



AI for procurement and supply chain

Improving sourcing, supplier management, demand signals and exception handling with supervised AI.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 41 / EMBER & TEAL

EXECUTIVE WORKING EDITION - 08 OCT 2026

41

AT A GLANCE

Improving sourcing, supplier management, demand signals and exception handling with supervised AI.

Procurement and supply chain teams work with messy documents, fragmented data and a steady stream of exceptions, which makes them a strong fit for supervised AI. This playbook explains where AI helps in spend analysis, supplier risk, contract review, demand planning and exception triage, and where classical analytics or plain process fixes are the better tool. It covers the operating model, the controls needed when AI touches supplier data and commercial terms, how to measure quality and value honestly, a staged implementation path and the questions to ask vendors. Throughout, people keep the decisions on money, contracts and supplier relationships.

WHO IT IS FOR

Chief procurement officers, category managers, supply chain and operations leaders, sourcing analysts, contract managers and the finance and technology partners who support them.

TIME TO USE

60 minutes to read; one workshop to select and scope a first use case

What you will have at the end

- A shortlist of use cases ranked by value, data readiness and risk
- A clear view of which problems need AI, which need analytics and which need process change
- An operating model with named roles, decision rights and review points
- A measurement plan with a baseline, a test set and pass marks
- A 90-day path from a narrow pilot to a decision to scale

Contents

01	What AI can and cannot do for procurement and supply chain	03
02	Where AI fits, and where it does not	04
03	How the main use cases work, in plain language	06
04	Operating model, roles and decision rights	08
05	Risks, controls and human review	09
06	Measuring quality and value	11
07	Implementation guide: one category, one process, one quarter	13
08	Common pitfalls and an illustrative scenario	15
09	A first quarter that produces evidence	17
10	Terms used in this publication	18

EXECUTIVE BRIEF

What AI can and cannot do for procurement and supply chain

01



AI is strongest where the work is reading, classifying, comparing and drafting at volume, and weakest where it must replace commercial judgment or fix bad data on its own.

Procurement is a function of documents and exceptions. Contracts, quotes, invoices, purchase orders, supplier questionnaires, certificates, delivery notes and emails carry most of the information, and much of it is unstructured. Supply chain teams add forecasts, inventory positions and disruption signals. People spend large parts of their week reading, matching, chasing and reconciling. Modern language models can read and compare text, classify items, summarize and draft, which maps directly onto this load.



The opportunity is real but narrow in the right ways. AI can cut the effort of cleaning a spend cube, extracting terms from a contract set, triaging a supplier inbox or summarizing why a shipment is late. It can surface patterns a busy analyst would miss. It cannot negotiate for you, decide a make-or-buy question, or turn poor master data into reliable data without human review. Where numeric forecasting is the task, established statistical and machine learning methods are often more suitable than a language model, and a good program uses both.



- Best near-term value: classification, extraction, comparison, summarization and drafting at volume
- Keep human decisions on award, price, contract terms, supplier onboarding and exceptions with financial impact
- Data foundations decide results: supplier names, item codes and category taxonomies must be understood before automating
- Quality must be measured on your own documents and data, not accepted from a demonstration
- Start with one category or one process, prove it, then extend



PLAIN TEST

If a skilled buyer could do the task by reading documents and applying a clear rule, AI can probably assist. If the task depends on a relationship, a negotiation stance or information nobody has recorded, it will not.

WORKING QUESTION

Which part of your procurement week consists of reading, matching or classifying documents that a capable assistant could prepare for a person to check?

WHERE IT FITS

Where AI fits, and where it does not

02

Match the tool to the problem. Language models suit text; statistical models suit numbers; rules and process change suit many problems people assume need AI.

A common error is to buy AI for a problem that a clean master data file, a simple rule or a changed approval step would solve. Before selecting a use case, describe the pain in plain terms, find out where the data lives and ask whether the task is judgment, pattern recognition, transformation or lookup. Each points to a different tool. The table below maps typical procurement and supply chain work to its best-fit approach.

