Can LLMs Evaluate What They Cannot Annotate?

Funded by the European Union

UK Research and Innovation

Usage/License

Download

The code and scripts can be downloaded from its GitHub repository: https://github.com/HYBRIDS-MSCA/hate-eval-agreement

Researchers

- Paloma Piot
- David Otero
- Patricia Martín Rodilla
- Javier Parapar.

These researchers are affiliated with the Information Retrieval Lab at the University of A Coruña.

Summary

This repository provides the implementation and evaluation framework for assessing the reliability of Large Language Models (LLMs) as annotators and evaluators in hate speech detection. It includes subjectivity-aware agreement metrics (xRR), ranking correlation analysis, and benchmarking tools to compare LLM-generated annotations with human judgments. The framework enables reproducible experiments to study whether LLMs can reliably preserve relative model performance trends despite annotation-level disagreement.

Key Features

- Subjectivity-aware evaluation: Uses xRR metrics to better capture agreement in subjective tasks.
- Human vs. LLM comparison: Assesses instruction-tuned LLMs against human hate speech annotations.
- Ranking consistency: Shows LLM labels preserve relative model performance ordering despite low instance agreement.
- Error analysis: Identifies performance differences and systematic biases across hate speech types and targets.

What We Offer

- Open-source resources: Public code, data, and evaluation pipelines.
- Reproducible methods: Clear implementation of subjectivity-aware metrics and ranking analysis.
- Benchmarking framework: Tools to evaluate LLMs as annotators and compare them with human judgments in hate speech detection.

Collaboration Objectives

- Fairness researchers: Apply subjectivity-aware evaluation to other NLP tasks like toxicity or bias detection.
- Moderation practitioners: Use LLMs as scalable evaluation tools where human review is costly.
- Dataset curators: Build richer multi-rater datasets to better analyze LLM behavior.
- Model developers: Create hybrid LLM–human evaluation systems for stronger benchmarking.

hybrids

HYBRIDS Website
www.hybridsproject.eu

Get in Touch
paloma.piot@udc.es