

AI Evaluation Bridge – AI Evaluation Framework Cheat Sheet

This quick reference guide summarizes the core concepts used to evaluate AI prompts, reasoning, and outputs. Use this sheet while practicing AI evaluation or reviewing model responses.

1. Prompt Structure

- Clearly defines the task for the AI.
- Specifies the output format and scope.
- Removes ambiguity about what the model should produce.

2. Reference Anchors

- Trusted sources used to verify factual accuracy.
- Examples: academic papers, trusted databases, official reports.

3. Evaluation Rubric

- Breaks the evaluation into specific criteria.
- Ensures consistent scoring across responses.

4. Dependencies

- Some criteria must be correct before others can be evaluated.
- Example: reasoning cannot be valid if factual extraction is incorrect.

5. Weighting

- Different criteria contribute different importance to the final score.
- Example: accuracy may carry more weight than formatting.

6. Golden Response

- The ideal answer used as the benchmark for comparison.
- Helps evaluators judge completeness and correctness.

7. Limiter Analogy

- Evaluation frameworks act like a limiter in audio engineering.
- They prevent responses from exceeding acceptable limits of logic, accuracy, or instruction compliance.

Quick Evaluation Flow

Step	Evaluation Question
Prompt	Is the task clearly defined?
Extraction	Are factual elements identified correctly?
Reasoning	Does the logic follow from the facts?
Comparison	How does the output compare to the golden response?
Scoring	Does the response meet rubric criteria?