The marketing AI eval rubric

A one-page rubric for grading AI-generated marketing output: the binary checks that need no model, the anchored scales for the subjective part, and the thresholds that decide what ships.

THE GOLDEN SET

- [ ] 30 to 50 real inputs from work you have already done, never invented examples
- [ ] Include the cases that have burned you: the regulated category, the confusing product name, the empty input
- [ ] Freeze it and version it. A set that quietly grows is not a baseline

BINARY CHECKS, GRADED BY RULES

A yes or a no, no partial credit. One no blocks the output.

- [ ] Every number, date, price, and specification appears in the source material
- [ ] No claim is made that the source does not support
- [ ] Length and format match the destination's constraints
- [ ] Required disclosures and legal text are present and unaltered
- [ ] No banned term, competitor name, or unapproved claim appears
- [ ] Product and brand names are spelled exactly as the brand spells them

SCALE ONE: BRAND VOICE

- [ ] 0 = would embarrass the person who owns the brand
- [ ] 1 = recognizably off, generic or borrowed from another category
- [ ] 2 = correct register, nothing distinctive
- [ ] 3 = indistinguishable from work the team would have written

SCALE TWO: RELEVANCE TO THE BRIEF

- [ ] 0 = answers a different question
- [ ] 1 = answers a nearby question, misses the actual ask
- [ ] 2 = answers the ask, misses the emphasis
- [ ] 3 = answers the ask with the emphasis the brief called for

SCALE THREE: USEFULNESS TO THE READER

- [ ] 0 = says nothing a reader could act on
- [ ] 1 = correct and obvious
- [ ] 2 = one specific, useful point
- [ ] 3 = the reader is better equipped than before they read it

THE THRESHOLDS

- [ ] Any binary check failed, block. It never reaches a human queue
- [ ] Every scale at 2 or above, ship
- [ ] Any scale at 1, route to a human
- [ ] Any scale at 0, block and add the case to the golden set

IF A MODEL IS DOING THE GRADING

- [ ] Calibrate against human ratings once, on 50 cases, and write the agreement rate down
- [ ] Run every pairwise comparison in both orders, keep only verdicts that survive the swap
- [ ] Score at least twice and aggregate, because single runs vary
- [ ] Never let a model grade its own family's output when choosing between models
- [ ] Recalibrate whenever the judge model changes

REGRESSION, THE ACTUAL POINT

- [ ] Rerun on every change to the prompt, the model, the retrieval source, or the tools
- [ ] Rerun on a schedule regardless, because dependencies change underneath you
- [ ] Compare against the last recorded score, not against a memory of it
- [ ] Treat a drop below threshold as a blocker, or accept that the eval is decorative