



AI for retail and e-commerce

Applying AI to merchandising, content, service, search and operations without losing customer trust.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 42 / EMBER & LIME

EXECUTIVE WORKING EDITION - 08 OCT 2026

42

AT A GLANCE

Applying AI to merchandising, content, service, search and operations without losing customer trust.



Retail and e-commerce run on catalogs, content, customer conversations and thin margins, which gives AI many places to help and several places to do damage. This playbook covers where AI earns its place in catalog quality, product content, customer service, merchandising insight, search and returns, and where it should stay out. It explains the mechanics in plain language, sets out roles and review points, describes the risks that matter to brands, such as inaccurate product claims and mishandled customer data, and shows how to measure quality and value honestly. It closes with an implementation path, pitfalls and questions to ask vendors.

WHO IT IS FOR

Heads of e-commerce, merchandising, digital, customer care and operations in retail and direct-to-consumer businesses, plus the marketing, data and technology colleagues who support them.



TIME TO USE

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

What you will have at the end

- A ranked shortlist of retail use cases by value, data readiness and brand risk
- A clear line between what AI drafts and what a person must approve
- An operating model with owners for catalog data, content, service and measurement
- A test set and scorecard for product content and service answers
- A 90-day plan from a narrow pilot to a scale decision



Contents

01	Where AI helps a retailer and where it puts trust at risk	03
02	Where AI fits in retail, and where it does not	04
03	How the main retail use cases work, in plain language	06
04	Operating model, roles and ownership	08
05	Risks, controls and human review	09
06	Measuring quality and value	11
07	Implementation guide: start behind the storefront	13
08	Common pitfalls, an illustrative scenario and vendor questions	15
09	A first quarter that produces evidence	17
10	Terms used in this publication	18



EXECUTIVE BRIEF

Where AI helps a retailer and where it puts trust at risk

01



AI helps most with the volume work behind the storefront: describing, tagging, answering and sorting. It hurts most when it states something about a product or a customer that is not true.

A retailer's work is dominated by repeated, detail-heavy tasks. Every product needs a title, a description, attributes, images, categories and often translations. Every customer question needs a fast, accurate answer. Every return needs a decision. Every season brings new assortments, promotions and reviews to read. These are tasks of reading, classifying and writing at volume, which is where language and vision models are most useful, provided a person checks the output where it matters.

The risks are equally concrete. A model that writes that a jacket is waterproof when it is only water resistant creates returns, complaints and possibly regulatory exposure. A service assistant that invents a return policy damages trust faster than a slow reply. Personalization that feels intrusive drives customers away. Retail brands live on trust and on accuracy at scale, so the winning approach is to let AI prepare and propose while the source of truth, the product data and policy, stays controlled by named owners.

- Strongest fits: catalog enrichment, product content drafts, service answers grounded in policy, review and feedback summarization, search improvement and returns triage
- Weakest fits: unsupervised pricing changes, unverified product claims, decisions about individual customers that affect their rights or access
- Product data quality is the foundation; AI on weak data produces confident mistakes at scale
- Customer-facing output needs stronger controls than internal drafts
- Measure results with experiments and baselines, not with impressions

PLAIN TEST

If a wrong output would reach a customer without a person seeing it first, the controls should be stronger than the drafting tool itself.



II

WORKING QUESTION

Which of your daily retail tasks is mostly reading, tagging and rewriting, and which of them would embarrass you most if done wrongly?

WHERE IT FITS

Where AI fits in retail, and where it does not

02

S.

Sort use cases by how checkable the output is and how costly an error is. Start where both are favorable.

Retail use cases fall into three groups. Back-office drafting, such as product descriptions or meeting summaries, has a human editor and low blast radius. Customer-facing assistance, such as chat answers, reaches people directly and needs grounding and guardrails. Decision support, such as demand or markdown suggestions, affects money and depends on solid data and often on classical forecasting rather than a language model. The best first projects usually live in the first group and move outward as controls mature.