USE CASE	BEST-FIT APPROACH	FIT AND NOTE
Spend classification into categories	Language model with human sampling, plus rules for known suppliers	Strong; accuracy depends on clean descriptions and a clear taxonomy
Contract term extraction and comparison	Language model with citations to clauses	Strong with review; legal interpretation stays with counsel

USE CASE	BEST-FIT APPROACH	FIT AND NOTE
Supplier risk monitoring from news and filings	Language model for summarizing plus defined alert rules	Moderate; source quality and false alarms need tuning
Demand and replenishment forecasting	Statistical or machine learning forecasting; language model for explanation	Moderate; do not use a chat model as the forecast engine
Exception triage (late, short, damaged, mismatched)	Rules for known cases; language model to read free text and route	Strong; start with classification and drafting, not autonomous action
Supplier questionnaires and RFP responses	Language model to compare answers against requirements	Strong with review; scoring remains a human decision
Price negotiation strategy	Analytics for facts; people for judgment	Weak for AI alone
Master data repair at scale	Matching tools plus human review	Moderate; wrong merges are costly queues

Signs a use case is ready

Readiness checks

- The task is repeated often enough to matter, such as weekly or daily
- Inputs are available in digital form and can be shared with the tool under your data rules
- A person can check an output in a fraction of the time it takes to create one
- Errors are recoverable before money moves or a commitment is made
- There is an owner who feels the pain and will use the result
- You can describe what a correct output looks like

COMMON MISTAKE

Starting with the most exciting use case, such as autonomous negotiation, instead of the most measurable one, such as classifying spend or extracting contract dates. Build trust and data discipline on the second before attempting the first.

WORKING QUESTION

For your top three pain points, is the real fix AI, analytics, master data cleanup or a changed process?

HOW IT WORKS

How the main use cases work, in plain language

03

Each use case follows the same shape: gather inputs, apply a model under instructions, produce a draft with evidence, and send it to a person for a decision.

Understanding the mechanics helps you ask better questions and spot weak designs. A language model reads the text you give it and produces text in return. It does not hold a live view of your suppliers or contracts unless the system fetches that information and supplies it. That is why good designs use retrieval: the system finds the relevant contract, record or policy and shows it to the model with the question. The model then drafts, classifies or extracts, ideally pointing to the exact passage it used.

Spend analysis

01 Consolidate

Combine purchase orders, invoices, card spend and payments into one table with supplier, amount, date, description and cost center.

02 Normalize suppliers

Match name variants, subsidiaries and typos to a single supplier record, with uncertain matches sent to a person.

03 Classify

Assign each line to a category in your taxonomy using description, supplier and amount, with a confidence flag.

04 Review

A category analyst samples the output by category and by confidence, corrects errors and feeds corrections back as rules or examples.

05 Analyze

Standard analytics then find fragmented spend, price variance, tail suppliers and savings opportunities.

Contract review and comparison

The system reads each contract, locates defined clauses such as term, renewal, price adjustment, termination, liability, data protection and service levels, and returns the wording with its location. A comparison view lines these up against your standard positions and flags deviations. The model should state 'not found' where a clause is absent. The output prepares a contract manager's work; it does not give legal advice or decide acceptability.



Supplier risk and exception triage

For risk, the system gathers signals from your own records, such as late deliveries, quality incidents and certificate expiry, and from external sources you have approved, then summarizes changes for the supplier owner. For exceptions, it reads emails, delivery notes and system messages, classifies the issue, proposes a next step and drafts a message. In both cases the value is speed to attention and consistency of handling, with a person deciding the action.

USE CASE	TYPICAL INPUTS	DRAFT OUTPUT	HUMAN DECISION
Spend classification	Line descriptions, supplier, amount	Category and confidence	Approve taxonomy changes; sample review
Contract extraction	Contract files	Clause table with citations	Accept terms; escalate to counsel
Supplier risk	Internal records, approved external feeds	Change summary and rating suggestion	Mitigation and sourcing decisions
Exception triage	Emails, notes, order data	Category, next step, draft reply	Approve action and any commitment
Demand support	Sales history, promotions, signals	Forecast plus plain explanation	Plan sign-off



WORKING QUESTION

For one use case, can you trace the path from raw input to the person who signs off, and name each step?



OPERATING MODEL

Operating model, roles and decision rights

04

Procurement AI works when category experts, data owners and risk functions share responsibility, and when decision rights are written down before launch.

