Same Accuracy, Different Decisions: Retesting Sheet

GRC Careers | AI Governance Jobs | ai-governance-jobs.com | November 7, 2026

> **The rule of thumb.** An overall accuracy score cannot tell you whether two configurations make the same decisions. Before you rely on a result, check the cases they label differently and the slices where an error costs the most.

1. Record for every test run

- Dated model snapshot (never only an alias)
- Prompt and example set version
- Decoding settings and output format
- Parsing or score-extraction code version
- Retrieval index or tool configuration
- Evaluation dataset version and run date
- Who ran it, who reviewed it, where results are kept

2. Reference labels

- Who labeled them and with what guidance
- How disputes between labelers were settled
- Count of disputed labels, and where they fall in the results

3. Slices and comparisons

- Material slices named *before* testing, each tied to a decision rule
- Counts as well as percentages for each slice
- Number of cases where configurations disagree, by slice
- Run-to-run variation of one unchanged configuration
- Cases where all configurations agree but differ from the reference

4. Decision record

- Configuration approved, by whom, under which rule
- Slices that failed or came close
- Known limits and open questions
- Retest triggers and the owner who watches for them

5. Retest triggers

Change or signal	Retest at least	Typical decision owner
Model snapshot or provider version, including a change behind an alias	All material slices; disagreement against the prior version	Model validation and governance
Prompt, examples or output format	All material slices; parsing checks	Application owner
Parsing, scoring or routing code	Slices plus share of cases routed to human review	Technical assurance
Reference labels or labeling guidance	Rescore past results on the new labels	Evaluation lead
Mix of incoming cases, users or purpose	Slice definitions and decision rules first, then tests	Business owner and compliance
Shift in label shares, staff overrides or complaints in one slice	That slice, then a wider check	Monitoring owner
Unknown change or unexplained drift	Suspend reuse of the old result; broad retest	Assurance lead

6. Red flags

- "Both score 70%, so they are equivalent."
- "All the configurations agree, so the answer must be right."
- A result that names a moving alias such as `latest` instead of a dated version.
- Dropping disputed cases from the test set, which raises the reported score without changing a single label.

Editorial checklist by GRC Careers, based on Louise Yorke, "Beyond accuracy: what AI assurance needs to establish in public services," arXiv:2611.06655 (research preprint, November 4, 2026). Not a standard or legal requirement. Thresholds and owners are yours to set and justify. Full guide: https://www.ai-governance-jobs.com/grc-careers-insights/same-accuracy-different-ai-decisions/