Eval	**LLM-as-a-judge**	**Position bias**
Non-determinism	**pass@k**	**pass^k**
Groundedness	**Hallucination**	**Red teaming**

POSITION BIAS

A judge preferring whichever candidate appears first, or second, regardless of quality.

Why it matters: Fixed by test design: run both orderings and require a consistent verdict.

LLM-AS-A-JUDGE

Using a strong model to score another model's output against a rubric or reference.

Why it matters: On the benchmark that named the technique, the judge agreed with expert humans 85% of the time against 81% human-to-human. Useful, and itself a component under test.

EVAL

A structured test of model output against defined criteria, run over a curated set of cases.

Why it matters: The AI equivalent of a test suite, except the unit of judgment is a rate across runs, not a single outcome.

PASS^K

The probability that ALL k attempts succeed.

Why it matters: The one that matters for anything an agent does unattended. The two are read aloud identically, so write them down.

PASS@K

The probability that at least one of k attempts succeeds.

Why it matters: Optimistic by construction. Useful for code generation, misleading as a reliability claim.

NON-DETERMINISM

The same input producing different output on repeated calls.

Why it matters: Temperature zero does NOT buy a reproducible test - production endpoints vary with server-side batch size. "It passed once" is not evidence.

RED TEAMING

Structured adversarial probing for harmful or policy-violating behavior.

Why it matters: Exploratory testing for harm. Its findings are the raw material for your safety regression suite - it is not itself a coverage claim.

HALLUCINATION

Confident output not supported by the source or by fact.

Why it matters: Not a bug you fix once. Treat it as a measured rate with a threshold, like any other quality property.

GROUNDEDNESS

Whether an answer is supported by the retrieved context it was given.

Why it matters: Distinct from factuality. An answer can be true and ungrounded, which is still a failure of the retrieval system.

Prompt injection	**Drift**	**Golden dataset**

GOLDEN DATASET

A hand-curated, expert-labeled set whose expected outputs are trusted as correct.

Why it matters: No standards body and no major vendor defines the phrase. The citable term is ground truth, and its label quality caps every metric you compute from it.

DRIFT

Change over time in input distribution (data drift) or in the relationship between input and outcome (concept drift).

Why it matters: The failure mode with no equivalent in classic software: a system that passed every test at release degrades with nobody changing a line.

PROMPT INJECTION

Input that alters behavior because instructions and data share one channel.

Why it matters: The highest-severity class for any tool-using agent. Indirect injection - from a page, ticket or document the system reads - is the harder half.