The common failure is to treat procurement AI as an IT project. The tools sit with technology, but the correctness of the output belongs to people who know the categories and contracts. Define who owns each use case, who reviews the outputs, who maintains the reference data and who may change the instructions. Without that, quality drifts and no one notices until a bad decision is traced back.

ROLE	OWNS	TYPICAL TASKS
Business owner (head of category or procurement operations)	Use case, value and adoption	Defines scope, accepts quality, reports results
Category manager	Domain correctness	Reviews classifications and summaries, supplies rules and examples
Contract manager or counsel	Contract positions and legal review	Defines clause library, reviews flagged deviations
Data steward	Supplier and item master data	Resolves duplicates, maintains taxonomy
Supply planner	Demand and exception handling	Validates forecasts and triage outcomes
Risk and compliance lead	Controls, supplier due diligence, audit trail	Approves data sources, reviews logs, sets limits
System owner	Configuration, access, monitoring	Manages releases, incidents and vendor relationship

Decision rights worth writing down

- Which outputs may be used as drafts and which may trigger any action automatically, if any

- Which dollar thresholds require a second reviewer
- Who may change prompts, rules, taxonomy and thresholds, and how changes are tested
- Who approves new data sources, especially external feeds about suppliers
- Who may share outputs with suppliers, and in what form
- How disagreements between the system and a category expert are resolved



Create a lightweight review rhythm. A weekly session during the pilot looks at errors, overrides and unusual cases; a monthly one covers metrics, drift and changes. Track every human override with a reason, since the pattern of overrides tells you whether the problem lies in data, instructions or the model. Treat supplier-facing communication with extra care: any message that states a commitment, a price or a deadline is approved by a named person before it is sent.

WORKING QUESTION

Who in your organization would be accountable if an AI-prepared summary led to a wrong supplier decision, and what would they have reviewed?



RISKS AND CONTROLS

Risks, controls and human review

05



The main risks are wrong extraction presented as fact, leakage of commercial data, biased or unfair supplier treatment and over-trust in fluent summaries.

Procurement data is commercially sensitive. Prices, volumes, margins, bids and negotiation positions have value to competitors and could be harmful if disclosed. Contracts often contain confidentiality clauses that restrict how information may be processed or shared. Before any supplier or contract data enters a tool, confirm what your agreements and policies allow and what the vendor does with inputs, including retention and use for training. Data protection, trade, sanctions, procurement and sector rules vary by place; check current requirements with



qualified advisers. Frameworks such as the NIST AI Risk Management Framework and ISO/IEC 42001 can structure governance, and obligations under laws such as the EU AI Act may apply depending on use. This playbook is not legal advice.

RISK	EXAMPLE	CONTROL
Wrong extraction	Renewal date or liability cap misread	Citations to clause text; sample review; mandatory check on high-value contracts
Commercial data leakage	Pricing pasted into an unapproved tool	Approved tool list; vendor terms reviewed; access limited by role
Unfair supplier treatment	Risk rating penalizes small or new suppliers for missing data	Review criteria; separate 'no data' from 'bad data'; periodic fairness check
Misclassification	Spend assigned to wrong category, distorting savings	Confidence thresholds; stratified sampling; correction loop
Supplier impersonation or fraud	Fake invoice or changed bank details in an email	Out-of-band verification; no AI approval of payment changes
Prompt injection	Supplier document contains hidden instructions	Treat files as data; limit tool actions; review inputs from outside
Over-trust	Summaries accepted without opening the source	Training; visible links to sources; audits

Controls to confirm before launch

- Data classes allowed in each tool are written down and enforced
- Outputs show the source passage or record they rely on
- Payment, bank detail and award decisions are never automated by the system
- High-value or high-risk items require a named second reviewer
- Logs capture inputs, outputs, reviewer and decision with retention agreed
- Supplier-facing messages are approved by a person before sending
- Supplier risk criteria are documented, explainable and open to challenge

EXAMPLE (ILLUSTRATIVE)

A contract tool flags a clause as an automatic renewal with ninety days notice. The contract manager opens the cited page, confirms the wording and records the date. The system never sends a termination notice; it only prepares the reminder.

WORKING QUESTION

Which single decision in your process would cause the most harm if made on an unchecked AI summary, and what is the check today?


MEASURING VALUE

Measuring quality and value

06



Test on your own documents and data, score accuracy separately from speed, and count the review effort that remains.

