



AI for manufacturing and field operations

Applying AI to documentation, quality,
maintenance, and frontline decision support.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 18 / EMBER & LIME

EXECUTIVE WORKING EDITION - 08 OCT 2026

18

AT A GLANCE

Applying AI to documentation, quality, maintenance, and frontline decision support.



Plants and field teams run on procedures, shift notes, work orders, inspection records and the memory of experienced people. AI can make that knowledge easier to find, summarize and use, and it can help spot patterns in quality and maintenance data. It does not replace the people who sign off on safety-critical and quality-critical decisions. This playbook explains where AI fits on the floor and in the field, how to scope and control it, how to measure it, and how to run a first pilot that operators trust.

WHO IT IS FOR

Plant managers, operations and maintenance leaders, quality and EHS leads, field service managers, and the IT and data colleagues who support them.

TIME TO USE

70 minutes to read; two working sessions with a shift supervisor and a quality lead to apply



What you will have at the end

- A ranked shortlist of two or three operations use cases with named owners
- A clear line between what AI supports and what qualified people decide
- A pilot plan with a baseline, a pass mark and a review rhythm
- A source and data inventory for documents, work orders and inspection records

Contents

01	What AI is good for on the plant floor and in the field	03
02	Where AI fits and where it does not	04
03	How these systems work, in plain language	06
04	Roles, ownership and the human in the loop	08
05	Risks, controls and human review	10
06	Measuring quality and value	12
07	A practical implementation guide	14
08	Common pitfalls and an illustrative scenario	16
09	Questions to ask before you buy or build	17
10	A first quarter that produces evidence	19
11	Terms used in this publication	20



EXECUTIVE BRIEF

What AI is good for on the plant floor and in the field

01



The strongest early value is in knowledge and paperwork around the work: finding the right procedure, summarizing the last shift, drafting the report. Control of machines and safety decisions is a different matter.



Manufacturing and field operations generate a large amount of text and semi-structured data: standard operating procedures, maintenance manuals, work orders, nonconformance reports, shift logs, inspection photos, supplier certificates, and the tribal knowledge of people who have fixed the same fault many times. Most of it is hard to search and slow to read at the moment someone needs it. Modern language models and vision models are useful here because they can read, summarize, compare and draft.

The same technology is a poor fit for anything that must be deterministic, certified or instantaneous. A language model should not decide whether a guard can be bypassed, whether a lot is released, or whether a pressure vessel is safe to run. Those decisions belong to qualified people working inside your quality system and safety procedures. The practical stance of this playbook is simple: AI supports, people decide, and the boundary is written down before any pilot starts.



Where the value usually sits

- Operational knowledge: answering questions from manuals, procedures and past fixes, with the source shown
- Shift handover: turning scattered notes and tickets into a structured summary the next shift can act on
- Quality documentation: drafting nonconformance reports, corrective action write-ups and audit evidence packs for human review
- Maintenance support: summarizing fault history, suggesting likely causes to check, and preparing work order text
- Visual inspection support: flagging candidate defects in images for a trained inspector to confirm



- Training and onboarding: turning procedures into practice questions and quick-reference guides


BOUNDARY TO WRITE ON DAY ONE

AI may suggest, summarize, draft and flag. It may not release product, approve a safety-related change, override an interlock or close a regulated record. Qualified people do, and their names are on the record.

The executive question is not whether AI is impressive. It is whether a specific task, done by a specific role, becomes faster, more consistent or less error-prone without adding risk. That question can be answered with a small pilot, a baseline and a few weeks of honest review.


WORKING QUESTION

Which paperwork or lookup task do your best operators and technicians complain about most, and who owns it?


WHERE IT FITS

Where AI fits and where it does not

02

Match the tool to the kind of problem. Documents and language suit language models; signals and images need different methods; control loops need none of them.



It helps to separate four families of problem. Language problems involve reading, writing and searching text. Vision problems involve images from cameras or phones. Signal problems involve time series from sensors and machines, where statistical and machine learning methods have a long record. Control problems involve real-time decisions on equipment, which belong to engineered control systems and safety systems, not to a chat assistant.

