

DISRUKTECH RESEARCH

AI Agent Readiness & Evaluation Toolkit

A practical executive toolkit to score use cases, design production architecture, define evaluations and compare three reference applications.

VERSION 1.0 | 9 OCTOBER 2026 | DISRUKTECH RESEARCH

Independent enterprise technology and AI firm

How to use this toolkit

Use the scorecard before selecting a model or platform. Score one bounded business task, not an entire department or transformation program.

- Bring a process owner, enterprise architect, security representative and system owner.
- Score from 0 to 3 using evidence, not aspiration.
- Record the source for every score.
- Stop or redesign use cases with weak ownership, prohibited data or irreversible high-risk actions.
- Re-score after architecture and control changes.

Scoring scale

Score	Meaning
0	Unknown or absent
1	Partially understood; major dependency
2	Adequate for a bounded pilot
3	Production-ready evidence and ownership

1. Enterprise Agent Readiness Scorecard

Dimension	Questions	Score 0-3
Business value	Named owner? Baseline? Measurable cycle time, quality, cost or risk outcome?	
Process boundary	Clear trigger, end state, exceptions and hand-offs?	
Data and knowledge	Allowed sources, permissions, quality and provenance known?	
Tools and systems	Supported APIs, validated inputs and test environment available?	
Control readiness	Identity, least privilege, approvals, audit, reversal and stop authority?	
Evaluation readiness	Representative test set, failure taxonomy and acceptance thresholds?	
Operating ownership	Named business and technical owners, monitoring and review cadence?	

Interpretation

Total	Recommended decision
17-21	Candidate for controlled production design
12-16	Pilot after closing named gaps
7-11	Fix foundations before building
0-6	Stop or reframe the use case