Start with a baseline. For each candidate use case, record how long the task takes now, how often errors occur, how many people touch it and what the delays cost. Then build a test set from real items: fifty to one hundred contracts clauses, spend lines or exception emails with answers that experts have agreed. Include awkward cases, such as scanned pages, unusual terminology, missing clauses and conflicting information. Without a test set, claims of accuracy are marketing.



USE CASE	QUALITY MEASURE	VALUE MEASURE
Spend classification	Accuracy by category and confidence band against expert labels	Analyst hours saved; share of spend classified; opportunities identified and acted on
Contract extraction	Field-level accuracy; correct 'not found' rate; citation accuracy	Review time per contract; time to answer clause questions; missed renewal incidents
Supplier risk	Share of alerts judged useful; missed events found later	Time from event to action; avoided disruption (estimated, with method stated)
Exception triage	Routing accuracy; draft acceptance rate	Time to first response; cycle time to resolution; expedite cost



USE CASE	QUALITY MEASURE	VALUE MEASURE
Demand support	Forecast error versus baseline method	Stockouts, excess inventory and working capital (measured, not assumed)

How to run an evaluation

01 Label the answers

Have two experts label the test items and resolve disagreements; their agreement sets the ceiling for what you can expect.

02 Run the system

Score each item as correct, partly correct or wrong, and record the type of error.

03 Weigh by consequence

A wrong liability cap matters more than a wrong cost center. Set higher pass marks for higher-consequence fields.

04 Compare with the baseline

Measure total time including review and correction, not just the generation step.

05 Monitor after launch

Re-sample monthly, since new suppliers, formats and category changes alter accuracy over time.

ILLUSTRATIVE FIGURES ONLY

If reviewing a contract takes 90 minutes manually and 35 minutes with prepared extraction and review, the saving is 55 minutes per contract. Multiply by your measured volume and subtract time spent fixing errors. Do not report savings you have not timed.

Be careful with claims about savings from better sourcing. Price improvements usually result from negotiation, competition and demand management; AI may help find the opportunity, but attribution should be agreed with finance before the program starts so that savings are not double counted. Report time and quality measures that you can observe directly, and treat estimated value as an estimate.

WORKING QUESTION

What accuracy would you need on contract dates and liability terms before a reviewer could skip a full read, and what would you do if you never reach it?



IMPLEMENTATION GUIDE

Implementation guide: one category, one process, one quarter

07



Prove value in a narrow slice with clear owners and honest measurement, then extend category by category.

01 Choose the slice

Pick one category or one process with enough volume, accessible data and an engaged owner, such as indirect spend classification or renewal tracking for a supplier group.



02 Prepare the data

Inventory sources, fix obvious master data problems, define the taxonomy and agree what data may be shared with the tool.

03 Write the instructions

Document the rules a good buyer applies, the clause library or the routing categories, with examples of correct and incorrect outputs.

04 Build the test set and baseline

Collect real items, label them and record current time and error rates.

05 Configure and tune

Run the tool on the test set, inspect errors, adjust rules and examples, and re-run until pass marks are met.



06 Pilot with reviewers

Release to a small group with an override log and weekly review, using the system as a draft only.

07 Decide

Compare with the baseline. Scale, change the scope or stop, and record why.

Data readiness

Before configuring anything

- A single supplier identifier exists, or a plan to create one
- Item and service descriptions are present and not mostly blank
- Category taxonomy is agreed and documented
- Contracts are findable, current and labelled with supplier and date
- Systems that hold orders, receipts and invoices can be joined
- Access rights and confidentiality limits are understood

Expect most effort to go into data and process, not software. If three systems each spell the same supplier differently, or half your contracts sit in personal folders, no model will rescue the work. Use the pilot to discover these issues cheaply. Plan change management as well: buyers and planners need to learn how to review an AI draft, when to override and how to give feedback that improves the system. Show them that the system removes drudgery rather than judgment.

PHASE	FOCUS	EXIT TEST
Days 1 to 30	Scope, data, baseline, test set	Test set labelled; baseline recorded; data rules agreed
Days 31 to 60	Configuration and evaluation	Pass marks met on test set; controls tested
Days 61 to 90	Supervised pilot	Measured gain with acceptable error and reviewer confidence
After day 90	Extend or stop	Business case based on observed results

WORKING QUESTION

Which single category or process, with accessible data and an engaged owner, could you pilot within the next quarter?

