Prompt Evaluation Checklist

This playbook judges a prompt by its outputs across many runs, after it has run, against a fixed set of inputs and a written rubric. It is the post-output companion to the Prompt Review Checklist, which checks a single prompt before you send it. Use this when one prompt feeds real work and you need to know whether a change made it better, worse, or just different.

Build the harness

- [ ] Lock a fixed input set of 20-50 real inputs, drawn from production and past failures, that do not change between runs.
- [ ] Write the rubric down: each dimension (correctness, grounded, format, tone, safety) has a definition and a pass bar.
- [ ] Attach a reference answer or clear pass criteria to each input so a scorer can tell pass from fail.
- [ ] Pin and record the conditions: model version, temperature, and system prompt stay the same every run.

Run and score

- [ ] Score each input 3-5 times, not once, so run-to-run variance is visible.
- [ ] Keep the score for every input, not just the average, so a fix that breaks one case cannot hide.
- [ ] Score the old prompt and the new prompt on the same inputs in the same run, and compare the per-case difference.
- [ ] Read the low-scoring transcripts yourself, because an average never tells you why something broke.

Decide and protect

- [ ] Block any change where an input that used to pass a safety or refusal check flips to fail, even once.
- [ ] Record the decision (ship, hold, or revert) with the number and the date so the next person inherits evidence.