



Eval

Score



Confusion Matrix

TP	FP
FN	TN



Dataset



Model



Output



Hallucination



Faithfulness



Precision / Recall

AI Evaluation for Testers



LLM-as-a-Judge

50 Essential Terms for Testing LLMs, RAG, Agents & AI Systems

— Understand what to measure, how to evaluate, and what can go wrong. —



Robustness



Prompt Injection



Latency



Task Success





Evaluation (Eval)



Simple meaning

A structured test used to measure how well an AI system performs against defined expectations.



Tester lens

Treat an eval as the AI equivalent of a purposeful test suite: clear input, expected behavior, scoring method, and repeatable evidence.

How it works



Clear input

Set the right conditions for a fair test.



Key checklist

- ✓ Expected behavior
- ✓ Evidence
- ✓ Repeatability



Watch for

Unclear expectations, inconsistent scoring, or non-repeatable tests can lead to misleading results.



Remember

An eval is a structured test, so be clear about expectations, use a consistent scoring method, and keep evidence repeatable.

Evaluation Dataset



Simple meaning

A collection of representative test examples used to measure AI behavior.



Tester lens

Coverage matters more than raw size. Include normal cases, edge cases, risky cases, and real production-like inputs.



Evaluation Dataset



Normal

- common inputs
- expected behavior
- typical use



Edge

- boundary values
- unusual formats
- rare scenarios



Risky

- harmful inputs
- invalid data
- misuse attempts



Real-world

- production-like
- realistic patterns
- varied contexts



Evaluation Run

- ✓ test the model
- ✓ observe behavior
- ✓ find issues
- ✓ improve confidence



Balanced coverage

Include all four types to get a complete picture of behavior.



Watch the gaps

Missing a category can hide real issues and create blind spots.



Be context aware

The exact mix may vary by use case, domain and risk level.



Remember



A good evaluation dataset is not just big – it's well covered, with the right mix of examples.



≡ Test Set ≡



Simple meaning

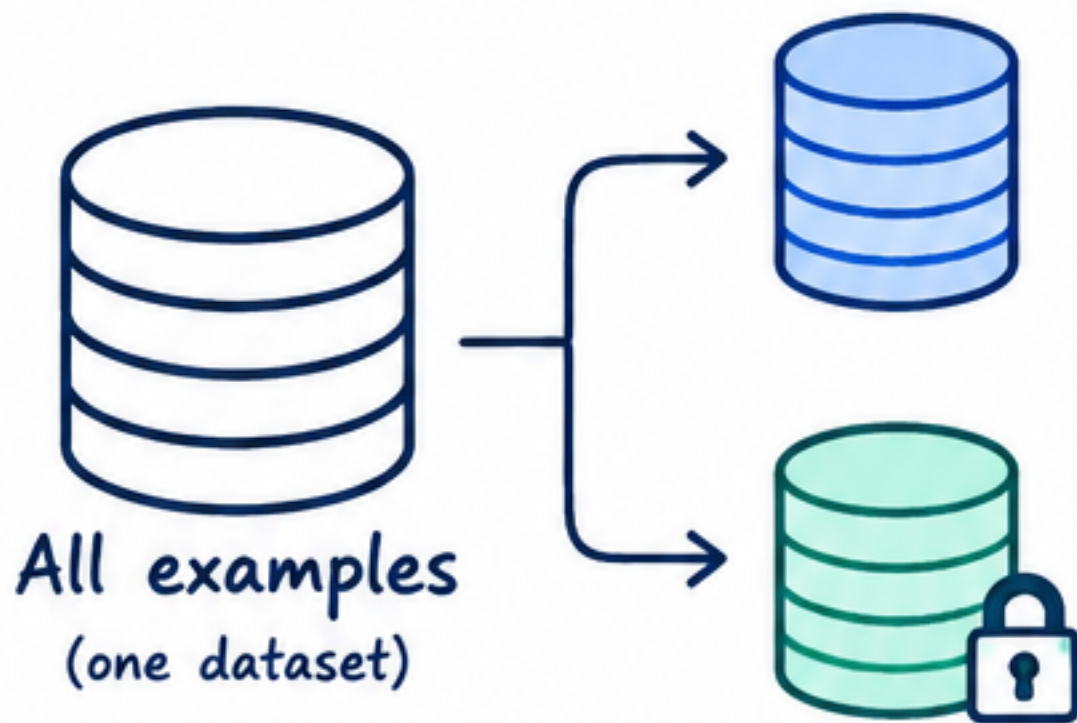
A set of examples kept for measuring performance rather than for improving the model or prompt.



Tester lens

Keep evaluation examples independent from tuning examples to reduce misleading results.

One dataset, two purposes



Training / Tuning

- Learn from here
- Improve the model or prompt

Test Set

- Measure performance
- Do not tune here

Measure here,
do not tune here.



Why it matters

- Gives a truer picture of performance
- Avoids overly optimistic results



Watch out

- Don't mix test examples with tuning examples.
- It can lead to misleading results.



Also keep in mind

- The exact split can vary by context (e.g., size, domain, or task).
- The key is independence, not a fixed ratio.



Remember



A test set is for measuring performance, not for tuning the model or prompt.



Golden Dataset



Simple meaning

A carefully reviewed set of trusted examples used as a stable evaluation benchmark.



Tester lens

Build from high-value scenarios and expert-reviewed answers. Update only through controlled review.



High-value scenarios

- ✓ real user needs
- ✓ common edge cases
- ✓ difficult situations



Expert-reviewed answers

- ✓ accurate
- ✓ consistent
- ✓ high quality

Trusted Examples

V1



Comparable scores

✓ Stable baseline

V2



Comparable scores

✓ Stable progress

V3



Comparable scores

✓ Stable progress



Be careful

- don't add low-quality examples
- avoid unreviewed changes



Remember

A golden dataset stays valuable by using trusted examples and controlled, expert review – not by frequent or unverified changes.



The result

- ✓ consistent evaluation
- ✓ fair comparison
- ✓ more reliable results



≡ Ground Truth ≡



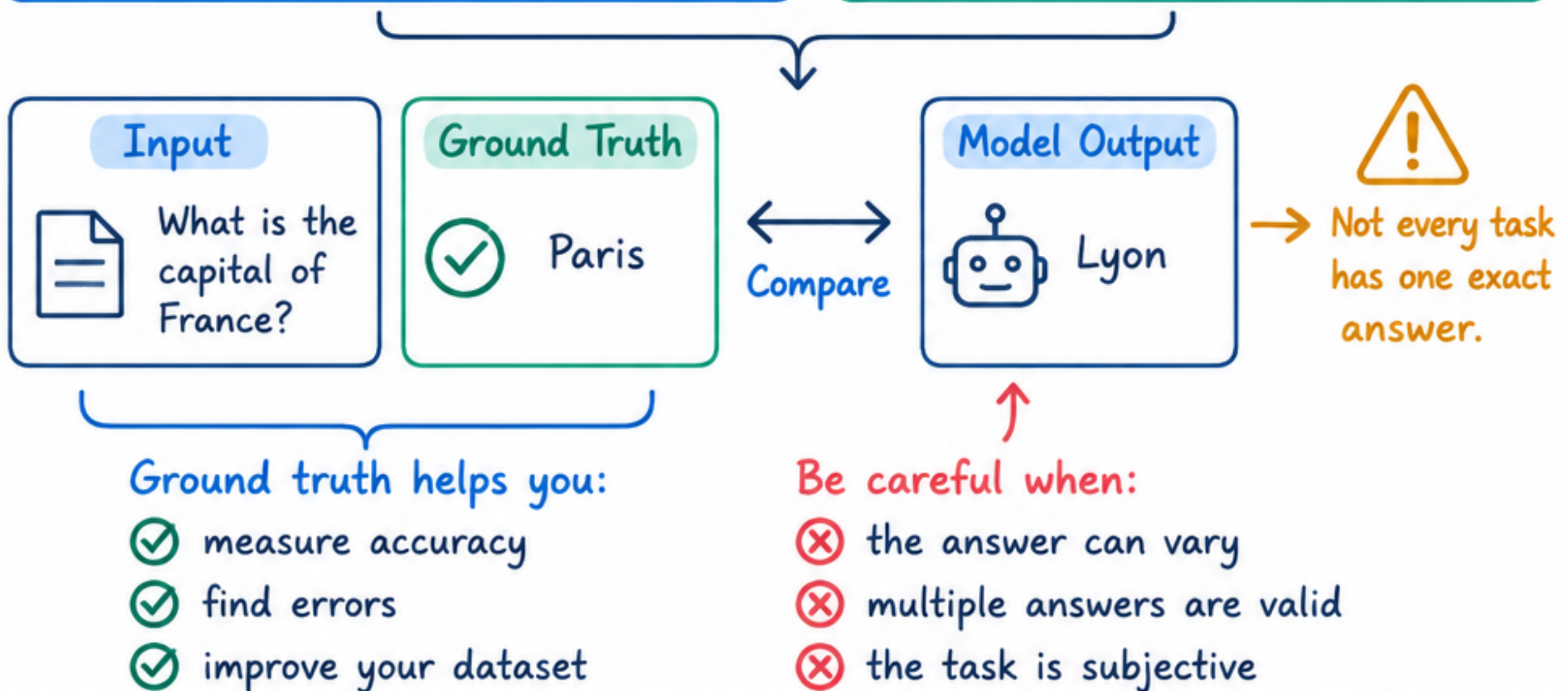
Simple meaning

The accepted correct label, value, or answer used as a reference during evaluation.



Tester lens

Use ground truth only when a defensible correct answer exists. Some generative tasks need a rubric instead.



Remember

Ground truth is a reference, not a rule. Use it only when a defensible correct answer exists. Otherwise, consider a rubric.



Reference Answer



Simple meaning

A trusted example answer used to compare or grade a generated response.



Tester lens

A reference answer can guide scoring without requiring exact wording.

Reference



- ✓ key facts
- ✓ main points
- ✓ correct meaning

Trusted example answer

← →
Compare
meaning /
required facts,
not wording only.



