



Choosing Metrics

Metrics and KPIs 01 | North star vs guardrail vs diagnostic metrics, what makes a metric good, offline vs online metrics and why they disagree, proxy metrics and how to validate them, KPI driver trees and the metric definition checklist

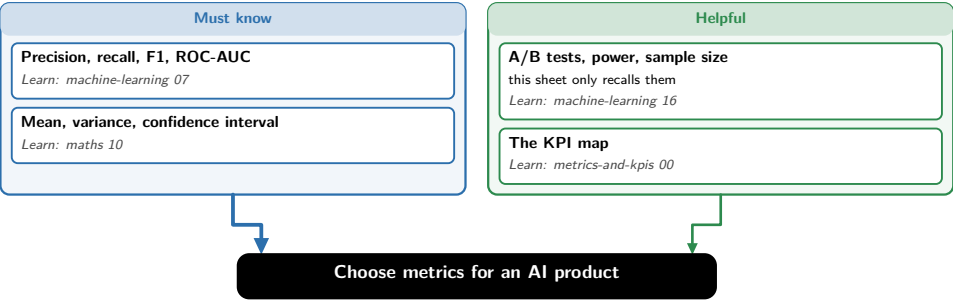
#Metrics #NorthStar #Guardrails #OfflineOnline #MetricTree

Three roles for a metric

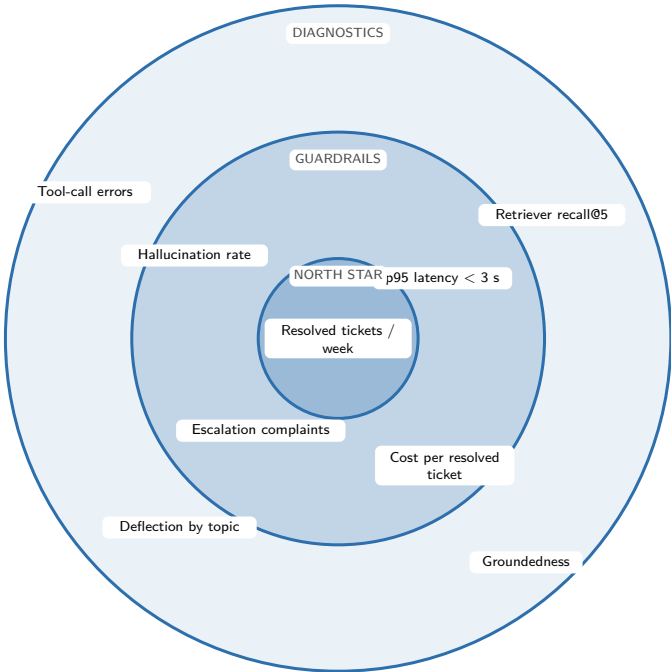
Driving a lorry to a city	Running an AI product
Arrival time at the destination	North star: the outcome users value (resolved tickets, retained users)
Speed limit, fuel level, engine temperature	Guardrails: latency, cost per request, error rate, complaint rate
Tyre pressure, gear, rev counter	Diagnostics: recall@10 of the retriever, step completion rates
Satnav estimate before leaving	Offline metric: predicted on yesterday's map
Actual traffic on the road	Online metric: what live users really do

WHERE THE ANALOGY BREAKS A driver's goal does not react to the dashboard. Users and models do react to what you measure, so a metric that becomes a target can drift away from the goal (Goodhart's law, sheet 11).

Choosing metrics is choosing what the dashboard of a long drive shows



What this sheet assumes



One north star at the centre, a handful of guardrails, as many diagnostics as you need. Example: a retrieval-augmented support assistant

Property	Test to apply	Fails when
Aligned	If it moves, does the user or business outcome move in the same direction?	clicks rise on clickbait
Sensitive	Can it detect a realistic change (say 1 %) with one or two weeks of traffic?	annual churn, revenue per user with heavy tails
Attributable	Can the team that owns it actually move it?	company revenue as a model team's target
Timely	Is it known within the decision window?	12-month retention for a weekly launch

Hard to game	Is the cheapest way to move it also the right way?	average handle time (agents hang up early)
Understood	Can a manager explain it and its direction in one line?	a 7-term weighted index

Properties of a good metric. A candidate that fails two of them is a diagnostic at best

North stars by product, and the driver tree			
Product	North star	Guardrails	Diagnostics
Search / RAG	successful sessions (answer used, no re-query)	p95 latency, zero-result rate	recall@k, MRR, NDCG@10
Recommender	long-term engagement per user (e.g. 7-day active)	diversity, creator coverage, complaints	CTR, dwell, catalogue coverage
Fraud detection	fraud loss avoided net of review cost	false-positive rate on good customers	recall at fixed precision, alert volume
Support assistant	tickets resolved without a human	cost per resolution, CSAT, hallucination rate	groundedness, tool-call accuracy
Coding assistant	accepted changes merged and kept	latency, security findings	acceptance rate, edit distance after accept
Demand forecast	inventory cost (stock-outs + overstock)	bias (mean error) per region	MAPE, pinball loss, interval coverage

Starting points by product type. Write the threshold of each guardrail next to it

💡 How to use the tree

Give each leaf to one owner: the model team owns the click rate and purchase-after-click, growth owns active users, merchandising owns order value. A model change is judged on its leaf, then confirmed on the root in an A/B test: a click-rate gain that lowers purchase-after-click (clickbait recommendations) shows up as a smaller root gain than the leaf predicts.

