

The Weekly Scorecard

One workflow, one week, ten rows. Record results, trends and the action or decision they require.

Workflow

Lane

Week ending

Owner

FIELD / MEASURE		YOUR ENTRY
SCALE	Adoption Runs this week; distinct users; share of eligible volume actually flowing through the workflow.	
SCALE	Cost per run (blended) Tokens in/out × rates + retrieval + orchestration overhead, blended across routed models.	
SCALE	Latency (p95) 95th-percentile end-to-end response time against the SLA target.	
SPECIFY	Schema validity % of outputs passing machine validation. Hard gate: 99.5%.	
GROUND	Groundedness Percentage of golden-set outputs passing the citation-or-refusal check: every factual claim has the required approved-source citation, or the response correctly follows the case's refusal/clarification path. Release gate: 100%.	
EVALUATE	Quality sample Human-rated helpfulness/accuracy on this week's audit sample (per the Ladder's sampling plan).	
GOVERN	Escalation & refusal rates % escalated to humans; % refused. Watched in both directions: a spike can mean the system stopped working; a slide can mean it started guessing.	
EVALUATE	Drift delta Golden-dataset score this week vs. baseline; alert on threshold crossings.	
GOVERN	Incidents & near misses Count + one-line descriptions; every incident becomes a golden-dataset case.	
SCALE/GOVERN	ROI proxy + autonomy status The workflow's outcome metric (e.g., cost per resolved ticket) and current Ladder rung, including any rung changes this week.	

Use separate lane measures where checks differ. Record sample sizes and denominators. The printed release gates are the book's worked standard; agree on your own before testing.