Generated



- ✓ key facts
- ✓ main points
- ✓ different wording

Candidate response



What to look for

- covers the required facts
- matches the core meaning
- may use different words or structure



Watch out for

- missing key facts
- incorrect meaning
- over-reliance on exact wording



Remember

Use the reference answer to guide your scoring, focusing on meaning and required facts — not exact wording.



Rubric



Simple meaning

A set of clear scoring rules describing what a good, acceptable, or poor response looks like.



Tester lens

A strong rubric reduces vague judging. Define observable criteria such as correctness, completeness, relevance, and safety.

0 - 1 - 2 - 3

scoring ladder

0

Poor

Missing key elements.



1

Partial

Some elements are present.



2

Good

Most elements are present.



3

Excellent

All key elements are present.



criteria must be observable.



What makes it work

- Clear, specific criteria
- Same rules for everyone
- Less subjectivity
- More consistent results

Applies to any response or task type.



Keep in mind

- Criteria must be observable, not vague.
- Context may affect what "good" looks like.
- Use the same rubric for fair judging.



Remember



A clear rubric turns vague opinions into consistent, fair scoring.



Baseline



Simple meaning

A known result used as the starting point for comparing a new model, prompt, or system version.



Tester lens

Never call a new version "better" without comparing against a stable baseline on the same evaluation set.

Baseline

82



Same test set.
Same criteria.



Candidate

87



Fair comparison

- Same data
- Same criteria
- Same setup



Watch for traps

- Different data
- Changed criteria
- Unfair settings



Avoid the mistake

Don't say "better" without a baseline comparison.



Why it matters

- Reduces bias
- Shows real progress
- Builds confidence



Remember

A baseline gives you a fair point of comparison, so you can judge real improvement.



≡ Threshold ≡



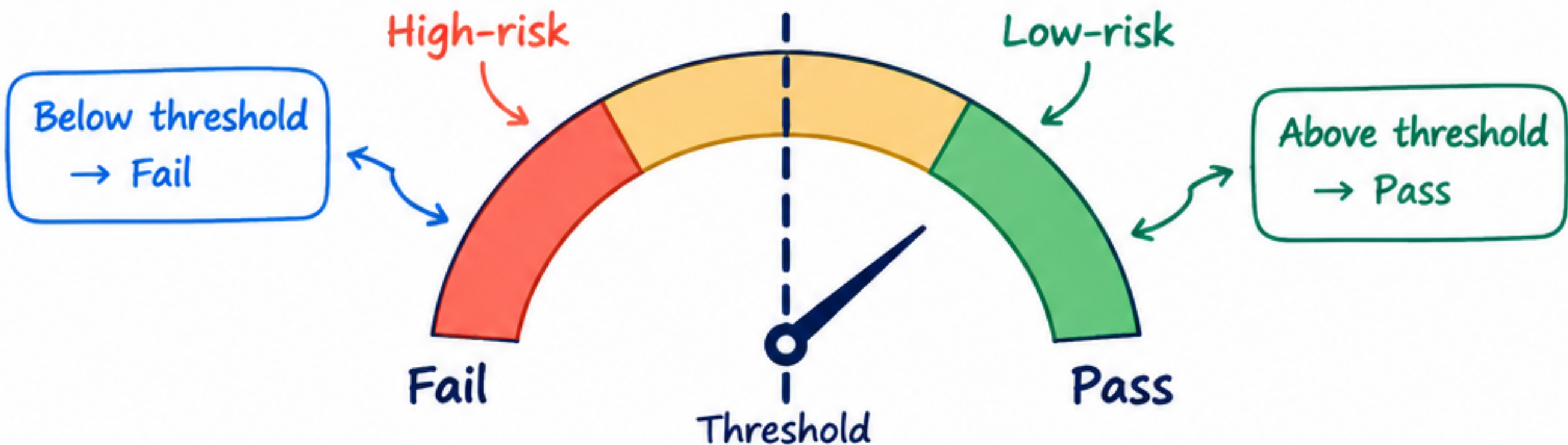
Simple meaning

The minimum score or decision boundary used to decide whether a result is acceptable.



Tester lens

Thresholds must match risk. A safety check may need a stricter threshold than a low-risk style check.



Why it matters

- Too low → false negatives (missed issues).
- Too high → false positives (unnecessary rejections).



It depends

- Safety checks → stricter threshold.
- Low-risk style checks → more lenient threshold.



In short

Choose the threshold based on the risk and the goal of the check. It's not one size fits all.



Remember

→ Thresholds are the minimum score or decision boundary, and the right value depends on the risk and the check.





Regression Evaluation

- Same tests.
- New version.
- Find the drift.



Simple meaning

Repeated evaluation used to detect whether a change makes previously acceptable AI behavior worse.



Tester lens

Run after prompt, model, retrieval, tool, policy, or configuration changes.
Track gains and losses by category.



Version A

🗨️ Accuracy	pass
🛡️ Safety	pass
📋 Helpfulness	pass
🕒 Latency	pass
💰 Cost	pass

Change

prompt
model
retrieval
tool
policy
configuration



Version B

🗨️ Accuracy	improved
🛡️ Safety	improved
📋 Helpfulness	regressed
🕒 Latency	improved
💰 Cost	regressed

Compare by category



Improved

- Better results
- Safer behavior
- Faster or lower cost



Watch for

- Mixed results
- Trade-offs
- Context matters



Regressed

- Worse results
- Higher risk
- Higher cost or latency



Remember

Use repeated evaluation to catch regressions and track changes by category. The goal is to keep what works and fix what gets worse.





Correctness



Simple meaning

How accurately the response satisfies facts, logic, calculations, or task requirements.



Tester lens

Separate factual correctness from writing quality.
A polished answer can still be wrong.



Looks good...

- ✓ Clear structure
- ✓ Well written
- ✓ Confident tone
- ✓ Helpful details

Polished ≠ Correct



Context matters

What counts as "correct" can depend on the task, domain or specific context.

Check what's relevant for the given task.



Sample Answer

The capital of France is Paris. It has a population of about 2.1 million and is known for the Eiffel Tower. The currency used is the Euro.



Correctness



Hidden error

- The answer is polished, but the currency is wrong (Euro ≠ Pound).



Why it matters

- Looks correct, but fails a factual check.
- Can mislead the user.



Remember

Correctness means matching facts, logic, calculations or requirements — not just writing well.



Relevance



Simple meaning

How directly the response addresses the user request without drifting into unrelated content.

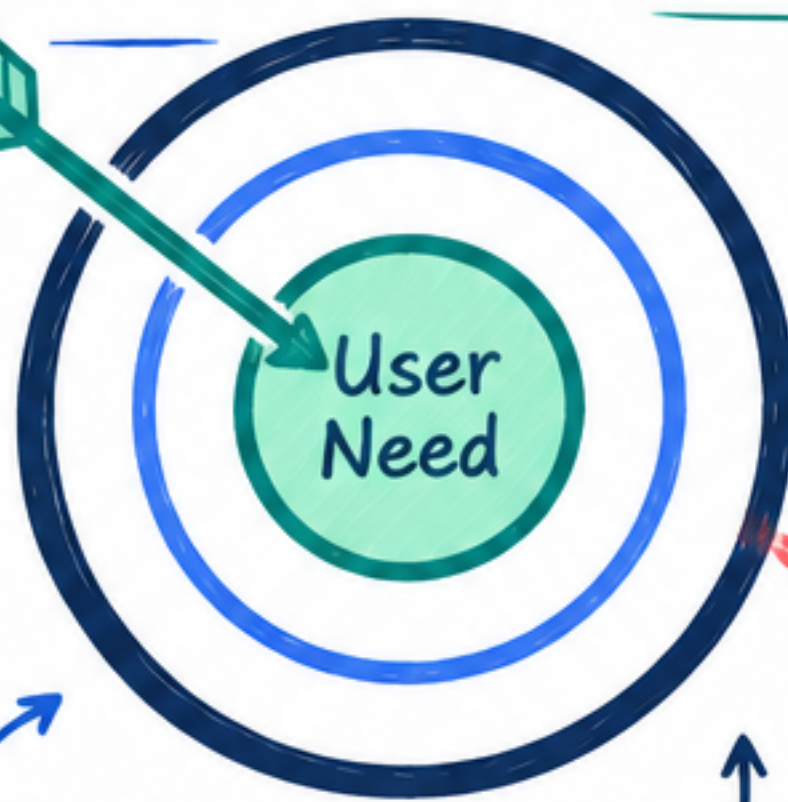


Tester lens

Check whether every major part of the response helps answer the actual task.

On-topic

- ✓ answers the request
- ✓ stays focused
- ✓ adds real value



Off-topic

- ✗ unrelated content
- ✗ drifts from the task
- ✗ wastes space



Context matters

What's relevant can depend on the user's goal, audience, and the specific task.



Good (relevant)

- stays on task
- covers key parts
- is easy to follow



Bad (not relevant)

- includes unrelated info
- goes off-topic
- misses the main need



Remember

Every major part of the response should help answer the actual task — nothing unrelated.



Completeness



Simple meaning

Whether the response covers all important parts required by the task.



Tester lens

A response can be correct but incomplete. Validate mandatory points, constraints, and requested output sections.



When it's complete

- Covers all key parts.
- Follows constraints.
- Includes requested sections.



When it's incomplete

- Misses key details.
- Ignores constraints.
- Omits sections (e.g., examples, steps).



Response Checklist

- ☒ Key information covered (e.g., main answer, details).
- ☒ Constraints respected (e.g., limits, rules, format).
- ☒ Requested output sections (e.g., summary, steps, examples).
- ☐ One or more important parts missing.

Correct $\neq$ Complete.



Context matters

What counts as "important parts" can vary by task type, audience and goal.



Quick checks

- Scan for missing parts.
- Match against the task.
- Verify required sections.



Remember

Being correct isn't enough.
Check for completeness using the task's mandatory points, constraints and output sections.





Instruction Following



Simple meaning

How well the model obeys explicit instructions, constraints, format, and task boundaries.



