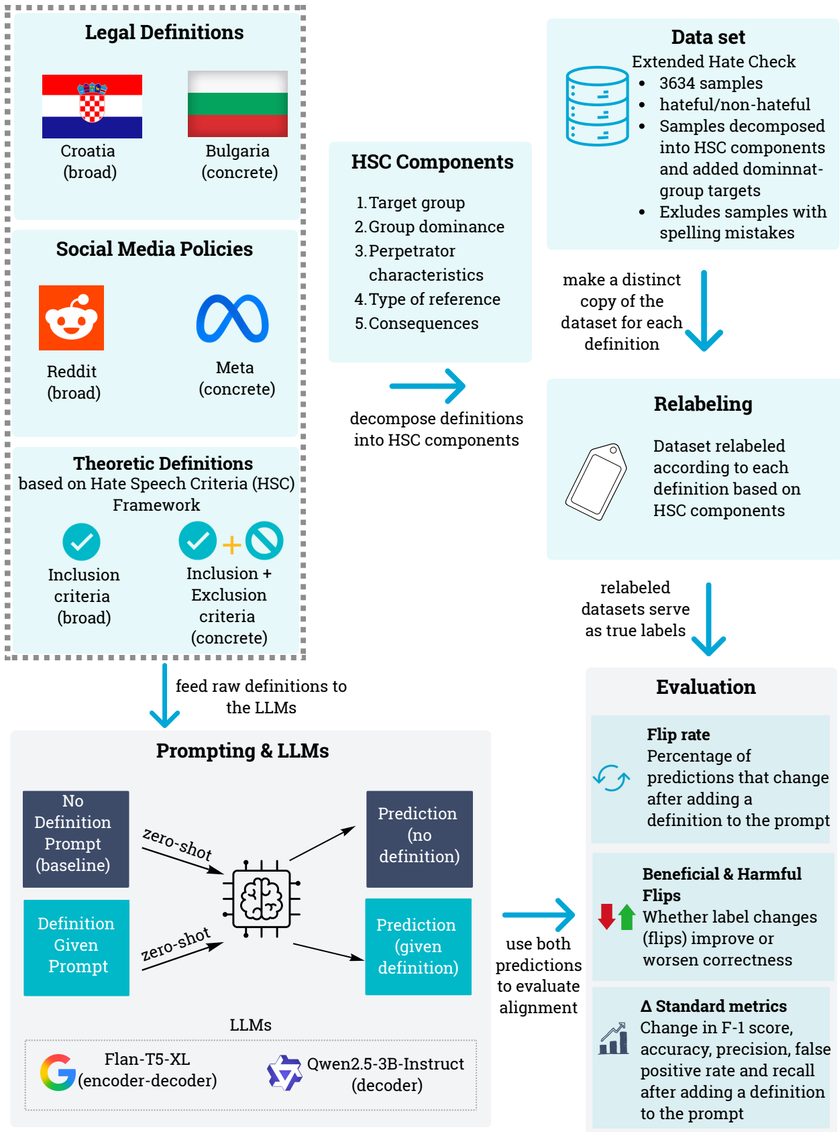
Research gap: Prior zero-shot studies using LLMs have evaluated framework-based (theoretical) hate speech definitions, but no work has directly compared them against real-world legal and platform-based definitions.

Research Question:

How well do LLMs align their hate speech classifications with different types of real-world-inspired definitions in a zero-shot setting?

RQ1: Do framework-based definitions lead to similar levels of LLM alignment as legal and social media policy definitions?
RQ2: Do broader or more concrete real-world-inspired definitions better support LLM alignment?
RQ3: Do samples associated with the hate speech components present in a real-world-inspired definition benefit from including that definition in the prompt?

