Human Evaluation Guidelines

Task Overview

The human evaluation assesses the quality and accuracy of GPT-4 generated explanations for *Faux Hate* comments. The *Faux Hate Explanation Task* analyzes online comments that resemble hate speech but are driven by misinformation, satire, or misdirected outrage. This evaluation helps validate whether automated systems can accurately identify subtle patterns of harmful or misleading discourse and provide meaningful explanations that reflect human-like reasoning.

Explanation Components

For each comment, the model generates explanations in three components:

- **Target**: Who or what is being attacked or discussed in the comment.
- **Intent**: The underlying motive of the author for posting the comment.
- **Implication**: The potential societal or reader-level impact of the comment.

Annotation Instructions

Annotators are tasked with **rating each component separately** on a scale from 1 (poor) to 5 (excellent):

- **1 - Poor**: The explanation is completely inaccurate or irrelevant.
- **2 - Fair**: The explanation partially aligns but misses key aspects.
- **3 - Moderate**: The explanation captures the gist but lacks precision or clarity.
- **4 - Good**: The explanation is mostly correct with minor gaps or ambiguities.
- **5 - Excellent**: The explanation fully and accurately reflects the meaning implied by the comment.

When rating:

- **Target**: Rate how accurately the explanation identifies the entity or group being referred to or attacked.
- **Intent**: Rate how well the explanation captures the underlying motive of the comment.
- **Implication**: Rate how well the explanation reflects the potential societal or reader-level impact.

If the context is not clear, annotators are encouraged to consult the internet, fact-checking websites, or any reliable background sources to better understand the narrative before annotating the comment. No automated tools should be used; this task is meant to evaluate how well GPT-generated explanations align with human judgment.

Ethical Considerations

The annotation process followed ethical standards. All participants were fully informed about the task and voluntarily agreed to participate. Annotators were also advised to take regular breaks to prevent fatigue and ensure consistent judgment quality.