PLANCKCYBER
INTELLIGENT SYSTEMS. PRACTICAL RESULTS.

LLM Evaluation Template

Representative cases, acceptance criteria and regression evidence.

INTELLIGENT SYSTEMS. PRACTICAL RESULTS.

How this asset operates

Funnel stage	Pilot, build and managed operations
Audience	Product, engineering, quality, risk and operating owners
Trigger	An LLM-enabled feature needs pilot or release evidence
Objective	Create a reproducible evaluation tied to the workflow
Primary CTA	Approve the test set and thresholds before release
Delivery method	PDF evaluation plan
CRM action	Link evaluation version and release decision to project record
Follow-up	Run evaluation, review failures and document release decision
Compliance	General technical template; evaluation results do not establish absolute accuracy, security or compliance.
KPI	Coverage, failure detection, regression rate and release-gate adherence

A benchmark should represent the actual workflow, users, data, exceptions and harms - not only general model performance.

Evaluation identity

Feature / workflow

____

____

Version

____

____

Model and configuration

____

____

Evaluation owner

____

____

Acceptance owner

____

____

Evaluation date

____

____

Case taxonomy

Case category	Purpose	Count	Source	Sensitive-data rule	Expected behavior
Normal cases					
Edge cases					
Known failures					
Adversarial / abuse					
Policy / refusal					
Human escalation					

Measures

Measure	Definition	Threshold	Method	Owner
Task success				
Factual support / citation				
Policy compliance				
Exception / escalation				
Latency				
Cost				
User acceptance				

Failure review

Material failures

____

____

Root cause patterns

____

____

Control or design change

____

____

Residual risk

____

____

Release recommendation

____

____

Release decision

[ ] Pass

[ ] Conditional pass with documented controls

[ ] Fail / do not release

[ ] Further evidence required

Decision record and limitations

Decision or conclusion

Evidence supporting the decision

Material limitations and assumptions

Accountable owner

Review date

Important

General technical template; evaluation results do not establish absolute accuracy, security or compliance.

PlanckCyber resources are general planning tools. They do not create a client relationship or substitute for qualified legal, privacy, security, financial, accounting, engineering, regulatory or procurement review.