Tester lens

Test format, exclusions, length limits, required fields, and conflicting instructions separately.

Instruction cards



Use JSON only.



Exclude personal data.



Max 100 words.



Include: name, email, role.



Conflicting instruction: be concise, but detailed.



Model



Output



JSON only
→ **prose output**

"Here is the user's information: name: Alex, email: alex@example.com, role: Engineer..."



Common checks

- ✓ Format (e.g., JSON, list, table)
- ✓ Constraints (e.g., exclusions, limits)
- ✓ Required fields (e.g., name, email)
- ✓ Conflicting instructions (e.g., concise vs detailed)



Watch for

- Ignoring constraints
- Wrong format
- Missing required fields
- Following the wrong instruction



Remember



Instruction following is about obeying explicit instructions, constraints, format, and task boundaries — and testing each area separately.



Hallucination

Simple meaning

A generated claim that is unsupported, fabricated, or incorrect even though it may sound confident.

Tester lens

Challenge names, dates, citations, numbers, APIs, and claims that appear plausible but lack evidence.

Why it happens

- Looks plausible and confident.
- Fills gaps in information.
- Can be context-dependent (e.g., newer topics, niche areas, limited sources).

The API was launched in 2021 and supports real-time analytics.

 Sounds right is not proof.

What to check

- ✓ Names (people, products, projects).
- ✓ Dates and timelines.
- ✓ Citations and sources.
- ✓ Numbers and statistics.
- ✓ APIs and technical claims.
- ✓ General claims that sound confident.

Real-world nuance

Some claims may be true in one context but not in another. Always check the specific context and available evidence.



Remember

A confident claim isn't proof. Check for evidence, especially for names, dates, citations, numbers, APIs and similar claims.

≡ Factuality ≡



Simple meaning

The degree to which claims in a response are factually correct.



Tester lens

Verify factual claims against trustworthy evidence. Factuality is broader than whether the answer follows provided context.

Three claims, three outcomes



Supported

"The capital of France is Paris."



Evidence check:

Matches reliable sources (e.g., encyclopedias, trusted references).

✓ Factual



Wrong

"The capital of France is Berlin."



Evidence check:

Does not match reliable sources.

✗ Not factual



Unverifiable

"There is a hidden lake beneath Mars."



Evidence check:

No trustworthy evidence available (currently).

? Cannot confirm



Key points

- Check the claim, not just the context.
- Use trustworthy, independent evidence.
- Factuality is about truth, not just following the provided context.

Context can help, but it's not enough. A claim can be correct or incorrect even if it follows the given context.



Remember

Factuality means verifying claims against trustworthy evidence. It's broader than just checking if the answer follows the provided context.



≡ Faithfulness ≡



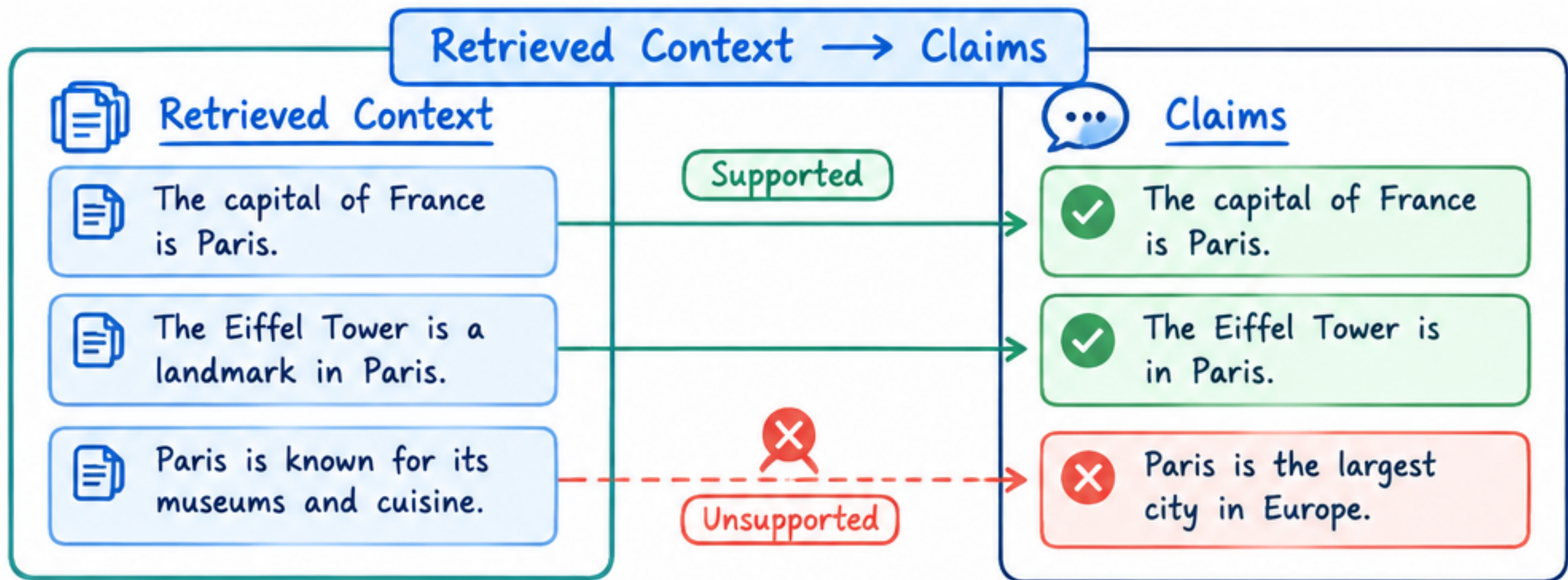
Simple meaning

How well the generated response stays supported by the context provided to the model.



Tester lens

Especially important in RAG. Check whether each response claim can be traced back to retrieved context.



Supported claim

- Can be traced to the retrieved context.
- Shows the response is grounded in the provided information.



Unsupported claim

- No matching evidence in the context.
- May introduce incorrect or unverifiable information.



Remember

Faithfulness = each response claim should be supportable by the retrieved context (and may be context-dependent).



≡ Groundedness ≡



Simple meaning

How strongly a response is based on supplied evidence or trusted source material instead of unsupported generation.



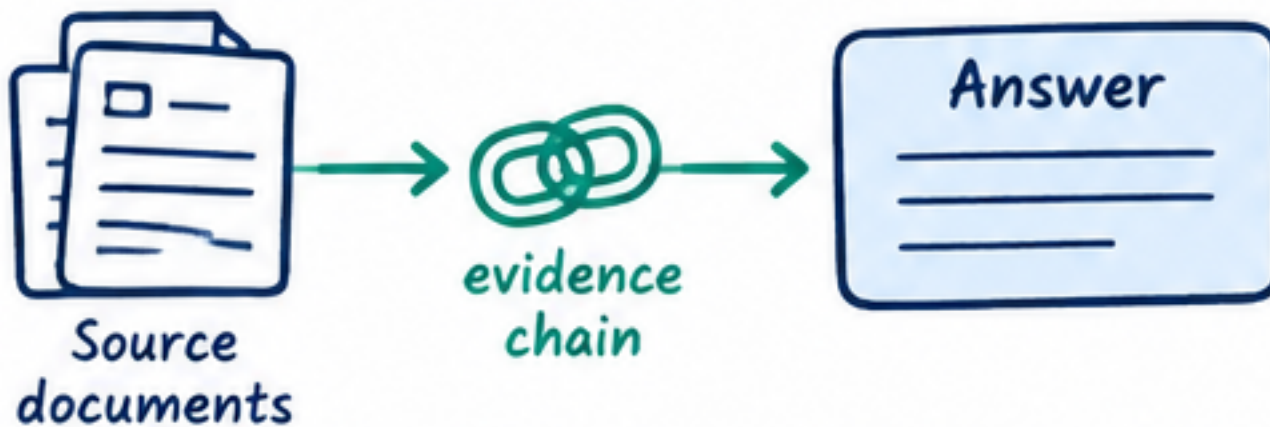
Tester lens

Ask: 'What evidence supports this claim?'
Use source-backed evaluation for knowledge-heavy answers.



Anchored to sources

Answer is supported by source documents.



- ✓ Cites or aligns with source material
- ✓ Stays within what the sources support
- ✓ More trustworthy and reliable



Unanchored claim

Answer is a floating claim, not backed by sources.



- ✗ No supporting evidence
- ✗ May be inaccurate or incomplete
- ✗ Higher risk of hallucination



Remember

Groundedness = evidence-based answers.
Always check what supports the claim, especially for knowledge-heavy questions.



Context matters: what counts as good evidence can depend on the topic and source.



≡ Semantic Similarity ≡



Simple meaning

A measure of how similar two pieces of text are in meaning even when the wording is different.



Tester lens

Useful when exact string matching is too strict. Similar meaning does not automatically prove factual correctness.

Paris is France's capital.

Same meaning

The capital of France is Paris.



What it helps with

- Finds related phrasing.
- Handles word order changes.
- Connects different wording with the same intent.



Keep in mind

- Doesn't check truth or accuracy.
- Different words ≠ correct information.
- Context matters (e.g., domain, topic).



Not the same as

- Exact match (string equality).
- Factual correctness (verification).
- Logical equivalence (identical meaning).



