# Candidate Guide

Many Mercor projects involve rubrics. Rubrics are used to evaluate model responses to difficult questions. This guide provides an overview of how to create rubrics and outlines some best practices for rubric creation.

## 1. What is a rubric?

Rubrics, also known as auto-graders, are detailed sets of rules outlining the key components of a model's ideal response to a user prompt. Rubrics allow for scalable and consistent feedback on model responses without the constant need for human input. For example, if a user were to ask a model, "Who won the 2025 Super Bowl, and what was the winning score?" an appropriate rubric might read:
"States that the Philadelphia Eagles won the Super Bowl."
"States that the winning score was 40-22."

Crucially, to create quality eval sets, rubrics and prompts must exist in pairs. Rubrics cannot meet the requisite levels of specificity detailed below without evaluating the model's answer to a specific question. Each rubric must be comprised of an exhaustive set of criteria that are clear, self-contained, and individual.

## 2. What is a criterion?

A criterion is one line item in a rubric. It describes *what the model should or should not do* in a specific turn of the conversation.

A good criterion must be:

- **A verifiable statement:** the criterion must be a statement which can be verified with a simple yes or no answer, but not phrased as question (e.g., "maintains a positive tone throughout the essay" is a verifiable statement)
- **Clear:** The criterion uses simple, unambiguous language that anyone can understand.
- **Measurable:** The criterion describes behavior that can be observed and judged (not vague like "be good").
- **Non-Redundant:** The criterion does not duplicate others and adds unique evaluation value.
- **Self-contained:** The criterion can be used to evaluated on its own without extra context or hidden knowledge.

## 3. What should a criterion include?

Each criterion must include (for section 2 of the assessment you will only need Tag and Description from this list):

- **Description**: What behavior is expected (e.g., “Model ignores background noise and responds only to user speech”).
- **Weight**: Major / Minor
- **Tag**: the top-level categories (Audio Understanding, Audio Generation, Text Reasoning, Text Knowledge)
- **Grade:**
  - **Fully met:** Response fully aligns with expectation.
  - **Partially met:** Response somewhat aligns but has noticeable gaps.
  - **Not met:** Response does not align with expectations.

## 3. Weighting Scale

- **Major**: Important. If unmet, the task fails.
- **Minor**: Nice-to-have; affects polish but not correctness.

## 4. Tagging System

- **Audio understanding**: How well does the model interpret audio inputs
- **Audio generation**: How well the model produces natural and appropriate speech output. This is influenced both by implicit or cultural expectations and by explicit requests in the prompt to respond a certain way.
- **Text reasoning:** How well the model reasons logically and follows instructions
- **Text knowledge**: How well the model preserves factual correctness