PROBLEM FAMILY	TYPICAL TOOLS	FIT FOR AI SUPPORT
Language: procedures, reports, handovers	Language models with retrieval from approved documents	Strong, with human review of outputs
Vision: surface defects, label checks, photos	Trained image models, sometimes multimodal models	Moderate to strong, needs labeled examples and an inspector in the loop
Signals: vibration, temperature, current	Anomaly detection and forecasting models	Moderate, depends on data quality and failure history
Planning: schedules, spares, routes	Optimization plus AI for explanation and what-if	Moderate, optimization often matters more than language
Control and safety functions	Engineered control systems and safety-rated components	Not a place for general-purpose AI

Use case screening

USE CASE	FIT	REASON
Procedure and manual question answering	Strong	Repeated lookups, clear sources, easy to check against the document
Shift handover summaries	Strong	High repetition, supervisor reviews before it is final
Nonconformance report drafting	Strong with review	Saves writing time, quality engineer owns the content
Visual inspection assistance	Moderate	Needs lighting, camera control, labeled defects and a review step
Predictive maintenance on rotating equipment	Moderate	Requires enough failure examples and reliable sensor data
Autonomous parameter changes on a line	Weak	Process engineering and change control apply, risk is high
Safety interlock logic or permit decisions	Not suitable	Functional safety and qualified sign-off apply

A useful screening habit is to ask what happens when the AI is wrong and nobody notices. If the answer is a slightly clumsy report, proceed. If the answer is an injury, a scrapped batch, a regulatory finding or a customer return, the use case needs either a much stronger control or a different approach. Check current requirements for your sector and jurisdiction, and involve your quality, safety and legal functions early.

COMMON MISTAKE

Starting with the most exciting use case, such as autonomous process tuning, instead of the most boring one, such as shift handover. The boring one teaches the team how to supervise AI at low risk.

WORKING QUESTION

For your top three candidate use cases, what is the worst realistic outcome if the AI is wrong and nobody notices?

HOW IT WORKS

How these systems work, in plain language

03

Knowing the moving parts helps you ask better questions of vendors and spot where errors enter.

A language model predicts useful text from the text it is given. It has no live knowledge of your plant. To answer a question about a specific pump, it must be shown the relevant manual page, the maintenance history and the question together. This pattern, often called retrieval-augmented generation, searches approved documents, hands the best passages to the model, and asks it to answer from them with citations. If the right passage is not found, the answer is weak, however fluent it sounds.

Vision systems for inspection are usually trained on labeled examples of good and defective parts. They are sensitive to lighting, camera angle, surface finish and new product variants. A model that works on the pilot line can degrade quietly when a supplier changes a coating. Signal models for maintenance learn what normal looks like for a machine and flag drift, but they need dependable sensor data and enough history of real failures to be useful beyond simple alerts.

01 Capture

Bring in the inputs: documents, work orders, shift notes, images or sensor readings. Note who owns each source and how current it is.

02 Prepare

Clean, label and structure the inputs. Remove superseded procedures. Attach metadata such as asset, line, revision and date.



03 Process

Use the model to retrieve, summarize, classify or flag. Constrain it to approved sources and to a defined output format.

04 Review

A named person checks the output against the source before it is used, filed or acted on.

05 Record

Log the input, output, reviewer and decision so errors can be traced and the system can be improved.



- Fluency is not accuracy: a confident answer can cite the wrong revision of a procedure
- Models can misread units, part numbers and tolerances, so spot checks on numbers matter
- Images and scanned documents add recognition errors before the model even reasons
- Edge cases such as rare faults and new products are exactly where models are weakest
- Connectivity matters in the field: decide what must work offline and what can wait for a signal

None of this requires your team to become machine learning experts. It requires them to know that outputs depend on inputs, that sources must be current, and that review is part of the process rather than an optional extra.



WORKING QUESTION

If a technician asked the assistant about a specific machine today, which document would it need to find, and is that document current and findable?



OPERATING MODEL

Roles, ownership and the human in the loop

04

AI in operations works when each output has a named reviewer and each source has a named owner.

The plant already has an operating model: supervisors, maintenance planners, quality engineers, EHS leads, document control, and IT or OT teams. AI should be placed inside it rather than beside it. The most common cause of failed pilots is that the tool lives in a corner with no owner, no review step and no connection to how work is actually released and recorded.