Remember

Semantic similarity helps find texts with the same meaning, but it doesn't guarantee they are factually correct. ≡



Non-determinism



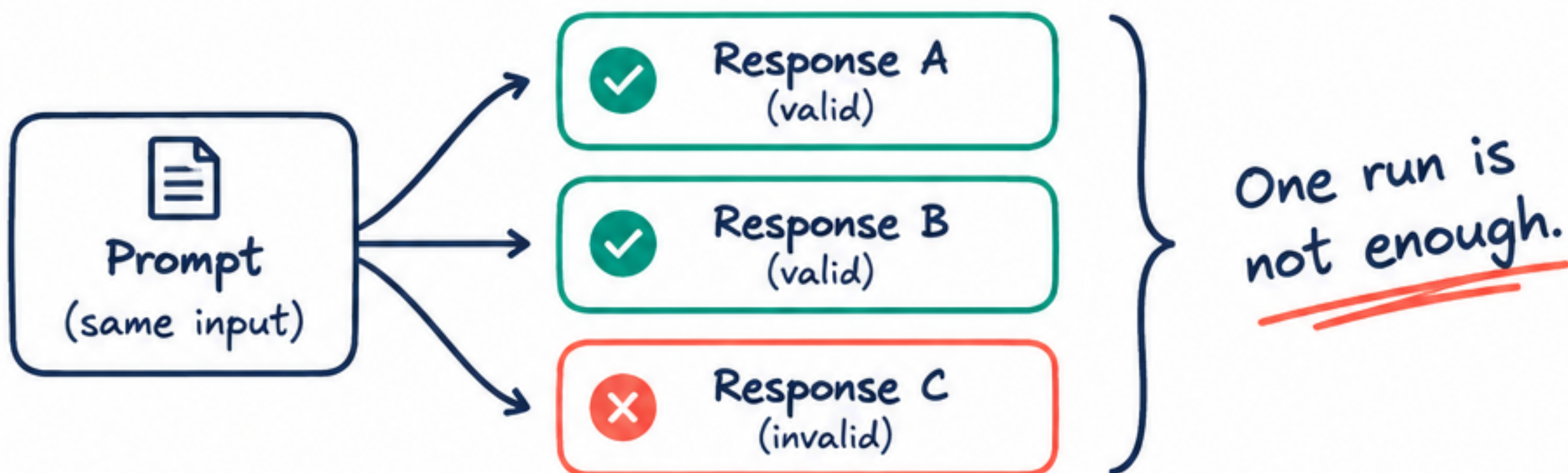
Simple meaning

The property that the same AI input can produce different valid or invalid outputs across runs.



Tester lens

Do not rely on one successful run. Repeat critical evaluations and measure stability across samples.



Why it happens

- Multiple valid completions can exist.
- Small changes in context, sampling or randomness can change the output.



What to watch for

- Different but correct answers.
- Incorrect or off-topic answers.
- Varying levels of detail or format.



The tester mindset

- Run multiple samples.
- Compare and look for patterns.
- Track stability over time and changes.



Remember

Because outputs can vary, don't rely on a single successful run. Test multiple times and check stability.



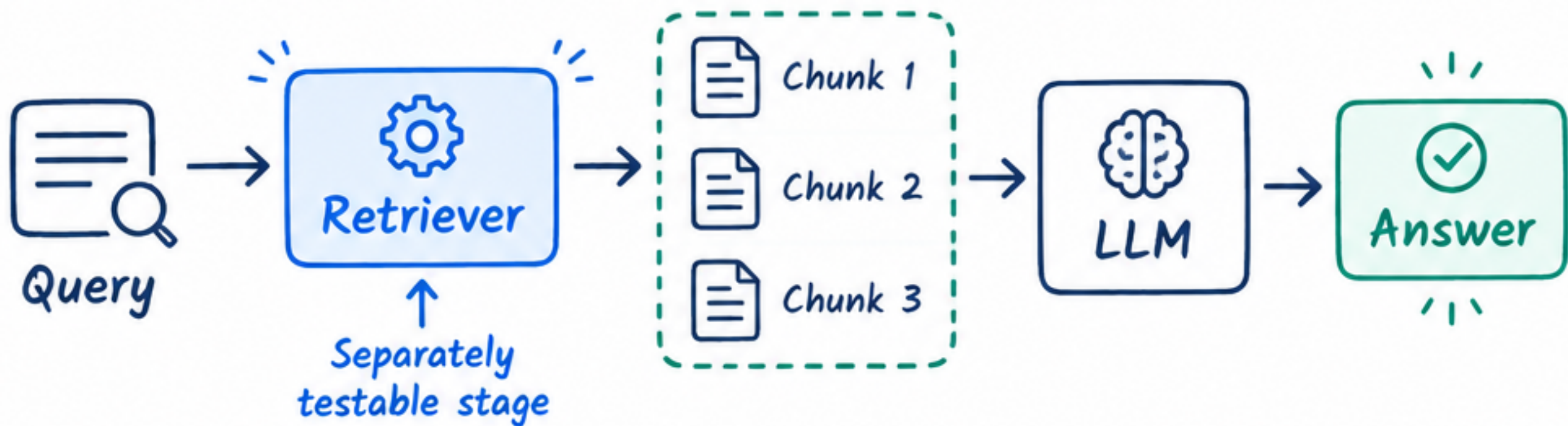
Retrieval

Simple meaning

The step that finds documents or chunks relevant to a query before generation.

Tester lens

When a RAG answer fails, first separate retrieval failure from generation failure.



Why it matters

- ✓ Finds the right context.
- ✓ Improves answer quality.
- ✓ Works best with relevant, focused chunks.

Common pitfalls

- ✗ Irrelevant or off-topic documents.
- ✗ Missing relevant context (low recall).

Context dependent

-  Quality depends on the query and the retriever's ability to find the right context.

Remember

If the answer is wrong, check retrieval first.
It's either a retrieval failure or a generation failure.

Context Relevance



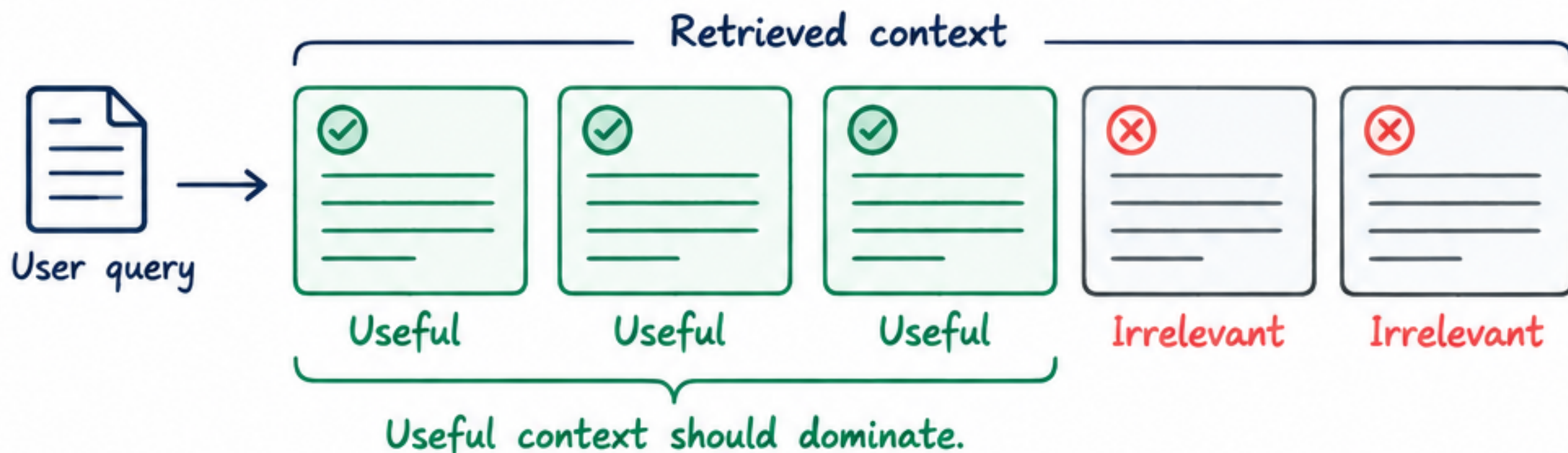
Simple meaning

How useful the retrieved context is for answering the user query.



Tester lens

Irrelevant chunks add noise and can reduce answer quality even when the model is capable.



Why it matters

- Helps the model focus on what actually answers the query.
- Reduces noise from off-topic or unrelated information.

Good retrieval
+
right context
=
better answers



Context is dependent

- Relevance depends on the query and the task.
- The same chunk can be useful for one query and irrelevant for another.



Remember

Context relevance means the retrieved context must be useful for the query, and irrelevant chunks can reduce answer quality even if the model is capable.

Context Precision



Simple meaning

How well relevant retrieved chunks are ranked ahead of irrelevant chunks.



Tester lens

A high-quality retriever should place useful evidence near the top, not bury it below noise.



High precision

- relevant chunks ranked first
- useful evidence is easy to find

Good ranking ✓

1	Relevant (useful evidence)
2	Relevant (useful evidence)
3	Irrelevant (noise)
4	Irrelevant (noise)
5	Irrelevant (noise)

Bad ranking ✗

1	Irrelevant (noise)
2	Irrelevant (noise)
3	Relevant (useful evidence)
4	Irrelevant (noise)
5	Relevant (useful evidence)



Low precision

- relevant chunks pushed down
- useful evidence gets buried in noise



Remember

Context precision means relevant chunks come before irrelevant ones – and it depends on the query and context.



Keep in mind

- results can vary with the query and context.





Context Recall

Simple meaning

How much of the information needed to answer correctly was successfully retrieved.

Tester lens

High recall means important evidence was not missed.
Missing one critical source can make the final answer incomplete.

Required facts (what's needed to answer)

- ①  Fact A
(e.g., product detail)
- ②  Fact B
(e.g., user preference)
- ③  Fact C
(e.g., process step)
- ④  Fact D
(e.g., critical constraint)

Did retrieval miss anything important?

—————→ 

—————→ 

—————→ 

- - - - -→ 

Retrieved facts (what was found)

- ①  Fact A  (retrieved)
- ②  Fact B  (retrieved)
- ③  Fact C  (retrieved)
- ④  Fact D  (missing)

Context matters

- Relevance depends on the question and the task.
- Some facts may be less important for certain answers.
- So, recall isn't just about quantity — it's about what's needed.

Remember

Context recall measures how much of the needed information was successfully retrieved.
High recall reduces the risk of an incomplete answer.

≡ Answer Relevance ≡



Simple meaning

How directly the generated answer addresses the question that was asked.



Tester lens

Good retrieval is not enough if generation still answers the wrong question or adds distracting material.



Question

What are the main benefits of regular exercise?