Offline vs online			
Offline metric	vs	Online metric	
fixed labelled set, past logs	Data	live traffic, real users ✓	
minutes, free to rerun	Speed and cost	days to weeks of traffic	
none for users	Risk	real users see the change	
invisible	Feedback effects	captured (users adapt, loops form) ✓	
inherits the old system's exposure	Bias	randomised, causal ✓	
ranking many candidates	Use	go / no-go on one or two	

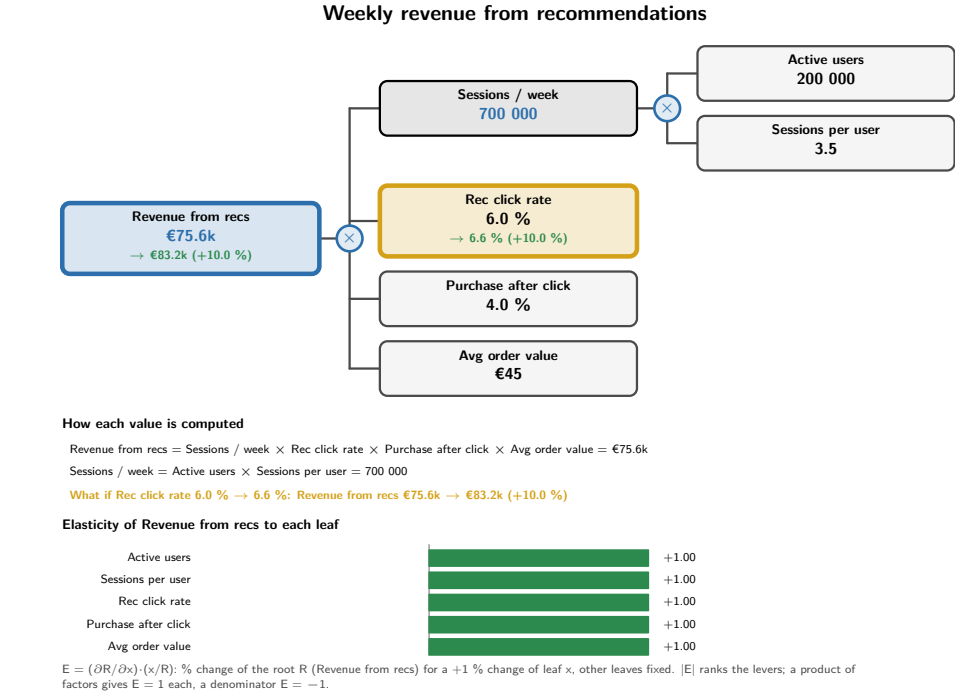
Two ways of measuring the same model. Use offline to filter candidates cheaply, online to decide

Cause	Mechanism	Mitigation
Logged-data bias	labels exist only for items the old model showed; position bias inflates them	randomised exploration slice, inverse-propensity weighting
Proxy mismatch	offline label (click) is not the online goal (purchase, retention)	validate the proxy on past tests (below)
Stale eval set	queries, catalogue and users changed since it was built	refresh from recent traffic, date-split
Latency	a bigger model is better but slower; slower pages lose users	latency guardrail, measure quality at fixed latency
Feedback loops	the model changes what users see and therefore future data	holdout group never exposed to the model
Novelty and learning effects	users react to newness, then settle	run 2+ weeks, plot the daily treatment effect

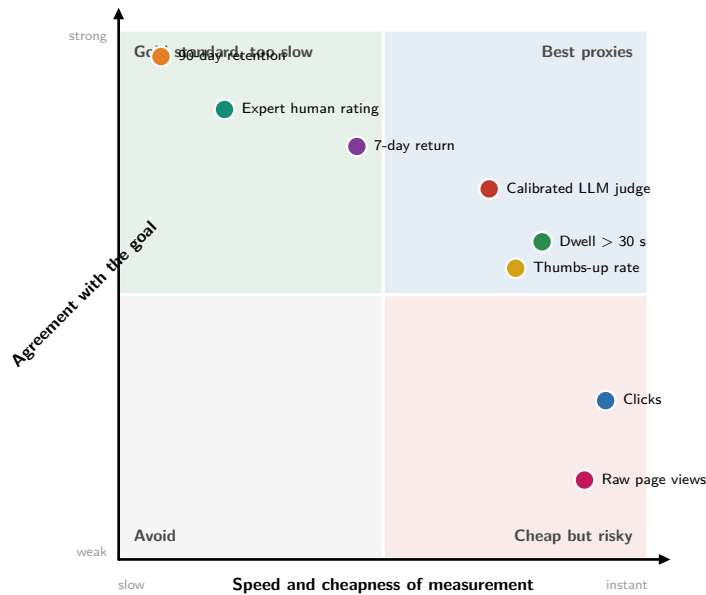
Why a model that wins offline loses online, and what to do about each cause

⚠ Measure the gap itself

Keep a log of every launch: offline delta and online delta of the north star. A low rank correlation between the two (Spearman below about 0.5 over ten or more launches) means the offline metric cannot rank candidates and the eval set or proxy needs work.



