CARE

WHITEPAPER

An Agent Passed Its Tests. Is It Ready?

Why evaluation must include the operating envelope

A benchmark can show that an agent performs a task. It rarely proves that the task, permissions, and escalation rules are appropriate for your business.

AGENTIC CARE

https://ai-managementsystem.com

Contents

01 Passing a test is not permission to operate

02 Benchmarks omit business authority and failure consequences

03 The invoice extractor that misses escalation

04 Evaluate behavior against the accepted operating model

05 A practical readiness decision

06 Separate test evidence from go-live authority

07 Who decides that an agent is ready?

08 Readiness is an acceptance decision supported by tests

09 Start with one agent

Passing a test is not permission to operate



Passing a benchmark is not the same as demonstrating operational readiness.

A benchmark can show that an agent performs a task. It rarely proves that the task, permissions, and escalation rule for your business.

A model can answer accurately while disclosing protected information or making a decision it should have escalate pass a happy-path test while failing in changed context.

Benchmarks omit business authority and failure consequences

Operational readiness includes task competence, explicit constraints, credible exception behavior, and the ability to evidence supports the result.

Construct test packages that supply only the accepted instructions and declared context. Test ordinary tasks and ad missing inputs, conflicting rules, requests for prohibited acts, and unavailable approval authorities.

An agent that only passes after receiving extra undocumented context is not a faithful implementation of its accept

> **MANAGEMENT IMPLICATION**
>
> A management approach should make the underlying decision visible without automatically m administrative work.

The invoice extractor that misses escalation



Tests can show task competence while still missing the business context that determines whether escala required.

> **EXAMPLE**
>
> An agent passes a benchmark for extracting invoice numbers. The same agent fails to escalate invoice instructions conflict with a purchasing limit.

Evaluate behavior against the accepted operating model

Automated judging can scale repeatability, but judges can hallucinate, reward verbosity, and miss unsafe side effec scenario, expected outcome, transcript, evaluator result, and any artifacts needed to reproduce the conclusion.

For consequential workflows, pair behavioral evaluations with independent verification of tool permissions and dow effects.

AGENT AND AUTOMATION

Prepare descriptions, changes, findings and proposals.

ACCOUNTABLE PEOPLE

Validate context, assess consequ accept decisions.

The meaningful test is whether the agent succeeds with only its accepted instruction and work package, including c exceptions.



ORGANIZATION

Taskboard

- Box workspace [redacted] holds the master AP sheet, vendor PDF folders and dated backups in `_meta/`.
- Agent messaging to AP Sync Orchestrator, ERP AP Sink and Steuerberater Sync.
- Canonical archive: [redacted]. The [redacted].
- Credentials: held in the runtime environment and never written into this SOP.

Inputs and triggers

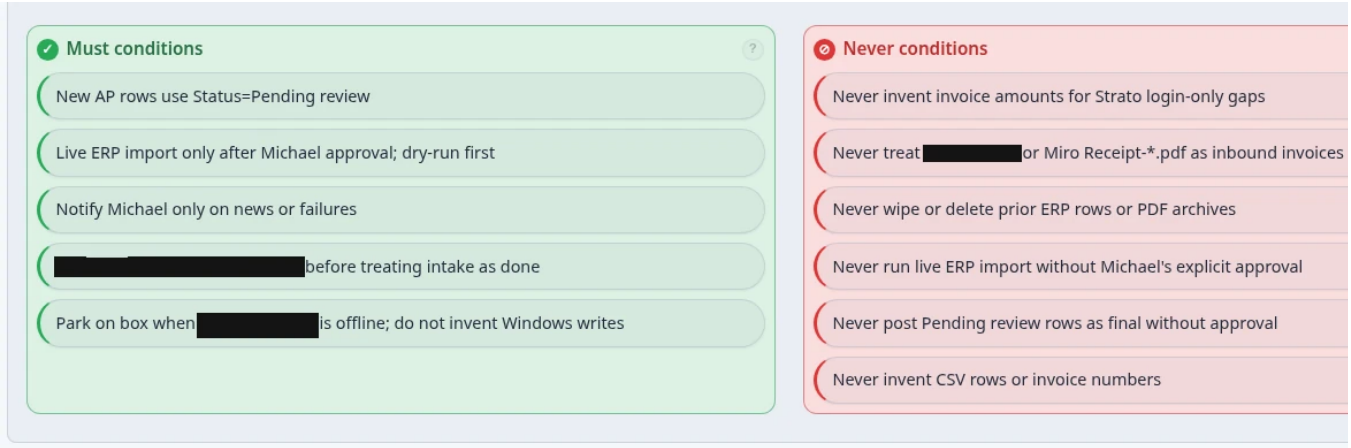
- Scheduled: AP Sync Orchestrator, Mon–Fri, twice daily (morning about 09:19 and afternoon about 15:19 Europe/Berlin). Each run covers the window since the previous run.
- Ad hoc: Michael sends a manual download path on [redacted] (e.g. [redacted]…) or attaches a PDF in chat.
- Inputs: new or unlabeled invoice-like Gmail with PDF attachments; the existing AP sheet and vendor folders; files parked on the box that are waiting for [redacted]

