Suggestion with Explanation

Canary gives you suggestions on how you should disclose risk based on what you have entered about your task so far. It will not save data about your task.

Here’s how I’ll make my recommendation:
1. Task analysis
 The task involves viewing images and text with hate speech, bullying, and harassment — known sources of psychological harm.
2. Risk classification
 These fall under the broader category of explicit or disturbing content, a standard risk label used to signal emotional or psychological difficulty.
3. Disclosure strategy
 To balance clarity and specificity, I recommends a general warning (“Exposure to explicit or disturbing content”) plus specific content tags (“hate speech” and “cyberbullying and harassment”).