LLM METRICS CHEAT SHEET

THE METRICS THAT ACTUALLY MATTER WHEN EVALUATING AI MODELS

1 SPEED & PERFORMANCE

How fast does the model respond?

METRIC	MEASURES	IF BAD...
Avg Tokens / sec	Output speed after generation starts	Slow streaming and batch processing
TTFT (Time to First Token)	Time before the first token appears	Feels broken before it even starts
P50 Latency	Typical response speed	Every interaction feels sluggish
P95 Latency	Slowest 5% of requests	Ageest stall and queues build
P99 Latency	Worst-case behavior	Timeouts and cascading failures

⚠ WHAT BREAKS? **PRODUCT EXPERIENCE**

2 COST & EFFICIENCY

How much does success actually cost?

METRIC	MEASURES	IF BAD...
Cost per Token	Raw model pricing	Cheap on paper, expensive in practice
Cost per Successful Task	End-to-end task cost	Budget leaks through retries and failures

⚠ WHAT BREAKS? **BUSINESS ECONOMICS**

LLM EVALUATION FRAMEWORK

3 RELIABILITY & QUALITY

Can you trust the output?

METRIC	MEASURES	IF BAD...
Instruction Following	Adherence to constraints	Outputs are almost correct
Schema / JSON Compliance	Structured output quality	Pipelines break
Determinism / Stability	Consistency across runs	Debugging becomes impossible
Variance Across Runs	Sensitivity to prompt changes	QA and auditing suffer
Long Context Effectiveness	Reasoning over large inputs	Important information gets ignored

⚠ WHAT BREAKS? **ENGINEERING WORKFLOWS**

4 AGENTS & OPERATIONS

Can it operate inside real systems?

METRIC	MEASURES	IF BAD...
Tool Selection Accuracy	Choosing the correct tool	Agents fail silently
Tool Argument Correctness	Passing valid parameters	Loops, retries, and errors
Recovery After Failure	Self-correction ability	Human intervention required
Refusal Accuracy	Correct safety behavior	Legitimate tasks get blocked
Refusal Style	How the model refuses	Poor user experience
Uptime & Reliability	Availability	Features stop working
Rate Limits / Throttling	Traffic handling	Growth bottlenecks
Version Stability	Behavior across releases	Silent regressions

⚠ WHAT BREAKS? **AUTOMATION & SCALE**

THE GROWTHROCKS RULE

If a metric breaks **USER TRUST** → Product problem

If a metric breaks **PIPELINES** → Engineering problem

If a metric breaks **BUDGETS** → Business problem

THE FINAL TAKEAWAY

Most teams evaluate models using one metric:

AVG TOKENS/SEC

The teams that scale evaluate:

SPEED + **COST** + **RELIABILITY** + **AGENTS**

Because the fastest model isn't always the best model.
The best model is the one that **fails in ways you can live with.**

GrowthRocks

Pick models. Build systems. Drive growth.
That's the GrowthRocks way.