3 Research Method



4 Results

RQ1 & RQ2:

F1-score and false positive rate (FPR) statistics for both models across definitions

G = definition given in the prompt
O = LLM uses own default definition
Δ = change in result (G-O)
bold = best result per model and definition
underlined = second best per model and definition

Definition	F1-score G	F1-score O	F1-score Δ	FPR G	FPR O	FPR Δ
Bulgaria (concrete)	0.5625	0.5221	0.0404	0.5414	0.6612	-0.1198
Croatia (broad)	0.8703	0.8550	0.0153	0.2889	0.4252	-0.1363
Meta (concrete)	0.8638	0.8550	0.0089	0.3235	0.4252	-0.1017
Reddit (broad)	0.8132	0.7657	**0.0476**	0.3861	0.5358	**-0.1496**
Theor. I+E (concrete)	0.7045	0.6655	0.0390	0.4914	0.6032	-0.1118
Theor. I (broad)	0.6983	0.6655	0.0328	0.5104	0.6032	-0.0928

(a) Flan-T5-XL

Definition	F1-score G	F1-score O	F1-score Δ	FPR G	FPR O	FPR Δ
Bulgaria (concrete)	0.4985	0.4826	0.0159	0.7681	0.8186	-0.0506
Croatia (broad)	0.8460	0.8322	0.0138	**0.5556**	**0.6603**	**-0.1046**
Meta (concrete)	0.8254	0.8322	-0.0068	0.6993	0.6603	0.0391
Reddit (broad)	0.7423	0.7262	**0.0161**	0.6810	0.7406	-0.0595
Theor. I+E (concrete)	0.6223	0.6217	0.0007	0.7810	0.7842	-0.0032
Theor. I (broad)	0.6245	0.6217	0.0028	0.7738	0.7842	-0.0104

(b) Qwen2.5-3B-Instruct

RQ1 & RQ2:

Flip rate statistics for both models across definitions.

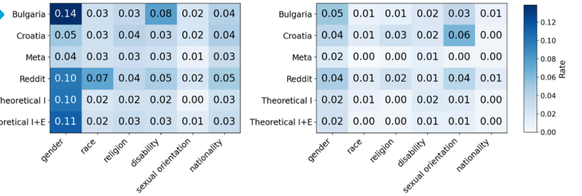
bold = best result per model and definition
underlined = second best per model and definition

Definition	Flan-T5-XL Flips	Flan-T5-XL Flip Rate	Flan-T5-XL Beneficial	Qwen2.5-3B-Instruct Flips	Qwen2.5-3B-Instruct Flip Rate	Qwen2.5-3B-Instruct Beneficial
Bulgaria (concrete)	358	**0.0985**	**0.9134**	193	0.0531	**0.8446**
Croatia (broad)	296	0.0815	0.6858	247	0.0680	0.7085
Meta (concrete)	256	0.0704	0.6406	193	0.0531	0.3782
Reddit (broad)	325	0.0894	0.9108	262	**0.0721**	0.7061
Theoretical I + E (concrete)	289	0.0795	0.9100	186	0.0512	0.5161
Theoretical I (broad)	259	0.0713	0.8842	188	0.0517	0.5585

RQ3:

Beneficial flip distribution for the Target group component

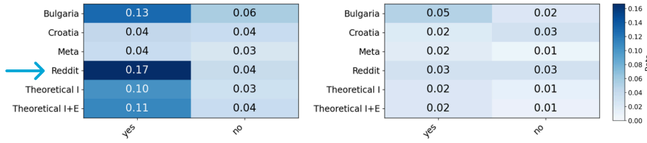
excludes gender- and disability-targeting samples
exclude gender-targeting samples



RQ3:

Beneficial flip distribution for the Dominance component

excludes dominant groups (white people + men)



5 Conclusion and Future Work

- For most cases, the framework-based definitions are outperformed in terms of BFR, but showed potential in some cases (RQ1)
- No evident distinction between broad and concrete definitions' effects (RQ2)
- Samples, whose targets are groups that are excluded by definitions, experience the biggest improvements when these definitions are added to the prompt for Flan-T5 (RQ3)
- Some components are connected, such as dominance and gender, preventing the isolation of effects to single components
- A subsequent qualitative study confirmed most of the observations
- Overall, interactions of model, definitions, and sample characteristics drive LLM alignment

Future work:

→ Explore the linguistic reasons behind the behavior and use the results to improve frameworks in terms of guidelines for structure and wording
→ Repeat the experiments for more models, prompts and definitions to generalize the results with the help of significance testing