USE CASE	FIT	REASON AND CAUTION
Catalog enrichment (attributes, categories, tags from text and images)	Strong with sampling	Large volume, checkable against source; wrong attributes harm search and returns
Product descriptions and titles	Strong with review	Fast drafts; claims must be drawn from verified specifications
Customer service drafting for agents	Strong	Agent reviews each reply; grounded in policy and order data
Customer-facing chat on policy and order questions	Moderate	Needs retrieval from current policy, handoff to people and testing



USE CASE	FIT	REASON AND CAUTION
Review and feedback summarization	Strong	Finds recurring themes; sample for accuracy; watch for fake reviews
Search and discovery improvement	Moderate	Meaning-based search helps long queries; measure with experiments
Merchandising and assortment insight	Moderate	Combine analytics with AI explanation; judgment stays with merchants
Returns triage and fraud signals	Moderate	Use to route cases; do not deny customers automatically on model output
Dynamic pricing	Weak without strict controls	Fairness, legal and brand risks; keep rules and approvals explicit

A simple way to choose

Pick a first use case that meets most of these

- High volume of similar tasks that currently take real staff time
- Source data exists and can be checked, such as specifications or policy text
- A person can approve output before a customer sees it
- Errors are cheap to catch and correct
- The result can be measured against a clear baseline
- An owner in the business wants it and will review the results weekly

COMMON MISTAKE

Starting with a customer-facing chatbot because it is visible. Most retailers get faster, safer value from catalog quality and agent assistance, which also supply the clean data and policy text that a customer-facing assistant later needs.

WORKING QUESTION

Which two use cases on your list would rank highest on volume and checkability, and which would rank lowest on risk?

HOW IT WORKS

How the main retail use cases work, in plain language

03

Good systems feed the model verified facts, give it clear instructions and a required format, and route the result to a person or a check.

A language model writes from what it is given. For retail, the key design choice is what you give it. If you ask it to describe a product from the title alone, it will invent plausible details. If you give it the verified specification table, the brand's style guide and an approved example, it will write a description that sticks to the facts. The same principle governs service answers: the assistant should retrieve the current policy and the customer's order, and answer from those, citing the policy.

Catalog enrichment and product content

- 01 Gather verified inputs**
Pull specifications from supplier feeds or product information management, plus images and the style guide.
- 02 Extract and normalize**
Use the model to read titles, spec sheets and images and propose attributes such as color, material, size and compatibility, mapped to your attribute list.
- 03 Draft content**
Generate titles, descriptions and bullet points from the verified attributes, with required sections and a word limit.
- 04 Check**
Run rule checks for banned claims and missing fields, then route to an editor by risk, with regulated categories always reviewed.
- 05 Publish and learn**
Publish approved items, record edits made by editors, and feed common corrections back into instructions.

Customer service

For agent assistance, the system reads the customer message, looks up the order and policy, and drafts a reply with the sources visible. The agent edits and sends. For customer-facing assistance, the same design is used, with narrower permissions: it can answer from policy, check order status and hand off to a person when the question is outside its scope or the customer is upset. It should not issue refunds, change orders or promise exceptions unless a defined rule and an approval step allow it.

Search, merchandising and returns

Meaning-based search converts product text and queries into numerical representations so that a shopper typing 'warm jacket for wet weather' finds relevant items even without matching words. Merchandising insight uses AI to summarize reviews, search terms and sales patterns into themes that merchants can test. Returns analysis classifies reasons from free text, such as sizing, quality or description mismatch, and shows which fixes in content or product would reduce them.

USE CASE	KEY INPUTS	DRAFT OUTPUT	WHO APPROVES
Product content	Specs, images, style guide	Title, description, bullets	Content editor; compliance for regulated lines
Catalog attributes	Titles, specs, images	Attribute values with confidence	Data steward by sample
Service reply	Message, order, policy	Draft with cited policy	Agent
Review insights	Reviews and tickets	Themes and examples	Merchant or product manager
Returns reasons	Return notes and codes	Reason categories and patterns	Operations lead

WORKING QUESTION

For product content today, which verified source would you hand a writer, and does the model get that same source?

OPERATING MODEL

Operating model, roles and ownership

04

Product data, brand voice and policy each need an owner. AI makes weak ownership visible very quickly.

Retail AI sits across merchandising, marketing, service, operations and technology, so the risk is that everyone uses it and nobody owns it. Decide which functions own which sources of truth. Product facts belong to merchandising or product data. Brand voice and claims standards belong to marketing and legal or compliance. Policies belong to operations and service. The AI program does not own these; it consumes them, and it should surface where they are unclear.

