1. Over-softening (22%): The system transforms the offensive sentence into an overly polite or mild statement, beyond what is necessary for toxicity removal, thus weakening the intended message.
Original: אתה אפס מאופס שלא מבין כלום בכלכלה ("You're a complete idiot who knows nothing about economics.")
Incorrect: "It's best to discuss economic topics in a positive and respectful manner." Preferred: "You don't understand economics."
2. Omission of key content (17%): The system removes not only toxic elements but also critical non-offensive content, resulting in semantic loss or excessive softening.
Original:
השרים האלה הם נוכלים והם רק דואגים לכיס שלהם ולא לעם
("These ministers are crooks; they only care about their own pockets, not the people.")
Incorrect: "The ministers care about their own pockets." Preferred: "These ministers care about their own pockets, not the people."
3. Addition of unoriginal content: (14%) The system introduces new information absent from the original sentence, thereby altering the intended meaning.
Original: האנשים האלה מטומטמים לחלוטין ("Those people are complete idiots.")
Incorrect: "Can we discuss the different aspects of this issue rationally?" Preferred: "Everyone who thinks that is completely wrong."
4. Unnecessary synonym replacement: (8%) The system substitutes words with less precise synonyms, unintentionally modifying tone or intensity.
Original: הוא היה גזען מגעיל שתמיד דיבר שטויות ("He was a disgusting racist who always talked nonsense.")
Incorrect: "He was an unpleasant person who always said incorrect things." Preferred: "He had racist views and always said incorrect things."
5. Retention of offensive words (2%): The system fails to remove toxic terms or retains them without contextual justification. In some cases, it may be acceptable to keep words with low toxicity if no insult is intended.
Original: האם כל השופטים הם מושחתים ונוכלים ("Are all the judges corrupt crooks?")
Incorrect: "Are all the politicians terrible liars?" Preferred: "Are all the politicians untruthful?"

Figure 1. Error categories.