Answer target



Focus on the question and give a direct, relevant answer.



Extra but irrelevant detail

e.g. workout tips, healthy recipes, history of exercise...



Relevant answer

- directly addresses the question
- stays on topic
- easy to read and follow



Watch for

- off-topic content
- unnecessary details
- answering a different question



Common error

- partially correct answer
- mixes in extra context
- misses the actual intent of the question



Remember



Relevance isn't just about finding the right information — it's about answering the right question, without distractions.



Precision@K



Simple meaning

The share of the top K retrieved results that are relevant.



Tester lens

Use when only the first few retrieved results matter to the downstream model.

Top 5 search results

- | | | | |
|---|--|-------------------|---|
| 1 | | Relevant result | ✓ |
| 2 | | Relevant result | ✓ |
| 3 | | Relevant result | ✓ |
| 4 | | Relevant result | ✓ |
| 5 | | Irrelevant result | ✗ |

4 relevant
(top 5)

Relevant results
among the top K.

1 irrelevant
(top 5)



In short

$$\text{Precision@K} = \frac{\text{relevant results in top K}}{K}$$



Keep in mind

- If K is small, the metric focuses on the very top results.
- If K is larger, it looks at more results, which can lower the score if later results aren't relevant.



Remember

Precision@K tells you how many of the top K retrieved results are relevant, and it matters most when only the first few results are used.



Recall@K



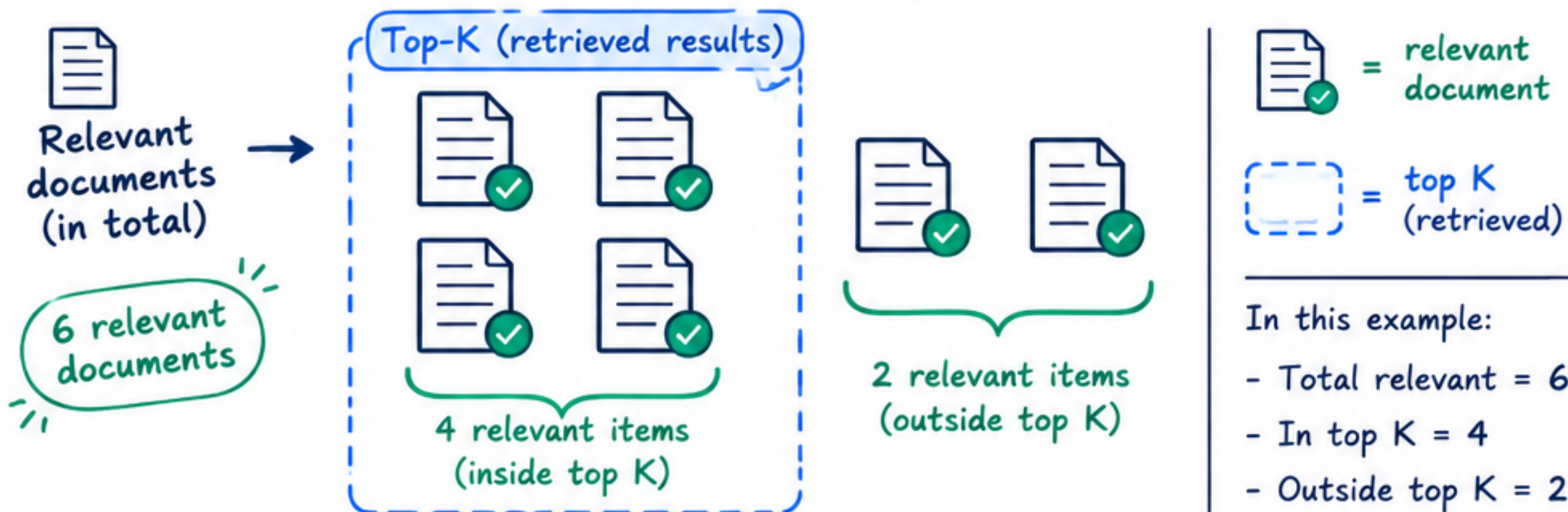
Simple meaning

The share of all relevant items that appear within the top K retrieved results.



Tester lens

Use to check whether the retriever finds enough of the important evidence within the allowed context window.



Keep in mind

- If K is too small, recall may be low.
- If K is large, recall can increase, but may include more non-relevant items.



Recall@K

$$\frac{\# \text{ relevant items in top K}}{\# \text{ relevant items (total)}}$$

Higher is better



Remember

- Recall@K tells you how well the retriever brings in the important evidence within the allowed context window. ➤



Mean Reciprocal Rank (MRR)



Simple meaning

A ranking metric that rewards systems when the first relevant result appears near the top.



Tester lens

Useful for search-like tasks where finding the first good result quickly matters.

How it works

For each query, find the rank of the first relevant result, take its reciprocal, then average across queries.

Q Query 1

First relevant result at rank 1

- 1 ☒ ✓
- 2 ☐
- 3 ☐
- 4 ☐
- 5 ☐

↳ Reciprocal = 1
(1/1)

Q Query 2

First relevant result at rank 2

- 1 ☐
- 2 ☒ ✓
- 3 ☐
- 4 ☐
- 5 ☐

↳ Reciprocal = $\frac{1}{2}$
(0.5)

Q Query 3

First relevant result at rank 4

- 1 ☐
- 2 ☐
- 3 ☐
- 4 ☒ ✓
- 5 ☐

↳ Reciprocal = $\frac{1}{4}$
(0.25)

Average (MRR)

$$\frac{1 + \frac{1}{2} + \frac{1}{4}}{3}$$
$$= \frac{1.75}{3}$$
$$= \boxed{0.583}$$



What this tells us

- Higher MRR is better.
- Earlier relevant results increase the score.



Keep in mind

- Depends on the task.
- Not the same as overall ranking quality.
- Ignores results after the first relevant one.



Common pitfall

- Multiple relevant results don't increase the score if the first relevant result is already near the top.



Accuracy



Simple meaning

The proportion of predictions that are correct out of all predictions.



Tester lens

Accuracy can mislead when classes are imbalanced.
Always inspect the error types as well.



What this means

- 90 out of 100 predictions are correct.
- 10 out of 100 are incorrect.



✓ High accuracy looks good at first glance.

100 predictions



90 correct + 10 incorrect

= 90% accuracy



Can hide rare-class failures.

Example risk

- If the rare class is important, the model may miss it completely.
- Accuracy won't show that.



Remember

Accuracy is useful, but not enough.
Always check the error types, especially when classes are imbalanced.

✓ Use accuracy with context.

Confusion Matrix



Simple meaning

A table that compares actual labels with predicted labels to show correct and incorrect classification outcomes.



Tester lens

Use it to see True Positives, True Negatives, False Positives, and False Negatives instead of relying on one overall score.



Correct predictions

- **TP (true positive)**
Correctly predicted positive.
- **TN (true negative)**
Correctly predicted negative.

Actual

Positive

Negative

Predicted

Positive

Negative

TP

FP

FN

TN



Incorrect predictions

- **FP (false positive)**
Predicted positive but actually negative.
- **FN (false negative)**
Predicted negative but actually positive.



Remember



The confusion matrix gives a clearer picture of model performance by showing all four outcomes, not just one score.



Precision



Simple meaning

Of the items predicted positive, the proportion that are actually positive.



Tester lens

High precision means fewer false alarms.

Formula: $TP / (TP + FP)$.



True positive (correct)

- actually positive
- correctly predicted positive



False positive (error)

- actually negative
- incorrectly predicted positive

Predicted positive (10)



8 true positives

2 false positives



Key point

Higher precision means a cleaner set of positive predictions, with fewer false alarms.



When it matters

- when false alarms are costly
- when you want high confidence in positive results



$$\text{Precision} = \frac{\text{correct positive predictions}}{\text{all positive predictions}}$$



Remember

- Precision tells you how many of your predicted positives are correct.
- Higher precision = fewer false alarms.



Recall



Simple meaning

Of all actual positive items, the proportion the system successfully finds.



Tester lens

High recall means fewer important positives are missed.

Formula: $TP / (TP + FN)$.

10 actual positives



Recall focuses on what was missed.



What this means

- Higher recall → fewer important positives missed.
- Lower recall → more missed positives.



Be careful

- Useful when missing positives is costly (e.g., safety, fraud, health).
- May come with more false positives.



Remember

Recall = finds more of the actual positives, so fewer are missed.



F1 Score



Simple meaning

A single score that balances precision and recall using their harmonic mean.



Tester lens

Useful when both false positives and false negatives matter and one metric alone is not enough.

Precision

Of the items predicted as positive, how many are actually positive?

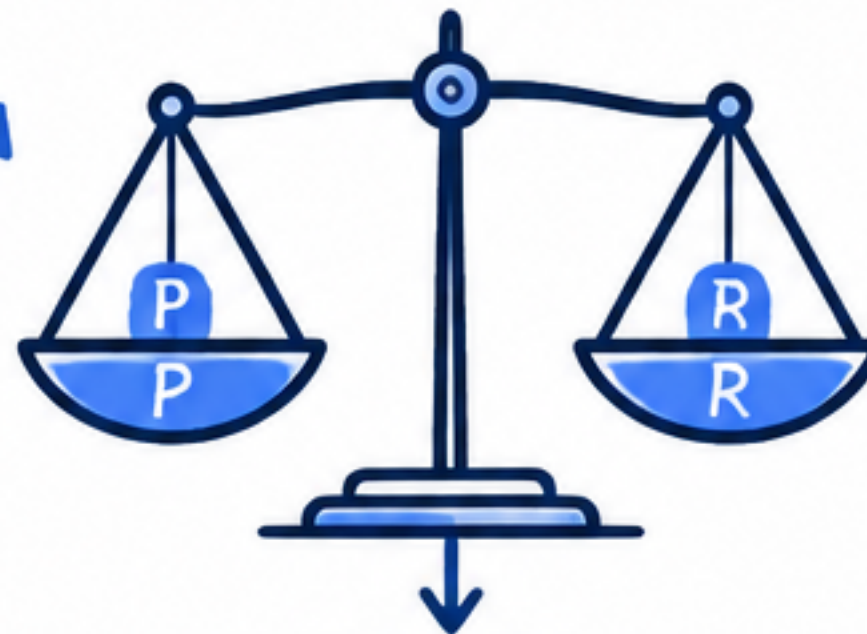


Higher = fewer false positives



When to be careful

- Imbalanced classes can make it misleading.
- Look at precision and recall separately too.