Operating procedure

Title	Action	Checkpoint	Expected outcome
Search Gmail	Search for invoice-like mail after the last run, including forwards. If an email is non-financial (server alert, newsletter, bank postbox notice), ignore it completely and do not report it	Time window set from the previous run	Candidate list with MessageId, sender, date and attachments
Classify and route	AP invoice → AP/[redacted]Lohnauswertung → [redacted] (and archive [redacted] only. Statement/non-invoice → [redacted] Miro receipt → [redacted]. Paddle `[TEST]` → [redacted]. Strato login-only → [redacted]	Each item has exactly one route	No payroll, archives or test items in AP
Deduplicate	Check Vendor+InvoiceNumber and MessageId against the sheet. One email can carry several invoices, so check each invoice separately	Every candidate is checked	No duplicate rows
Save PDFs	Save to [redacted]. For manual downloads, copy from [redacted] first	PDF readable	PDF stored under its vendor
Extract and append	Extract vendor, number, date, currency, net, tax and gross. Back up the CSV/XLSX to `_meta/`, append rows with Status= `Pending review`, rebuild the xlsx	Amounts come from the document; empty with a note if absent	Sheet updated, with backup
Label Gmail	Apply `Rechnungen` + partner sublabel (or `Sonstige`) and mark read. Never skip unlabeled invoices	Row written	Mail trail complete
Sync Windows	If [redacted] is online: copy via [redacted]…, check the Windows CSV line count is not above the box copy, make a `.bak`, move into [redacted] (PowerShell syntax). If offline, park on the box and sync on the next run	Line-count check passed	[redacted] matches the box
Report to orchestrator	SendToAgent to AP Sync Orchestrator, priority false: nothing new, rows added (vendor/number/amounts), or the error	Run finished	Orchestrator informed
Hand off	If there are new AP rows: ask ERP AP Sink for a dry-run (live only after Michael approves in ERP AP Sink's own chat), and send supplier invoices to Steuerberater Sync for the [redacted] Never send [redacted] payroll or `[TEST]` items	New rows exist	Sinks informed
Notify Michael	Message Michael only with real results, decisions or blockers. Stay silent on "nothing new" and on non-financial mail	Something material happened	Michael informed only when it matters

Business rules and exceptions

Observed CARE Standard Operating Procedure for Finance Syncing: scheduled and ad-hoc triggers plus a step table c classify, deduplicate, save PDFs, extract, sync, hand off, and notify.



Observed CARE business rules: Must conditions such as Status=Pending review and dry-run before live ERP import, an such as never inventing invoice amounts or running live import without explicit approval.

A practical readiness decision

Separate a passed evaluation from the decision to allow production use. The latter must consider ownership, residu controls, and operational coverage.

CARE's evaluation-oriented workflow can make these artifacts part of a legible review, without implying that any sir certifies safety.

> **MANAGEMENT IMPLICATION**
>
> Approve readiness against the actual declared operating model, not a generic competence sco

Separate test evidence from go-live authority

Use this comparison to distinguish the current operating habit from the management capability the organization n

Question	Evidence required
Can it perform the task?	Representative work-package outcomes
Will it respect limits?	Prohibited-action and escalation tests
Is it acceptable to deploy?	Owner decision and operational controls

Use the comparison to test one real agent rather than discussing an abstract enterprise-wide target state. Where th unknown, record that uncertainty and assign responsibility for resolving it.

Readiness is an acceptance decision supported by tests

> Readiness is a management judgment informed by tests, not a test score promoted to authorit

The first practical improvement is to create one current, accountable record and use it to make a better manageme scalable system follows when that discipline is repeated across agents and maintained as conditions change.

Start with one agent

Use a real agent and review what it produces.

1. the agent to https://ai-managementsystem.com/under-care.
2. it is finished, log into CARE and review its documentation.

Images and text generated with the assistance of ChatGPT by OpenAI.

AGENTIC CARE

Instant AI Management System for your Agents, as a Managed Service.

https://ai-managementsystem.com