Hate Speech Detection in Large Language Models
Using Zero-shot/Few-shot Prompt Engineering

Author:

Parham Bateni
m.bateni@student.tudelft.nl

Supervisor:

Urja Khurana
U.Khurana@tudelft.nl

Responsible Professor:

Pradeep Murukannaiah
p.k.murukannaiah@tudelft.nl

TU Delft
EEMCS, Delft University of Technology, The Netherlands

Introduction

- **Large Language Models** (LLMs) are widely used for hate speech detection, but hate speech lacks a single agreed-upon definition.
- Prior work has mainly studied definition-aware prompting in **zero-shot** settings.
- How **few-shot** prompting interacts with definition format, exemplar selection, and model choice remains largely unexplored.

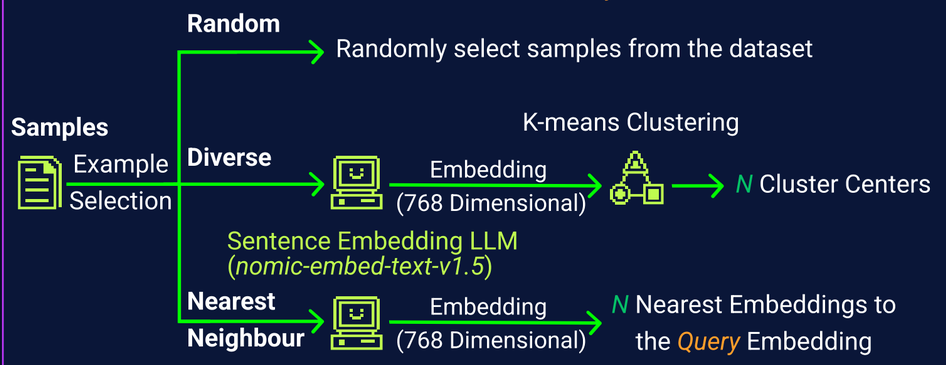
Research Questions

Main Research Question: How do zero-shot and few-shot prompting affect LLM performance in hate speech detection when prompts are enriched with dataset-specific definitions?

- **RQ1**: Does injecting explicit hate speech **definitions** improve performance compared to prompts without them?
- **RQ2**: Which **prompting strategy** is more effective under definition-aware settings?
- **RQ3**: How does performance vary across **LLMs**?

Few-shot Prompting

Goal: Inject *N* examples of hateful and non-hateful text into the prompt to guide classification of the *Query*.



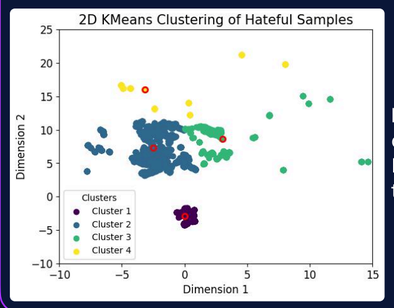


Figure 2. Two-dimensional projection of K-means clusters over hateful samples. Colors indicate cluster assignments. In the Diverse sampling strategy, exemplars are selected as the examples nearest to the cluster centroids, outlined in red.

Results

Prompting	Definition	Macro F1 (%)		
		Gemma-2	Qwen2.5	Llama-3.2
Zero-shot	No Definition	75.0	83.1	80.2
	Vanilla	82.5	83.1	80.0
	HSC	82.0	80.9	79.2
Few-shot Random	No Definition	77.8	83.8	84.0
	Vanilla	78.5	82.8	82.9
	HSC	77.6	83.2	82.1
Few-shot Diverse	No Definition	80.9	84.0	83.9
	Vanilla	80.0	83.5	82.0
	HSC	78.7	83.4	80.9
Few-shot Nearest	No Definition	**87.7**	**91.2**	**94.3**
	Vanilla	84.1	91.0	93.3
	HSC	83.6	90.7	93.8

Table 1. Macro F1 (%) results across prompting strategies and definition conditions for each model. Bold indicates best performance per model.