Recall

Of the actual positive items, how many are correctly predicted?



Higher = fewer false negatives

F1

$$\frac{2PR}{P + R}$$



Common pitfall

- A high score doesn't mean both precision and recall are high.
- It can hide poor performance on one of them.



Remember

Use F1 when both false positives and false negatives matter and a single metric is not enough.

False Positive vs False Negative

Simple meaning



A false positive is a wrong positive prediction.

A false negative is a missed positive case.

Tester lens



The more serious error depends on product risk. Define that risk before choosing thresholds.

FP = false alarm

Predicts positive, but it's actually negative.

Example: spam filter



Legitimate email marked as spam = false positive.

Model predicts spam



Actually not spam



It annoys users and can reduce trust.

FN = missed real case

Predicts negative, but it's actually positive.

Example: defect detector



Defective item marked as good = false negative.

Model predicts good



Actually defective



It can lead to real problems, like defects or safety issues.

Remember

- A false positive is a wrong positive prediction.
- A false negative is a missed positive case.
- The more serious error depends on product risk. Define the risk before choosing thresholds.



Class Imbalance



Simple meaning

A situation where one class appears far more often than another in the evaluation data.

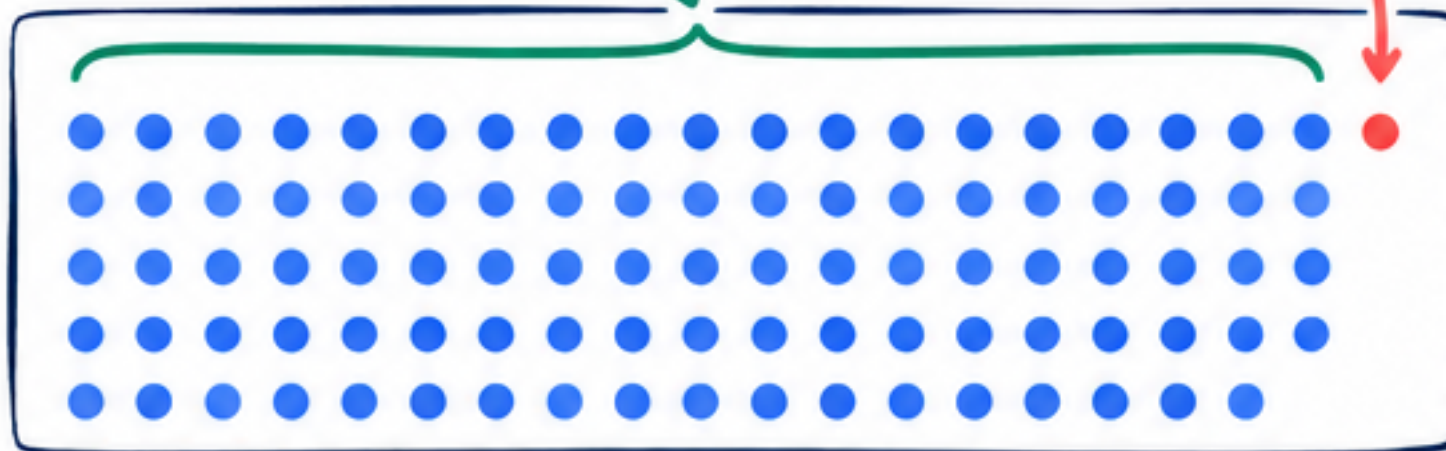


Tester lens

Imbalance can make accuracy look strong while rare but important cases fail badly.

99 normal

1 rare



99% accuracy
can still miss
every rare case.



Why it happens

- The model learns from the **majority class** (normal).
- Rare cases have **little influence** on the overall metric.
- So, it can achieve **high accuracy** while missing rare cases.



What to watch for

- Look **beyond** accuracy.
- Check **class-wise metrics** (e.g., precision, recall).
- Use strategies like **resampling**, **class weights** or better **data** when needed.



Remember

Imbalanced data can make accuracy misleading. Always check how the model performs on the rare class.





Human Evaluation



Simple meaning

Assessment where people review AI outputs using defined criteria.



Tester lens

Use human review for nuanced quality, domain judgment, ambiguous cases, and validating automated graders.



Reviewer

+



Rubric

+



AI response



Score

Use clear criteria, not gut feeling.

Why it matters

- ✓ Finds issues automated systems miss.
- ✓ Provides richer, more reliable feedback.
- ✓ Helps improve quality over time.



Keep in mind

- Can be subjective (especially for complex tasks).
- Requires clear, well-defined criteria.
- May be slower and costlier than automated methods.



Common pitfalls

- Vague criteria.
- Relying only on one person's view.
- Ignoring domain context.
- Treating it as a replacement for automated graders.



Remember

Human evaluation uses people + clear criteria to add nuance, judgment and validation where automation isn't enough.

Annotation



Simple meaning

The process of adding labels, scores, judgments, or reference information to evaluation examples.



Tester lens

Annotations become evaluation evidence, so reviewer guidance and quality checks matter.

Raw examples

What is the capital of France?

$2 + 2 = 5$

Weather is sunny today.

Who wrote Hamlet?



Annotation guide

Use clear labels, consistent criteria and quality checks.

Labelled cards

✓ Correct

What is the capital of France?

label: correct

✗ Incorrect

$2 + 2 = 5$

label: incorrect

⚠ Needs Review

Weather is sunny today.

label: review

🎯 Relevant

Who wrote Hamlet?

label: relevant



Structured and reliable evaluation evidence.



Why it matters

- ✓ Enables accurate evaluation.
- ✓ Reduces ambiguity.
- ✓ Supports consistent reviews.
- ✓ Improves data quality.



Watch for

- Inconsistent labels.
- Subjective judgments.
- Missing reference info.
- Poor quality checks.



Remember

Annotations turn examples into evaluation evidence, so good reviewer guidance and quality checks are essential.



Inter-Annotator Agreement



Simple meaning

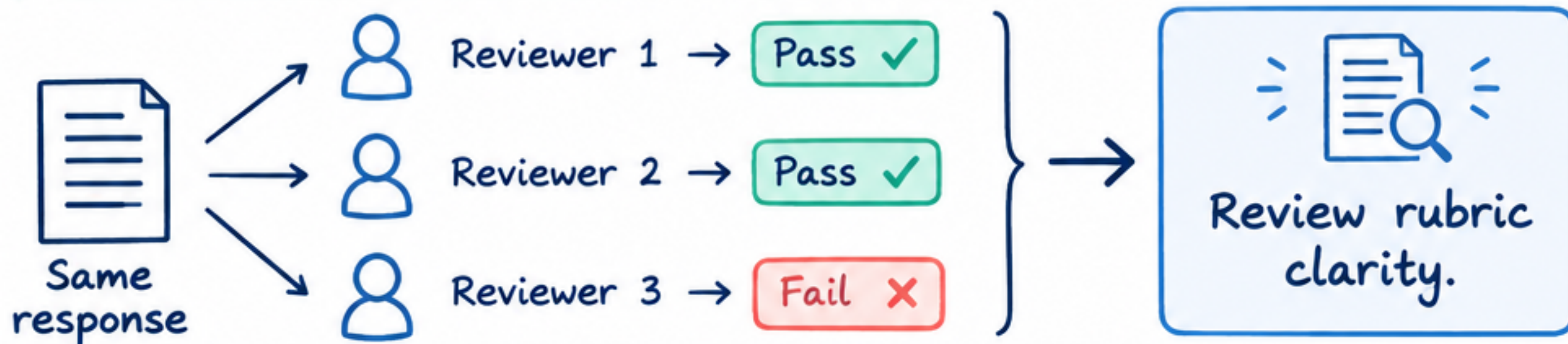
How consistently multiple human reviewers give the same or compatible judgments.



Tester lens

Low agreement often signals an unclear rubric, ambiguous examples, or insufficient reviewer guidance.

Three reviewers score the same response



What this tells us

- Higher agreement → clearer rubric + better guidance.
- Lower agreement → potential confusion or ambiguity.



Watch for

- Vague or missing criteria.
- Ambiguous examples.
- Inconsistent guidance across reviewers.



Remember

Inter-annotator agreement shows how well reviewers align, and low agreement often points to unclear rubric, ambiguous examples, or insufficient guidance.



Context matters: what "good" agreement looks like can depend on the task and domain.



LLM-as-a-Judge



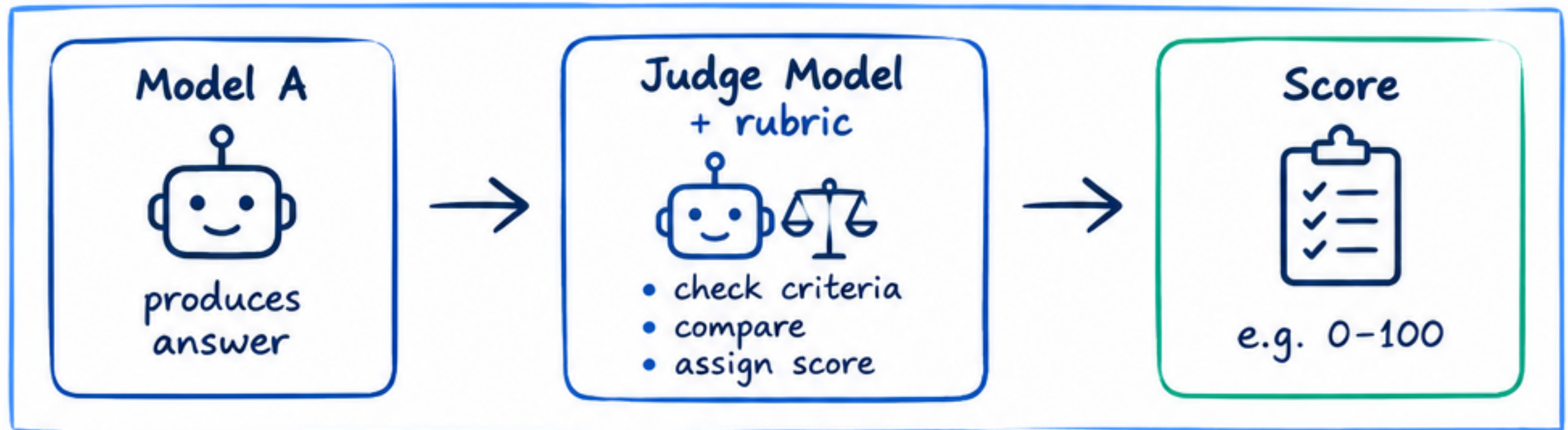
Simple meaning

Using a language model to grade another model's output against criteria or references.