ROLE	RESPONSIBILITY FOR AI-SUPPORTED WORK	DECISION THEY KEEP
Operations sponsor	Chooses use cases, removes blockers, owns the value case	Whether to scale, change or stop
Shift supervisor	Reviews handover summaries and flags errors	What the next shift is told
Quality engineer	Reviews drafted reports and defects flags	Disposition of nonconforming product
Maintenance planner	Reviews fault summaries and drafted work orders	Work scope and priority
EHS lead	Defines where AI must not be used and reviews incidents	Safety-related approvals and permits
Document control owner	Keeps sources current and removes superseded files	What counts as the approved version
IT and OT security	Reviews access, networks and data flows	What may connect to production systems

Levels of autonomy

- Level 0, inform: the assistant answers questions and shows sources, and the user decides everything

- Level 1, draft: the assistant prepares a document or record and a named person edits and approves
- Level 2, flag: the assistant raises candidate issues, such as a possible defect, and a trained person confirms
- Level 3, act within limits: the assistant takes a reversible, low-risk action inside strict rules, with logging
- Anything touching safety functions or product release stays at Level 0 to 2 unless qualified people and your formal change process decide otherwise

Working with unions, works councils or employee representatives early is good practice where they exist. Operators who understand what the tool does, and what it does not do, are far more likely to report its errors honestly. Make clear that the aim is to reduce paperwork and searching, and describe how any monitoring data will and will not be used.

EXAMPLE, ILLUSTRATIVE

A maintenance planner reviews AI-drafted work order text each morning. The AI proposes; the planner edits and releases. If the planner disagrees often, that is a signal to fix the source documents or the prompt, not to trust the draft.

WORKING QUESTION

For your first use case, who is the named reviewer, and do they have time in their week to do the review properly?

RISKS AND CONTROLS

Risks, controls and human review

05

Most operational risk comes from stale sources, unchecked numbers and over-trust. Controls are mostly ordinary good practice applied to a new tool.

RISK	HOW IT SHOWS UP	CONTROL
Stale or superseded procedures	Assistant cites an old revision	Index only controlled documents; show revision and date; remove old files
Wrong numbers or units	Torque, tolerance or dosage misquoted	Require source link; spot check numbers; block numeric answers without a citation
Over-trust by operators	Draft accepted without reading	Review sampling; require edits or sign-off; track acceptance rates
Data leakage	Supplier or customer data pasted into a public tool	Approved tools only; data classification; access rules
Model drift in vision	Defect detection degrades after process change	Scheduled re-validation; trigger review on any material or supplier change
Safety misuse	Someone asks the tool whether a guarded step is safe	Clear signage and policy; assistant refuses and points to the safety procedure
Record integrity	AI-written text in a regulated record	Named human author; audit trail; follow your quality system rules

Safety and regulated environments

Functional safety, machinery safety and sector quality regimes each have their own requirements for how safety-related functions are designed, validated and changed. General-purpose AI is not designed or certified for those functions, and this playbook does not provide certification or compliance advice. The principle is to keep AI outside safety-rated functions, to keep qualified people responsible for sign-off, and to check current requirements

and involve your safety and quality professionals before any AI touches a regulated process or record.

Before any pilot goes live

- The list of prohibited uses is written and shared with the team
- Every source is controlled, owned and current, with revision visible
- Outputs show their sources and are easy to check
- A named reviewer exists for each output type
- Data handling has been reviewed for supplier, customer and personal data
- There is a simple way for users to report a wrong or unsafe answer
- A stop rule exists: who can pause the pilot, and on what trigger
- Change control applies to prompts, models and sources, not only to code

Incident handling should be rehearsed in advance. If the assistant gives a wrong instruction and someone acts on it, the team needs to know whom to tell, how to preserve the interaction log and how the near miss feeds back into controls. Treat AI errors like any other near miss: report them without blame, learn from them and change the system.

WORKING QUESTION

If the assistant were wrong about a procedure tomorrow, how would your team find out, and how fast?

MEASURING VALUE

Measuring quality and value

06

Set a baseline before the pilot, measure quality separately from speed, and count the cost of review.

Operations leaders are used to metrics, which is an advantage. The discipline is to choose them before the pilot and to avoid claiming credit for improvements that came from something else, such as a new line, a new supervisor or a seasonal change. Capture a baseline for four to six weeks if you can, or run the AI-supported process on one line or shift while a comparable one continues as before.

MEASURE	WHAT IT TELLS YOU	HOW TO COLLECT
Time to find an answer	Search and lookup effort saved	Timed samples before and after, by role
Handover completeness	Whether the next shift has what it needs	Supervisor checklist score on a sample
Report cycle time	Paperwork burden	Timestamps from draft to approval
Answer accuracy	Whether outputs match the source	Reviewer scoring on a fixed test set
Correction rate	How much reviewers change	Edit tracking or reviewer tally
Escapes and misses	Defects or faults the process missed	Quality and maintenance records
User trust and use	Whether people actually rely on it	Short survey and usage logs

Build a test set

01 Collect real questions

Ask operators and technicians for 40 to 60 actual questions they asked in the last month.

