AI evaluation plan

A working guide from Beeston Labs. Version date: 8 November 2026.

Use this document before running the evaluation. It connects tests to a decision and records the conditions needed to interpret the result.

Decision and scope

Decision to support:

System, model and configuration:

Intended use and users:

Uses or environments outside scope:

Evaluation owner and decision owner:

Baseline for comparison: Existing process, current system or other agreed alternative.

Claims to examine

Claim	Why it matters	Test method	Acceptance criterion	Evidence location

Write a claim narrowly enough to test. Replace “works accurately” with the specific task and conditions. Agree acceptance criteria before examining results, including consequences that cannot be averaged away.

Test data and scenarios

Describe the population or operating environment the test is intended to represent. Record the source and permitted use of the material, the selection method and the relevant gaps.

Include ordinary examples, difficult conditions, expected refusals or abstentions, incomplete inputs and plausible misuse within the agreed testing scope. Keep tuning examples separate from the material used to judge acceptance.

For physical AI, include sensor degradation, occlusion, delayed or missing results, relevant environmental variation and the required system response. Plan hazardous-condition testing through appropriate simulation, controlled environments and specialist oversight; do not create dangerous live conditions to satisfy a generic checklist.

Measures

Choose measures that reflect the decision. Possible measures include task completion, critical-error count, false alerts per operating period, missed events, time to a usable result, human correction effort and successful recovery from a simulated failure.

State the denominator and observation period. Record important subgroups or conditions separately where an average would hide a material failure. Do not set a universal percentage threshold without a reason connected to the use.

Configuration record

Record model and application versions, prompts or policy configuration, data-processing steps, tools and permissions. For edge AI, include the host, accelerator, compiled model, quantisation/calibration approach, runtime, power mode and relevant thermal conditions.

For NVIDIA/Hailo comparisons, use the same task definition and explain any differences in precision, preprocessing, batch size or supporting host work. Headline compute ratings alone are not the evaluation result.

Conduct and exceptions

Who runs the test?

Who establishes the reference answers?

How are ambiguous cases handled?

How are failed runs, configuration changes and exclusions recorded?

Which issue would stop testing and trigger review?

Results and decision

Finding	Supporting result	Limitation	Effect on decision	Next action/owner

Report failures as well as the headline result. A small sample with no observed failure does not prove that failure is impossible. Identify conditions not yet evaluated.

The decision should specify the permitted use, conditions, remaining work and reassessment triggers. Store enough evidence to repeat the relevant test after a change.

This plan is a general evaluation aid. It is not a complete functional-safety validation protocol or a substitute for required specialist assessment.

You may adapt this guide for use within your own organisation. Keep the Beeston Labs attribution and review date when sharing an unmodified copy.