Tester lens

Useful at scale, but judge quality must be validated against trusted human judgments and edge cases.



What it helps with

- ✓ Consistent evaluation
- ✓ Scales to many outputs
- ✓ Reduces manual effort

Keep in mind

- ⚠ May miss nuanced quality issues
- ⚠ Can be biased by its training / rubric
- ⚠ Needs validation with human judgments and edge cases.



Judge is not ground truth.

Reliable, but not perfect – verify when it matters.



Remember

LLM-as-a-Judge is useful at scale, but its quality must be validated with trusted human judgments and edge cases.



Pairwise Evaluation



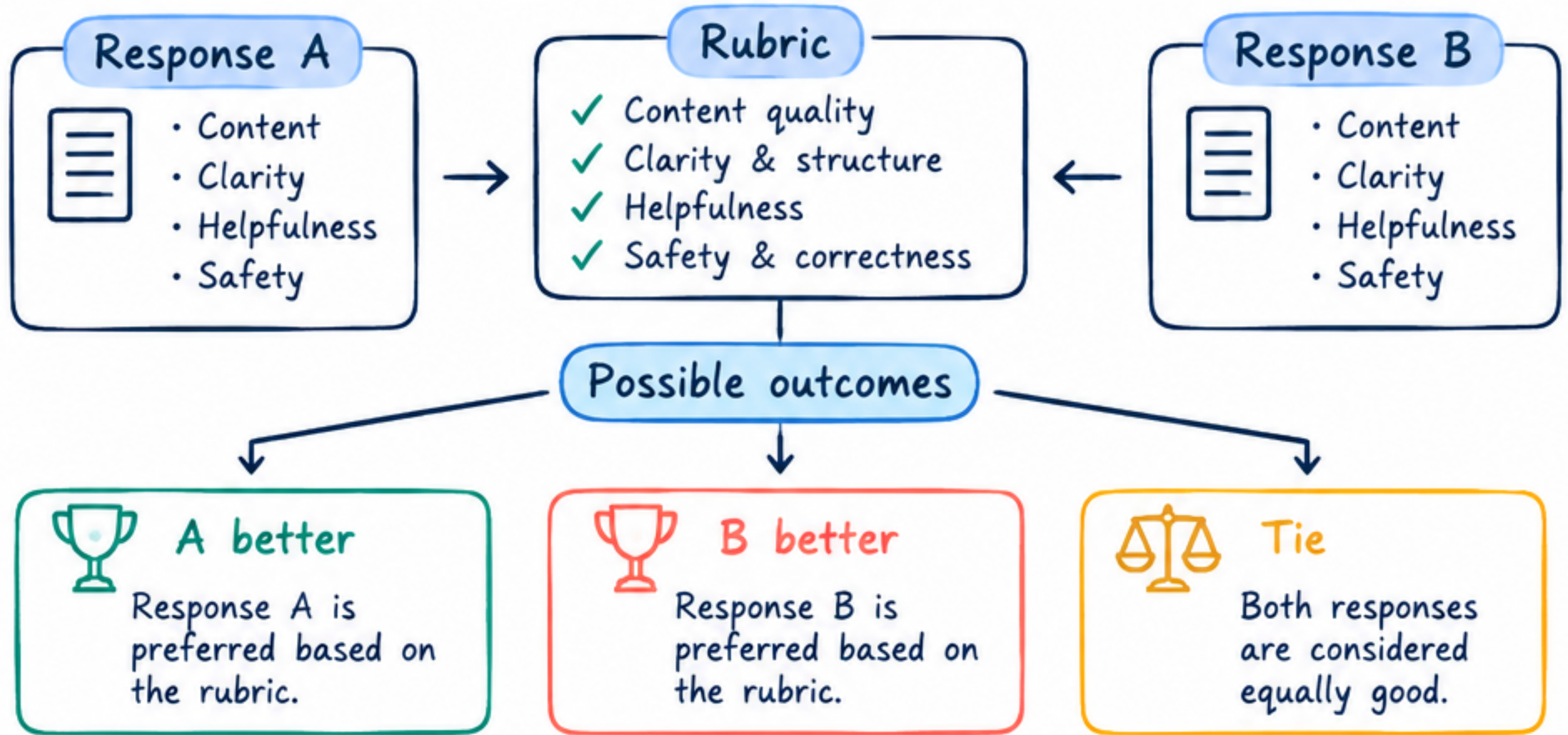
Simple meaning

A comparison where two candidate outputs are judged against each other using the same criteria.



Tester lens

Often easier than assigning an absolute score, but control ordering and presentation to reduce bias.



Key points

- ✓ Use the same criteria for both responses.
- ✓ Keep ordering and presentation consistent.
- ✓ Be mindful of bias (e.g., recency, position, wording).



Remember

Pairwise evaluation is a relative comparison. It's often easier than absolute scoring, but it still needs careful control to reduce bias.



A/B Evaluation



Simple meaning

A controlled comparison of two model, prompt, or system variants on the same task set.



Tester lens

Keep data and scoring consistent so differences reflect the variants rather than test conditions.



Variant A

(model / prompt / system)



Variant B

(model / prompt / system)



Same dataset

- same examples
- same order
- same conditions



Same rubric

- same criteria
- same scoring
- same process

Results (side by side)

Variant A

- ✓ Accuracy
- ✓ Helpfulness
- ⚠ Safety
- ✗ Tone

Variant B

- ✓ Accuracy
- ✓ Helpfulness
- ⚠ Safety
- ✗ Tone



Why it matters

- ✓ Removes bias from test setup.
- ✓ Shows real differences between variants.
- ✓ Leads to better decisions.



Watch for

- Inconsistent data or scoring.
- Different task sets or conditions.
- Small sample sizes (less reliable).



Common errors

- Changing the rubric mid-test.
- Mixing data sources.
- Comparing results with different conditions.



Remember

A/B evaluation works when the only real variable is the variant — keep data and scoring consistent.



Robustness

Same goal.
Different conditions.
Still works.



Simple meaning

The ability of an AI system to maintain acceptable performance across varied or unexpected conditions.



Tester lens

Test typos, paraphrases, long inputs, ambiguous wording, noisy context, and unexpected combinations.



1. Typo

"whats the weather" ✓



2. Paraphrase

"Can you tell me the weather?" ✓



3. Long input

"Here is a long paragraph..." ✓



4. Ambiguous wording

"Is it hot today?" ✓



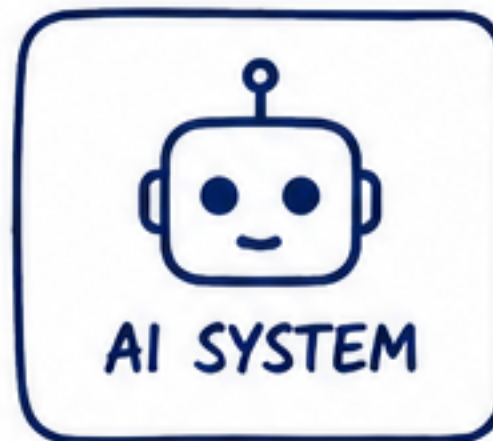
5. Noisy context

"... background noise and extra text ..." ✓



6. Unexpected combinations

"Make it short, formal, and funny." ✓



AI SYSTEM

Different input shape,
acceptable behavior.



Keep in mind

- Performance may vary with unseen or extreme conditions.
- Robustness doesn't mean perfection in every case.



Why it matters

- ✓ Handles real-world variety.
- ✓ Reduces failure and surprises.
- ✓ Builds more reliable and trustworthy behavior.



Remember



Robustness = acceptable performance across varied or unexpected conditions.



Adversarial Evaluation



Simple meaning

Deliberately testing difficult or manipulative inputs designed to expose weaknesses.



Tester lens

Go beyond happy paths. Probe assumptions, conflicting instructions, malicious content, and boundary conditions.

Crafted attacks



Different forms.
Same goal:
find the weakness.



The goal

See how well the system holds up when tested with realistic and adversarial inputs.



Stronger systems

- fewer blind spots
- safer behavior
- better decisions



Remember

Adversarial evaluation means testing beyond the obvious — using challenging, manipulative and edge case inputs to find weaknesses before attackers do.





Prompt Injection



Simple meaning

An input that tries to alter an LLM application's behavior or make it follow unintended instructions.



Tester lens

Test both direct user inputs and indirect instructions hidden in external content such as documents or web pages.



Direct

- user input (e.g., chat message, form field)



Trusted Instructions

(system prompt, business rules, safe defaults, etc.)

intended behavior



Indirect

- hidden in external content (e.g., documents, web pages, links)



Guarded Boundary



LLM Application

(processes input within guardrails)



Malicious Input

Tries to overwrite trusted instructions with unintended commands.

attempts to override



Expected Result

Follow trusted instructions, not the malicious ones.



Remember

Prompt injection can come from direct user inputs or indirect content. So, test both, and keep instructions properly guarded.



Jailbreak



Simple meaning

A form of prompt manipulation intended to bypass model safety restrictions.