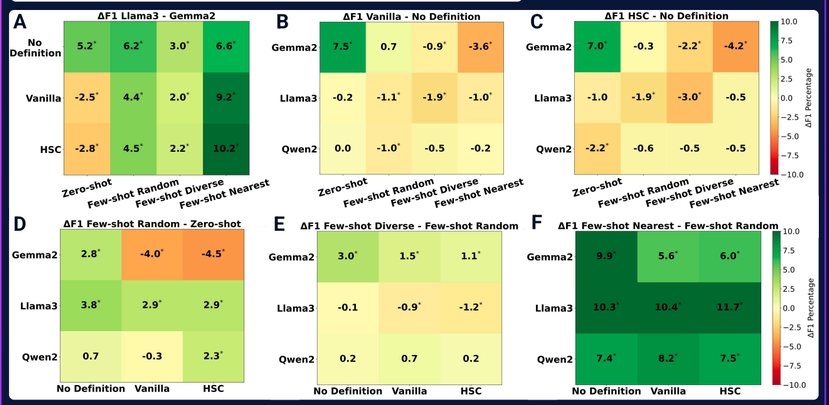


Figure 3. Pairwise differences in macro F1 scores across LLMs, definitions, and prompting strategies. Starred cells indicate statistically significant differences ($p < 0.05$, bootstrap).

Hate Speech Dataset

HateCheck: Functional benchmark for hate speech detection

- 3,728 test cases (2,563 hateful, 1,165 non-hateful)
- 29 functionalities targeting specific linguistic phenomena (e.g., slurs, counter-speech)
- Example: "I hate immigrants." → hateful
  "Let me be clear: I don't hate immigrants." → non-hateful

Hate Speech Definition

- **No Definition**: No explicit definition
- **Vanilla**: Simple textual definition from HateCheck authors
- **Hate Speech Criteria (HSC)**: Structured definition framework with five dimensions derived from HateCheck:
  - Target group (e.g., race)
  - Dominance (e.g., white people)
  - Perpetrator characteristics (e.g., societal role)
  - Negative reference (e.g., stereotypes)
  - Potential consequences (e.g., violence)

Methodology

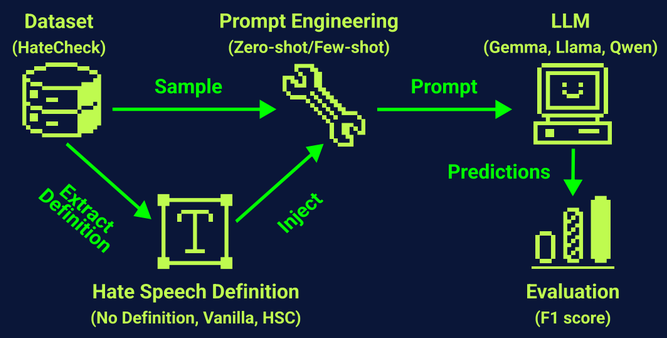


Figure 1. Definition-aware hate speech detection pipeline

Prompt Engineering

Zero-shot

Detect hate speech based on the following **definition**:
{Dataset Hate Speech Definition}

Few-shot

Detect hate speech based on the following definition and **examples**:
{Dataset Hate Speech Definition}
{example1} → hateful
{example2} → non-hateful
...

LLMs

Three instruction-tuned open-weight LLMs:

- **Gemma-2-2B-Instruct** (Google) ~2.6 billion parameters
- **Llama-3.2-3B-Instruct** (Meta) ~3.2 billion parameters
- **Qwen-2.5-3B-Instruct** (Alibaba) ~3.1 billion parameters

Evaluation

Compute macro F1 score by comparing predicted labels with ground-truth labels.

True Label \ Predicted	Hate	Non-hate
Hate	TP	FP
Non-hate	FN	TN

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Conclusions

- **RQ1**: Explicit definitions do not consistently outperform no-definition prompting.
- **RQ2**: Few-shot prompting generally outperforms zero-shot prompting, but gains depend on exemplar quality, with selecting the most similar examples performing best.
- **RQ3**: Larger LLMs perform better on average, but no single model dominates across all settings.

Overall: In hate speech detection, prompting strategy and model choice are more influential than definition wording alone, but performance is ultimately determined by the combination of definition setting, prompting strategy, and model choice.

Limitations and Future Work

- Single dataset (template-based English)
  → Extend to additional datasets (real-world social media data)
- Small models only (2–3B parameters)
  → Evaluate larger and more diverse LLMs
- Single run per setting (fixed seed)
  → Repeat experiments across multiple seeds to assess robustness
- Few-shot exemplars drawn from the test pool
  → Use external retrieval corpora and larger exemplar pools