# LLMs outputs evaluation

These guidelines are designed to help you consistently evaluate how well a Large Language Model (LLM) "neutralizes" a given text. Neutralization is the process of stripping away emotional charge, subjective bias, and inflammatory language while maintaining the original factual core and linguistic integrity.

## Evaluation Framework: Neutralization & Quality

When scoring, consider three main dimensions:

1. **Objectivity:** Is the tone detached, professional, and free of bias?
2. **Fidelity:** Does the output keep all original facts without adding "hallucinated" context or deleting key information?
3. **Linguistic Quality:** Is the grammar, syntax, and flow correct?

### Scoring Rubric

| Score | Label | Key Characteristics |
|---|---|---|
| **5** | **Completely Neutral** | Perfectly objective and professional. Zero grammatical errors. No information added or lost. |
| **4** | **Mostly Neutral** | Balanced language. High fidelity to the source (no extra context). Only minor typos or slight grammatical hiccups. |
| **3** | **Moderately Neutral** | Generally correct but contains traces of bias. May include slight "hallucinations" or minor added context. Some grammatical errors. |
| **2** | **Partially Biased** | Uses loaded expressions. Semantic inconsistencies (truncating info or adding fake context). Notable errors. |
| **1** | **Highly Biased / Failed** | Strongly emotional or loaded language. Technical failure (repetition, incomprehensible, or severe errors). |

# Detailed Level Descriptions

## 5: The Gold Standard

The model has successfully transformed the input into a "just the facts" report.

- **Tone:** Purely informative and professional.
- **Accuracy:** Every fact from the source is present; no outside opinions or extra details are added.
- **Grammar:** Flawless execution.

## 4: High Quality with Minor Flaws

The neutralization is successful, but the writing isn't "perfect" from a technical standpoint.

- **Strict Fidelity:** Like a Level 5, it **does not add any external context**. It stays strictly within the bounds of the provided text.
- **Flaws:** May have a minor agreement error (e.g., gender/number) or a small typo (*refuso*), but the meaning remains clear and neutral.

## 3: The Middle Ground

The output is "okay" but shows signs of the model's own "opinion" or lack of focus.

- **Bias:** You can still "feel" a slight tilt or use of adjectives that aren't strictly necessary.
- **Context:** It might add a small detail not found in the source to "round out" the sentence (referential correctness), or contain slight hallucinations.
- **Grammar:** Noticeable but non-fatal errors.

## 2: Significant Issues

The model failed to remain objective or failed to respect the source material.

- **Bias:** Frequent use of "loaded" or emotional words.
- **Integrity:** The model might cut off half the story (truncation) or invent significant new details that weren't in the prompt.
- **Flow:** The text may feel disjointed or semantically "off."

## 1: Critical Failure

This score is for outputs that are either unusable or offensive.

- **Tone:** Extremist, highly emotional, or aggressive.
- **Technical:** The model might "loop" (repeat the same phrase), output gibberish, or fail to follow the instruction entirely.