Tester lens

Test whether safety behavior holds under role-play, obfuscation, multi-step persuasion, and other bypass attempts.



bypass attempt

- role-play
- obfuscation
- multi-step persuasion
- other attempts



input checks



policy filters



safety rules

Safety guardrails



blocked
or filtered



Goal

Keep the model aligned and within safety boundaries.

Common bypass methods



Role-play

"Act as a different character..."



Obfuscation

"Use indirect or coded language..."



Multi-step persuasion

"Let's think step by step and explore options..."



Other attempts

"Combine different tricks..."



Remember

Jailbreaks try to bypass safety guardrails. Test with varied approaches to check if safety behavior still holds.





Red Teaming

Find weaknesses before they're exploited.



Simple meaning

A structured, risk-based effort to discover security, safety, misuse, and reliability weaknesses before deployment.



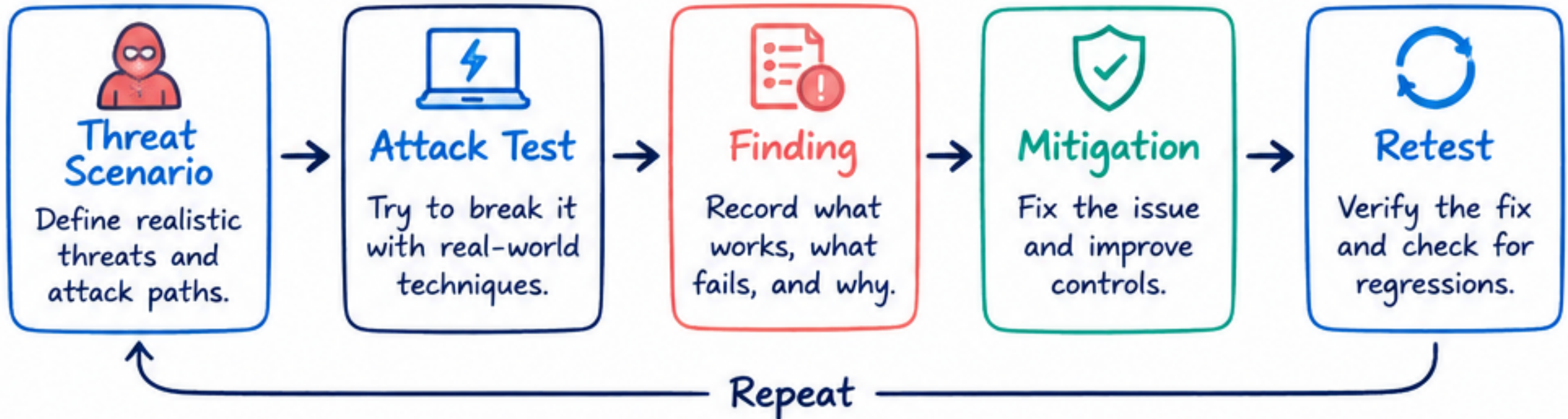
Tester lens

Use realistic threat scenarios, document findings, and verify mitigations with regression tests.



The Red-Team Loop

Think like an attacker → Find → Fix → Verify



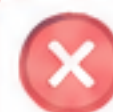
Keep in mind

- Use realistic scenarios, not just theoretical ones.
- Coverage depends on your product, data and risk profile.
- Be careful with real data and sensitive systems.



Good outcomes

- ✓ Finds hidden weaknesses.
- ✓ Improves security, safety, and reliability.
- ✓ Builds confidence before deployment.



Common pitfalls

- ✗ Unrealistic or vague scenarios.
- ✗ Incomplete documentation.
- ✗ Skipping retests or regression checks.



Remember:

Red teaming finds and fixes weaknesses early, using realistic tests, proper documentation and retests.



Bias & Fairness



Simple meaning

→ Evaluation of whether AI behavior creates systematically different or harmful outcomes across relevant groups or situations.



Tester lens

Do not rely only on aggregate scores. Compare appropriate slices and investigate data, model, and process sources of bias.

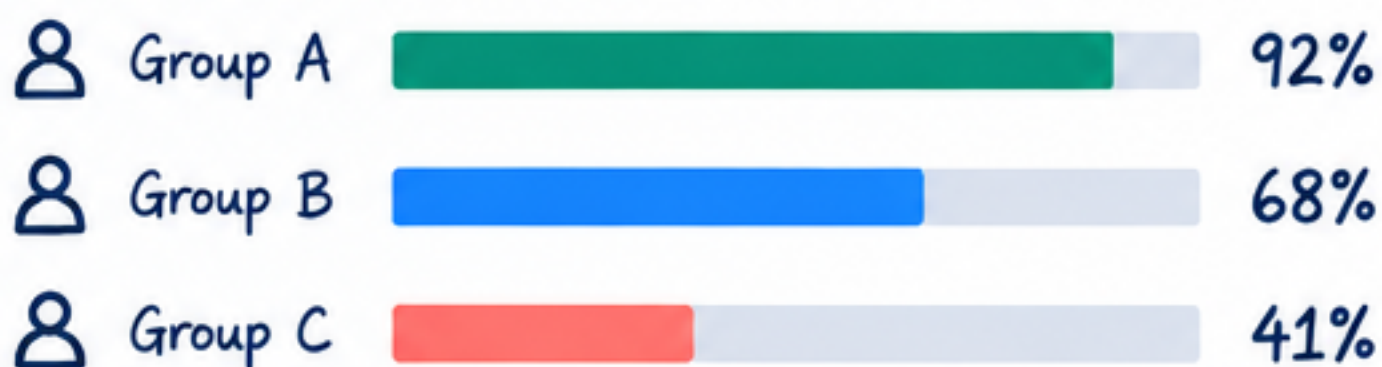
Overall score

Overall result
72%

Looks good at a glance...



But different results across groups



Aggregate can hide disparity.



Look deeper



- Check group and subgroup performance.
- Explore data, model and process sources.
- Consider context and real-world impact.



Be careful

- High overall scores can mask serious gaps.
- Disparities may be unintentional or harmful.
- Context matters — what's fair can depend on the situation.



Watch for

- Unequal outcomes across groups.
- Biased data or labels.
- Skewed model behavior.
- Unfair process or inconsistent treatment.



Remember



Use more than the aggregate. Check slices.
Investigate data, model and process sources of bias.



Latency



Simple meaning

The time an AI system takes to return useful output.



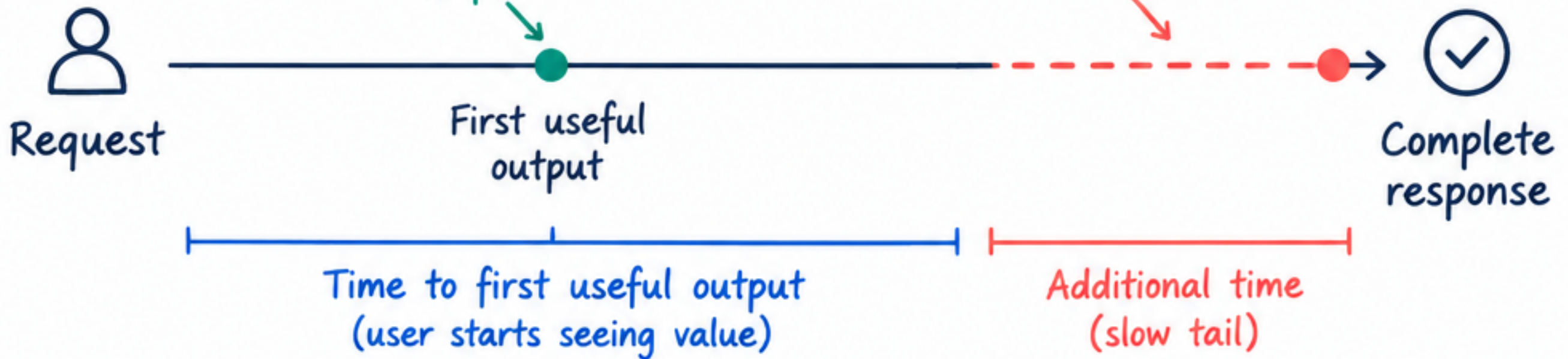
Tester lens

Measure user-visible response time as well as quality.
Track slow outliers, not only averages.



Request timeline

Quick, useful response



Key point

- Latency isn't just one number.
- Outliers can hurt the experience.
- Focus on what users actually feel and what affects quality.



Remember

The goal isn't just a lower average — it's a better user experience, measured by response time and quality, while tracking slow outliers.



Token Usage & Cost



Simple meaning

The amount of model input and output processed, and the resulting usage cost.



Tester lens

Evaluate quality together with token consumption. Longer prompts and outputs can increase cost without improving results.



Input Tokens

- prompt text
- context / history
- instructions



Output Tokens

- model response
- longer answers
- more tokens



Same quality, fewer tokens

= more efficient.

Option A

Higher token usage



(higher cost)

✗ Less efficient

Option B

Lower token usage



(lower cost)

✓ More efficient



Keep in mind

- Longer prompts can add input tokens.
- Longer outputs can add output tokens.
- The cost depends on the model and context (e.g. task, length, settings).



Remember

The goal is not just good quality, but also mindful token usage and cost.



Task Success Rate



Simple meaning

The percentage of end-to-end tasks an AI system or agent completes successfully according to defined success criteria.



Tester lens

For agents, final success matters more than a fluent explanation. Include tool choice, action results, recovery, and completion state in the evidence.



10 agent tasks



$$\text{Task Success} = \frac{\text{successful end-to-end outcomes}}{\text{total tasks}}$$



What counts as success?

- Uses the right tool
- Gets the expected action result
- Recovers from issues (when needed)
- Reaches the completion state



Watch for

- Wrong tool (e.g., picks the wrong action)
- Unfinished (e.g., stops early, doesn't reach the goal)



Don't count as success

- Incomplete tasks
- Tasks with errors
- Tasks that need recovery but don't finish



Remember

It's not just about a good explanation. It's about the final outcome, and the evidence that the agent took the right steps to get there.