02 Write expected answers

Have an experienced person record the right answer and the source document for each.

03 Score the assistant

Grade correct, partly correct, wrong and unsafe separately. Unsafe answers are tracked on their own line.

04 Repeat after changes

Re-run the set whenever sources, prompts or the model change.

Value should be stated in operational terms first and money second. Hours of searching removed per technician per week, report turnaround, repeat visits avoided, or defects caught earlier are all credible. Illustrative example, not a benchmark: if 30 technicians each save 20 minutes a day, that is about 10 hours of capacity per day, but it only becomes value if that time is redeployed to work that matters. Subtract the review time and the cost of maintaining sources.

HONEST ACCOUNTING

If reviewers rewrite most of what the AI drafts, the pilot is not saving time. Report that plainly. A negative result on one use case is useful and cheap.

WORKING QUESTION

What number would convince your plant manager that this use case is working, and have you measured the current value of it?

IMPLEMENTATION GUIDE

A practical implementation guide

07

Start with one line, one team and one use case. Prove it, then widen.

01 Choose the use case

Pick a task with frequent repetition, clear sources, low consequence of error and an engaged supervisor. Shift handover and procedure lookup are common starting points.

02 Inventory sources

List documents, systems and data needed. Name an owner for each, mark controlled versions, and retire duplicates.

03 Define the boundary

Write what the tool may and may not do. Review it with safety, quality and the people who will use it.

04 Build a thin pilot

Use an approved tool with access controls. Keep scope small: one line or one asset family, 10 to 20 users, 6 to 8 weeks.

05 Review weekly

Look at misses, edits, user comments and any near misses. Fix sources first, prompts second, model choice last.

06 Decide with evidence

At the end, choose to scale, change or stop, using the baseline and the test set.

Field operations considerations

- Connectivity: decide what must work offline, such as downloaded manuals, and what can sync later
- Devices: confirm the tool works on the phones, tablets or rugged devices technicians already carry
- Voice and photo input: dictation and photos reduce typing, but check accuracy with gloves, noise and glare

- Language: crews may work in several languages, so test translations of technical terms with native speakers
- Parts and asset naming: inconsistent names in work orders cause retrieval misses, so clean naming early
- Privacy: photos taken on site may include people or customer property, so set rules for capture and storage



Pilot readiness

- Sponsor, reviewer and source owners are named
- Baseline is captured or a comparison group is chosen
- Test set of real questions exists with expected answers
- Training session is booked for users and reviewers
- Feedback channel is visible on the floor or in the app
- Stop rule and escalation path are agreed
- Date for the scale, change or stop decision is in diaries



Training should be practical and short: show real examples, show a wrong answer and how it was caught, and explain where the tool must not be used. Supervisors and experienced operators are the best champions because their judgment sets the tone. Give them a way to say that a use case is not working.

WORKING QUESTION

Which line or crew is ready to pilot, and who is the supervisor who will genuinely use and challenge the tool?



PITFALLS

Common pitfalls and an illustrative scenario

08

Most failures are ordinary: dirty sources, no reviewer, unclear boundaries and no one measuring.

- Indexing every shared folder, including drafts and superseded procedures
- Treating a vision demo on clean samples as proof for the production environment
- Leaving review to whoever has time, so it becomes optional
- Skipping the baseline and then being unable to show value
- Letting the pilot run on without a decision date
- Ignoring operator feedback because it arrives informally
- Connecting tools to production systems before security and OT teams are involved

ILLUSTRATIVE SCENARIO, HYPOTHETICAL

A plant pilots a procedure assistant. Early answers cite two revisions of the same cleaning procedure. Operators lose trust within a week. The fix is not a better model: document control removes the old revision, the assistant shows revision and date on every answer, and a supervisor re-runs the test set.

What recovery looks like

Recovery from a rough start usually follows a pattern. The team pauses to look at the log of misses rather than defending the tool. They group failures into source problems, retrieval problems, prompt problems and genuine model limits. Most turn out to be source or retrieval problems that are cheap to fix. They report the findings to users openly, because people forgive a tool that visibly improves and do not forgive one that is defended.