PITFALLS AND SCENARIO

Common pitfalls and an illustrative scenario

08

S.

Most disappointments come from unclear scope, weak data and unmeasured claims. They are avoidable.

- Buying a platform before defining the problem and the data it will run on
- Judging accuracy on a demonstration rather than a representative test set
- Letting the system act, such as sending supplier messages, before drafts have been proven
- Ignoring the confidentiality terms in supplier contracts when choosing tools
- Counting savings that belong to negotiation or demand changes as AI savings
- Forgetting small suppliers, new suppliers and multiple languages when testing
- Failing to retrain reviewers, so errors are rubber-stamped



ILLUSTRATIVE SCENARIO (HYPOTHETICAL)

A distributor builds a tool to classify indirect spend. In the demo it looks accurate. On a labelled sample of its own invoices it misclassifies many lines described only by part numbers. The team adds supplier-based rules, asks category managers to supply forty examples, routes low-confidence lines to a person and re-tests. Accuracy rises on the second run, and the analysts now review a short queue instead of every line.



Buying questions for vendors

QUESTION	WHAT TO LOOK FOR
How does the tool handle our confidentiality obligations and where is data stored?	Written terms on retention, location, training use and subprocessors
Can it cite the clause, record visible evidence for each field or statement line behind every output?	



QUESTION	WHAT TO LOOK FOR
How do we test accuracy on A pilot on your documents with agreed metrics our own data before committing?	
What controls exist for roles,Role-based access, review queues and exportable logs approvals and audit logs?	
How are taxonomies, clause Configurable by your staff without vendor tickets libraries and rules maintained by our team?	
What happens to our configuration if we leave?	Export of data, rules and history
What is the full cost at our volumes, including reviewing and integration?	A model at pilot and production scale, with assumptions stated

Build or buy

- Your ERP or source-to-pay suite already offers the capability and can be tested
- You need custom handling for unusual document types or strict controls
- You can staff evaluation and maintenance after launch
- Integration with existing systems is understood and costed
- An exit plan exists if quality or terms change

WORKING QUESTION

If a vendor showed you ninety percent accuracy tomorrow, what would you ask to learn what it was measured on?

30 / 60 / 90

A first quarter that produces evidence

09



Prove value on one category and one process under supervision, then extend with evidence.

First 30 days: ground it

- Select one category or process and name the business owner
- Map data sources, fix the supplier identifier and agree the taxonomy
- Record the baseline for time, errors and cycle time
- Build and label a test set from real documents and records
- Publish data rules for which tools may see supplier and contract information



Days 31 to 60: test it

- Configure the tool with rules, examples and a clause library
- Evaluate against the test set and tune until pass marks are met
- Test controls: access, citations, logs and second-review thresholds
- Train reviewers on checking drafts and logging overrides



Days 61 to 90: decide

- Run a supervised pilot with weekly error review
- Compare results with the baseline on time, quality and cycle time
- Agree with finance how value will be attributed
- Decide to extend to the next category, change the approach or stop



GLOSSARY

Terms used in this publication

10

TERM	MEANING
Spend cube	A consolidated view of spend by supplier, category, time and business unit.
Taxonomy	The agreed list of categories used to classify spend and items.
Master data	Core reference records such as suppliers and items that other systems rely on.
Entity resolution	Matching different spellings or records that refer to the same supplier.
Clause library	A set of standard contract positions used to compare agreements.
Retrieval	Finding the relevant record or passage and supplying it to the model with the question.
Citation	A link from an output to the passage or record that supports it.
Exception triage	Sorting problems such as late or short shipments by type and urgency.
Confidence band	A grouping of outputs by how sure the system is, used to direct review.
Baseline	Measured performance before the change, used for comparison.
Test set	Labelled real cases used to measure accuracy before and after changes.
Prompt injection	Hidden instructions in a document that try to steer the model.
Source-to-pay	The process from sourcing through contracting, ordering, receiving and payment.

A note on use

This publication is a public orientation resource, not legal, technical, financial or professional advice. Adapt it to your people, information, workflows and obligations, and check current tool terms, regulations and risks before acting.

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.

