AutoQA criteria — action list

All 68 live criteria classified against the capability limits (audit of the authenticated 2026-07-18 export; config verified unchanged 2026-07-23). Totals: 1 automatable as written · 1 after rewrite · 27 advisory/triage pending per-criterion validation · 17 not safe for automated adverse use · 22 not reconstructable from the export.

Block from adverse use — advisory or human-only until rebuilt and validated (17)

ID	CRITERION	LEVEL	WHY BLOCKED
CRQ–3	False Ties	Turn	"Very clear quality gap" is a global preference judgment in the at-chance regime; also infers tasker effort (a mental state).
PWQ–1	Dictating Implementation Details	Sub	"What vs how" boundary is not objective.
PWQ–3	Anti-Production Steering	Sub	Anti-production steering is an open engineering norm.
RCD–1A	Model A Pros Verification	Turn	Unbounded semantic verification over whole trajectories.
RCD–1B	Model A Cons Verification	Turn	Same class.
RCD–2A	Model B Pros Verification	Turn	Same class, Model B.
RCD–2B	Model B Cons Verification	Turn	Same class, Model B.
RCD–3	Preference Justification Verification	Turn	Whole-trajectory comparative alignment of preference and rationale.
RCD–7	Exaggeration Detection	Turn	Exaggeration is a severity-taste call, not a record fact.
RCD–8A	Model A Contrived Cons	Turn	"Contrived / non-issue in practice" is normative.
RCD–8B	Model B Contrived Cons	Turn	Same, Model B.
RCD–9A	Model A Contrived Feedback	Turn	Same class as RCD-8.
RCD–9B	Model B Contrived Feedback	Turn	Same, Model B.
TR–13A	Incorrect Over-engineering A	Sub	Over-engineering judgment: needs norms, intent, and project taste the record does not fix.
TR–13B	Incorrect Over-engineering B	Sub	Same as TR-13A, Model B.
TR–7A	Incorrect Destructive Ops A	Sub	Destructive-ops correctness both directions: needs permission model and environment facts often absent from the record.
TR–7B	Incorrect Destructive Ops B	Sub	Same as TR-7A, Model B.

No authority until auditable — operative prompt not reconstructable from the export (22)

ID	CRITERION	LEVEL	REQUIRED FIRST
AI–N–1	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
AIV–6	Hedging Language	Sub	Export full prompt body, I/O schema, and version; then classify like the rest.
CRQ–N–1A	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
CRQ–N–1B	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
DCM–N–1	Change Rollback	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
DCM–N–2	Prescribing Behavior Ideals	Sub	Export full prompt body, I/O schema, and version; then classify like the rest.
NAIV–7	Not the Real Thing	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.

NAIV-8	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
PWQ-N-1A	H	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
PWQ-N-1B	H	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
PWQ-N-2	Poor Collaboration	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-1A	L	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-1B	Incorrect Environment Handling (Model B)	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-2A	Incorrect Clarifying Question Penalty (Model A)	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-2B	Incorrect Clarifying Question Penalty (Model B)	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-3A	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-3B	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RCD-N-4	M	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RWQ-N-1A	Incorrect Destructive Actions (Model A)	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RWQ-N-1B	Incorrect Destructive Actions (Model B)	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.
RWQ-N-2	Inconsistent Honesty Reporting	Sub	Export full prompt body, I/O schema, and version; then classify like the rest.
RWQ-N-3	Shallow Analysis	Turn	Export full prompt body, I/O schema, and version; then classify like the rest.

Cleared for exact detection, with caveats (2)

ID	CRITERION	LEVEL	CONDITION
AIV-1	Em Dash Usage	Sub	Deterministic signal only. Never sole proof of AI authorship; quoting the model's own output must not count (citation rule).
NAIV-6	Honeypot Catch	Sub	Deterministic after the sentinel set is explicitly enumerated; not safe for adverse use as written.

Fix before expanding any authority

1. One preference-scale constant everywhere; the live 0–3/4–7 no-tie mapping contradicts the instructed 0–2 / 3–4 tie / 5–7 canon and mis-scores compliant ties.
2. An enforcement field on every criterion: advise / flag for a person / may count against work.
3. Treat labeler free text as data in judge prompts (injection hardening + planted-injection tests).
4. Bind every criterion and prompt bundle to an instruction version; retire era-mixed taxonomies.
5. Move deterministic halves (empty fields, rating variance, tie clustering, character checks) from model calls into code.
6. Rename worker-facing criteria to survive being read aloud to the person flagged (e.g. "Colon Cancer").

The rule in one line: most of AutoQA is appropriate to run; large parts are inappropriate to punish with — the 17 judgment-heavy checks, the 22 that cannot be read from the export, any path from an uncalibrated flag to pay, and any check running the broken preference scale. Live cross-evidence on record: on the one retained real case, the platform judge marked all three labeler flags "Wrong," the human lead found all three grounded, and a record-check sided with the human — contradicting one automated verdict with a quote.