SYMPTOM	LIKELY CAUSE	FIRST FIX
Contradictory answers	Multiple versions of a document	Remove duplicates; index only controlled versions

SYMPTOM	LIKELY CAUSE	FIRST FIX
Answers miss obvious documents	Poor asset naming or chunking	Add metadata; test search on real questions
Users stop using it	Slow, unreliable or untrusted	Fix speed and accuracy on the top ten questions first
Reviewers rubber-stamp	Volume too high or review unclear	Sample review; clarify what to check
Vision false alarms	Lighting or product variation	Standardize capture; add examples; retrain with inspector input

Another quiet pitfall is vendor claims. Be wary of accuracy figures that were not measured on your parts, your lighting or your documents. Ask for a trial on your own data with your own test set, and compare results against your baseline rather than a brochure.

WORKING QUESTION

Which of these pitfalls is most likely in your plant, and what would an early warning sign look like?

BUYING AND SETUP

Questions to ask before you buy or build

09

Good questions expose assumptions about data, control and responsibility before money is spent.

Questions for any vendor or internal team

- Which of our sources will it use, and how does it handle revisions and removals?
- Can every answer show its source passage and revision?
- How is our data stored, who can access it, and is it used to train shared models?
- What happens offline, and how does it sync?

- How does it integrate with our maintenance, quality and document systems, and who maintains the integration?
- What evidence of performance on data like ours can you show, and can we run our own test set in a trial?
- What logs do we get, and can we export them for audit?
- How are updates to models and prompts announced and tested?
- What is the exit plan if we stop, and can we take our data and configuration with us?

DECISION	BUY	BUILD
Common documents and handovers	Often faster and good enough	Rarely worth it at the start
Plant-specific vision inspection	Specialist vendors may fit	Needs ML skills and labeled data
Integration with your systems	Check connectors carefully	Gives control, but you own maintenance
Governance and review workflow	Configure to your rules	Often unavoidable internal work

Cost should be viewed across licenses or usage fees, integration, source cleanup, training, review time and ongoing evaluation. The cheapest part is often the tool itself. Ask what the vendor expects from you on day one, such as clean data or an administrator, and be realistic about whether you can provide it.

If you want an outside view, Halden & Company offers free assessments at halden.company/assessments, and you can reach the team at halden.company/contact. Whatever route you take, keep the same principles: a narrow start, named owners, human sign-off for anything that matters, and a decision based on evidence.

WORKING QUESTION

If the vendor disappeared in a year, what would you keep, and what would you lose?

30 / 60 / 90

A first quarter that produces evidence

10



Prove one low-risk use case with named reviewers, then widen only where the evidence supports it.

First 30 days: ground it

- Choose one use case and name sponsor, reviewer and source owners
- Write the boundary of what AI may and may not do, with safety and quality
- Inventory and clean the sources needed for the pilot
- Capture a baseline and build a test set of real questions



Days 31 to 60: test it

- Run the pilot with a small group on one line or crew
- Review misses, edits and near misses every week
- Fix sources and prompts before considering a different model
- Hold a short training and feedback session with users



Days 61 to 90: decide

- Compare results with the baseline and the test set
- Decide to scale, change or stop and record why
- Document the operating model, reviewers and controls
- Select the next use case using what the pilot taught you



GLOSSARY

Terms used in this publication

11

TERM	MEANING
Retrieval-augmented generation	A pattern where the system finds relevant passages in approved documents and has the model answer from them, ideally with citations.
Shift handover	The transfer of status, issues and priorities from one shift to the next, often a source of lost information.
Nonconformance report	A record of product or process that does not meet requirements, with its disposition and corrective action.
Predictive maintenance	Using condition data and models to estimate when equipment is likely to need attention.
Anomaly detection	Methods that flag readings that differ from learned normal behavior.
Human in the loop	A design where a named person reviews or approves AI output before it is used.
Functional safety	The discipline of ensuring safety-related control functions work correctly; it has its own standards and qualified practitioners.
Document control	The process that keeps procedures versioned, approved and current.
Baseline	A measurement of current performance taken before a change so results can be compared fairly.
Test set	A fixed group of real questions or cases with known answers used to score a system repeatedly.
Model drift	A decline in model performance as conditions, materials or data change over time.
Change control	The formal process for approving and recording changes to systems, procedures or processes.

A note on use

This playbook is general guidance and not legal, safety or certification advice. Safety-critical and quality-critical decisions must be made and signed off by qualified people. Check current requirements for your sector and jurisdiction.

TAKE THE NEXT STEP

Take the free Halden assessment at halden.company/assessments, or talk to us at halden.company/contact about turning this into a plan for your organization.

