Deceptive Humor HITL Guidelines

Task Summary and Motivation

Deceptive Humor refers to humorous content that deliberately embeds fabricated or misleading claims in a playful or satirical way. While such comments may appear entertaining on the surface, they subtly propagate misinformation by lowering the reader's guard and reframing false claims as jokes. This unique blend of humor and deception makes it both socially engaging and difficult to detect, posing risks for online discourse and public trust.

However, identifying large-scale real-world examples of deceptive humor is extremely challenging (as understood in our pilot work). On social media, these comments often appear ambiguous, context-dependent, or intertwined with cultural cues. Even expert annotators struggle to establish ground truth without extensive background knowledge of the claims. As a result, real-world data remains sparse, noisy, and unreliable for systematic study.

To overcome this limitation, we adopt a synthetic generation process. Using large language models (LLMs), we generate humorous comments grounded in verified fake claims. This approach ensures scalability, diversity across languages, and precise alignment with known false narratives. Yet, because synthetic text may introduce errors, inconsistencies, or culturally inappropriate humor, a Human-in-the-Loop (HITL) process is essential.

The HITL step ensures that the generated comments are:

- Linguistically natural and grammatically correct,
- Culturally appropriate for the target language or region,
- Genuinely humorous and human-like, and
- Faithful to the deceptive humor nature (humor masking a fabricated claim).

By integrating human evaluation at every stage, we refine the synthetic dataset into a reliable and high-quality benchmark that better reflects real-world deceptive humor while avoiding the pitfalls of raw model outputs.

Review Instructions:

At each iteration (batch of 100 generated comments), annotators must carefully review and validate each comment based on the following dimensions:

1. Spelling and Grammar Check

    - Verify that the comment is free from spelling mistakes, typographical errors, or structural inconsistencies.
    - Ensure that sentences are grammatically well-formed and resemble natural human writing.

2. Humor Quality and Human-likeness

    - Assess whether the comment genuinely resembles human-like humor rather than robotic or templated phrasing.
    - The humor should align with one of the predefined categories (Irony, Absurdity, Social Commentary, Dark Humor, Wordplay) and reflect the intended satire level (Low, Moderate, High).

3. Cultural and Contextual Fit

    - For code-mixed or non-English languages, evaluate whether the humor fits local linguistic and cultural norms.
    - Recognize that cultural references, idioms, or satire may differ significantly across regions and languages. The humor should make sense to a native speaker or culturally aware audience.

4. Deceptive Humor Nature

    - Confirm that the humor is anchored in a fabricated claim, i.e., it uses humor as a vehicle to mask or subtly convey misinformation.
    - The comment should not be a direct statement of the claim without humorous framing, nor should it simply be a random joke unrelated to the claim.

Flagging Protocol:

Annotators must flag and reject any comment that meets one or more of the following conditions:

- **Not humorous**: The comment fails to convey any clear humorous element.
- **Explicit hate or direct attack**: The comment expresses hateful or harmful content toward a group or individual without the deceptive humor framing.
- **Grammatical or spelling errors**: Serious errors that make the comment difficult to read or unnatural.
- **Culturally inappropriate or nonsensical**: The comment does not make sense in the language or cultural context.
- **Incorrect labeling**: The assigned humor attribute or satire level is clearly inconsistent with the content (at the abstract level).

Flagged comments are sent back into the iterative loop with corrective feedback to the model, ensuring quality improvement in subsequent generations.

Contact and Support:

If annotators encounter any kind doubts, edge cases, or ambiguities during the review process, feel free to consult the assigned PhD students for clarification. For complex or unresolved cases, the faculty supervisor is available for further discussion. Open communication is strongly encouraged to maintain consistency and resolve uncertainties in a timely manner.