Proxy metrics



Candidate proxies for a content product's goal (long-term satisfaction), placed by speed of measurement and by how well they track the goal. Positions are illustrative; measure your own

PROBLEM

Is click-through rate (CTR) a usable proxy for 7-day retention? Six past A/B tests have both numbers.

1. Compare signs

⇒ A +/+, B +/+, C +/-, D -/-, E +/-, F +/+
why: a proxy must at least point the same way

2. Count agreements

⇒ A, B, D, F agree = 4 of 6
why: C and E disagree

3. Rate

⇒ $4 / 6 = 0.67$
why: 0.5 is what a coin would give

4. Inspect the misses

⇒ E had the second-largest CTR gain and the worst retention
why: a clickbait-style change: the proxy rewards exactly what hurts the goal

GIVEN

- Δ CTR per test A-F (percentage points): +2.0, +1.1, +0.4, -0.5, +1.5, +0.8
- Δ 7-day retention (points): +0.6, +0.2, -0.1, -0.3, -0.4, +0.3
- Directional agreement = tests where both deltas have the same sign ÷ tests

ANSWER

CTR agrees with retention on 67 % of tests and fails on the riskiest one: keep CTR as a diagnostic, decide on dwell or 7-day return, add retention as a guardrail

Check: Six tests are few: the 95 % interval on 4/6 runs from about 0.30 to 0.90 (Wilson score interval).
Re-check as tests accumulate

Validating a proxy on past experiments before trusting it

Metric definition checklist (write it before the first dashboard)

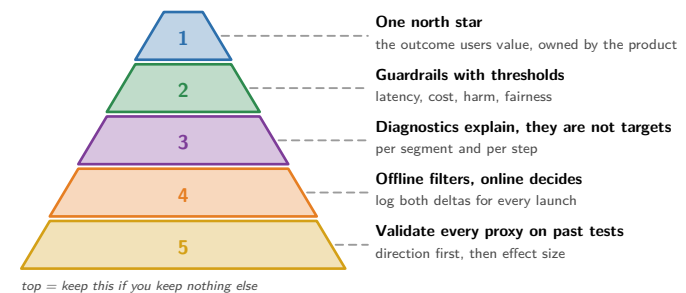
✓ baseline / mandatory □ recommended – verify in your context

- ✓ Numerator and denominator in words and SQL resolved tickets ÷ tickets opened, not ÷ conversations
- ✓ Unit of analysis equals the randomisation unit per user if users were randomised
- ✓ Time window and timezone 7 days after first exposure, UTC
- ✓ Good direction and target up, $\geq 40\%$ by Q2
- **MUST** Filters and exclusions bots, employees, test accounts
- **MUST** Owner and review rhythm
- **SHOULD** Minimum detectable effect with current traffic computed as in machine-learning 16
- **SHOULD** Known ways to game it, and the guardrail that catches each

Myths, recap and self-test

Statement	True	False
1. A team should have three north-star metrics, one per stakeholder → Several north stars conflict and nobody can decide; combine them into one OEC or demote two to guardrails.	☐	☒
2. Guardrails need explicit thresholds → 'Latency must not get worse' is untestable; 'p95 $\leq$ 800 ms and no worse than +5 %' is.	☒	☐
3. If offline NDCG improves, online engagement will too → Logged-data bias, proxy mismatch and latency often reverse the sign; track offline-online agreement.	☐	☒
4. A proxy is valid when it agrees in direction with the goal on most past tests → Directional agreement (and correlation of effect sizes) over many past experiments is the standard check.	☒	☐
5. Revenue per user is always the best north star → Heavy-tailed, slow and easy to buy with short-term tricks that hurt retention.	☐	☒

Check each statement; answers and reasons are printed



What to remember, most important first

Q1 North star vs guardrail in one line each?

A North star: the outcome you push up. Guardrail: a metric with a threshold that must not be crossed while you do.

Q2 What is an OEC?

A Overall evaluation criterion: the one metric, or fixed weighted combination, that decides an experiment.

Q3 Elasticity of the root to any factor of a pure product tree?

A 1: a +10 % change in one factor moves the root by +10 %.

Q4 Name three causes of an offline-online gap.

A Logged-data (position) bias, proxy mismatch, stale eval set; also latency, feedback loops, novelty.

Q5 Why is average handle time a bad target for support agents?

A The cheapest way to lower it is to end conversations early, which hurts resolution.

Q6 How do you validate a proxy?

A On past experiments: same direction of change as the goal on most tests, and correlated effect sizes.

Q7 What should the unit of analysis match?

A The randomisation unit of the experiment (user, session, request).

Q8 Where do diagnostics belong: targets or dashboards?

A Dashboards: they explain movements; targeting them invites gaming.

Q9 When does a proxy become a surrogate metric?

A When past experiments show it predicts the long-term outcome well enough to decide on it alone.