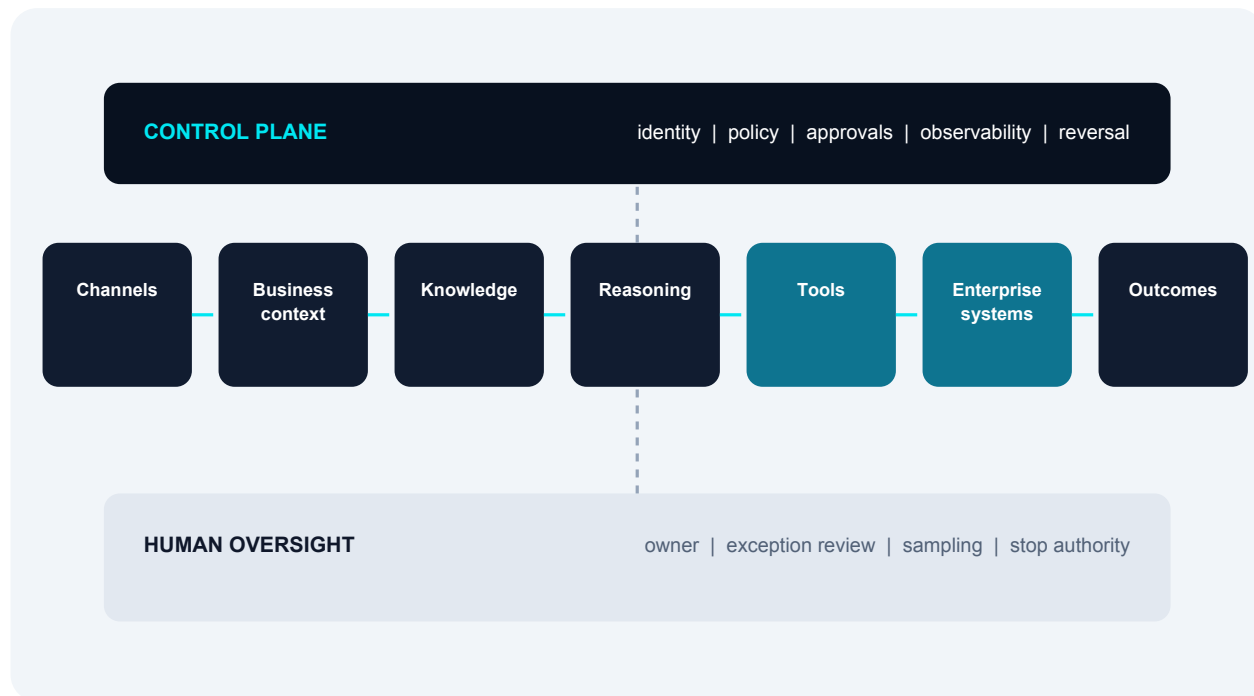
2. AI Portfolio Decision Matrix

Decision	Business value	Readiness	What happens next
Scale	High	High	Production controls, operating owner and rollout plan
Fix first	High	Low	Resolve data, access, process or control gaps
Pilot	Moderate	Moderate or high	Time-boxed test with explicit learning goals
Stop	Low or unclear	Any	Remove from backlog or reformulate the business problem

Portfolio questions

- What specific metric should move?
- What must be true before the agent receives system access?
- Which failure would create unacceptable business or compliance exposure?
- Who can stop the agent?
- What is the ongoing cost per completed business outcome?

3. Production Architecture



The architecture separates model capability from enterprise authority. Models reason; authenticated tools act. The control plane defines which actions are permitted, under which identity, with which evidence and approval.

Minimum production artifacts

- Agent brief and system boundary
- Knowledge-source and permission map
- Tool contracts with validated inputs
- Autonomy and control matrix
- Evaluation suite and acceptance thresholds
- Runbook, dashboard, escalation and stop process

4. Evaluation Framework

Measure	What it tests	Example acceptance evidence
Task success	Completed outcome under defined conditions	Correct resolution across representative cases
Grounding	Claims trace to approved sources	Source coverage and unsupported-claim rate
Policy adherence	Rules and thresholds are followed	Zero unauthorized actions in control tests
Tool reliability	Correct tool, parameters and sequence	Successful validated calls and safe failure handling
Human escalation	Agent stops when outside authority	Correct escalation rate on boundary cases
Security	Resistance to prompt injection and privilege expansion	Adversarial test results and access review
Operations	Cost, latency, drift and exceptions	Dashboard thresholds and named response owner

Evaluation is not a one-time benchmark. Re-run the suite after model, prompt, knowledge, tool or policy changes.

5A. Reference Case - Procurement Exception

Element	Reference design
Trigger	Supplier invoice blocked by price variance
Knowledge	PO, receipt, amendment, supplier terms and policy
Agent action	Explain cause; prepare release or escalation
Controls	Dedicated identity, variance threshold, segregation of duties, approval queue, audit and reversal
Outcome	Shorter exception cycle time without weakening payment policy
Evaluation	Resolution accuracy, evidence coverage, unauthorized-release rate, escalation quality

Illustrative reference design using fictitious data.

5B. Reference Case - Manufacturing Planning

Element	Reference design
Trigger	MRP exception messages exceed planner capacity
Knowledge	Material plan, inventory, open orders, lead times, production schedule and approved planning rules
Agent action	Prioritize exceptions and recommend reschedule, expedite or review
Controls	Read-only first; plant and material scope; explanation; approval before order changes
Outcome	Faster planner attention on exceptions with material business impact
Evaluation	Priority precision, missed-critical-exception rate, recommendation acceptance and planner time saved

Illustrative reference design using fictitious data.

5C. Reference Case - Customer Service

Element	Reference design
Trigger	Customer request arrives by portal or email
Knowledge	Customer, order status, service policy, product documentation and prior cases
Agent action	Classify, draft grounded response, retrieve status and route exceptions
Controls	Customer-level permissions, approved sources, personal-data rules and mandatory human hand-off
Outcome	Faster first response and more consistent case routing
Evaluation	Classification accuracy, grounded-response rate, hand-off precision, response time and customer correction rate

Illustrative reference design using fictitious data.

Engagement outputs and citation

A DISRUKTECH readiness or use-case engagement can produce:

- Scored use-case portfolio
- Agent brief for the top candidate
- Production reference architecture
- Autonomy and control matrix
- Evaluation plan and go-live criteria
- Twelve-month roadmap and operating ownership model

Suggested citation

DISRUKTECH Research. AI Agent Readiness & Evaluation Toolkit. Version 1.0, 9 October 2026.

<https://disruktech.com/research>

Status

Practitioner toolkit. Examples are reference designs with fictitious data and are not represented as client deployments.

disruktech.com | info@disruktech.com