ROLE	OWNS	TYPICAL TASKS
Head of e-commerce (business owner)	Use cases, outcomes and priorities	Chooses pilots, accepts results, manages trade-offs
Product data steward	Attribute definitions, specifications, taxonomy	Approves enrichment rules, samples outputs, fixes source data
Content lead and editor	Voice, accuracy and claims in content	Maintain style guide and examples, review drafts
Compliance or legal reviewer	Claims standards and regulated categories	Defines banned and restricted claims, reviews exceptions
Customer care lead	Policy answers and escalation	Maintains policy text, reviews service quality samples
Merchandising and pricing lead	Assortment and pricing decisions	Interprets insights; owns every price change
Data and privacy lead	Customer data use, consent and retention	Approves data flows, reviews vendors and logs
System owner	Tools, access and monitoring	Handles releases, incidents and vendor contact

Decision rights to settle early

- Which content may publish after a spot check and which always needs full review
- Which customer actions an assistant may take and which require a person
- Who may change prompts, banned-claim lists and policy text, and how changes are tested
- How edits by editors and agents are captured and fed back
- What triggers a pause: a spike in complaints, a wrong claim or a policy change
- Who speaks to customers or regulators if an AI-related error occurs



Keep a visible log of corrections. If editors repeatedly change the same phrase or attribute, the instruction or the data needs fixing. If agents regularly override a policy answer, the policy text may be ambiguous. Hold a short weekly review during a pilot and a monthly one afterward, with merchandising, content, service and data represented.

WORKING QUESTION

If an AI-written description made a wrong safety or material claim today, who would find out first and who would be accountable?



RISKS AND CONTROLS

Risks, controls and human review

05



The risks that matter most to retailers are inaccurate claims, wrong policy answers, mishandled customer data, unfair treatment and quality that drifts quietly.

Retail sits under consumer protection, advertising, product safety, privacy and sector rules that differ by country and by product category. Claims about materials, safety, origin, sustainability and health are especially sensitive. Handling of customer data, marketing consent, and automated decisions about individuals may be regulated. Frameworks such as the NIST AI Risk Management Framework and ISO/IEC 42001 help structure governance, and laws such as the EU AI Act may apply depending on use. This playbook is not legal advice; check current requirements with qualified advisers and translate them into controls.



RISK	EXAMPLE	CONTROL
Inaccurate product claims	Description says waterproof, specification says water resistant	Generate only from verified attributes; banned and restricted claim lists; review in regulated categories
Wrong policy answers	Assistant states a return window that has changed	Retrieve current policy; show source; test on policy changes; handoff to a person
Customer data exposure	Order details pasted into an unapproved tool	Approved tools; minimum necessary data; access by role; vendor terms reviewed
Unfair or opaque treatment	Returns flagged for certain customer groups	Use signals to route to a person; review outcomes by segment; no automatic denial
Brand voice drift	Tone becomes generic or off-brand	Style guide and examples; edit-rate tracking; periodic samples
Fake or manipulated reviews	Summaries treat spam as signal	Source filtering; flag suspicious patterns; sample reading
Prompt injection	Customer message or supplier file tries to change assistant behavior	Treat inputs as data; limit actions; monitor logs

Controls to confirm before customers see AI output

- Content is generated from verified product attributes, not guessed
- Banned and restricted claims are defined and checked by rule
- Policy answers come from the current policy, with a visible source
- The assistant hands off to a person for complaints, safety, legal threats and vulnerable customers
- The assistant cannot refund, discount or change orders outside a defined rule
- Customers are told when they are talking to an automated assistant, where required or appropriate
- Logs of inputs, outputs and handoffs are kept with agreed retention
- A fast switch exists to pause the assistant or revert content

EXAMPLE (ILLUSTRATIVE)

A home goods retailer lets a model draft descriptions only from the supplier specification fields. A rule check blocks words such as 'non-toxic' and 'hypoallergenic' unless an approved certification field is present. Editors see the blocked terms highlighted before approval.



WORKING QUESTION

What is the most damaging sentence your assistant could plausibly write about a product or a policy, and which control stops it?



MEASURING VALUE

Measuring quality and value

06

Judge quality on a test set of your own products and conversations, and judge value with before-and-after comparisons or controlled tests.



Retail has the advantage of rich data and the ability to run experiments. Use both. First measure quality directly: are the attributes right, are the descriptions faithful to the specification, are the service answers correct and in policy? Then measure business effects: time saved, conversion, returns, contact volume and resolution. Keep the two separate. A fast system that is wrong has negative value, and a modest accuracy gain may not move sales at all.

USE CASE	QUALITY MEASURE	VALUE MEASURE
Catalog enrichment	Attribute accuracy against a labeled sample; completeness	Share of products with complete attributes; search success; time per product



USE CASE	QUALITY MEASURE	VALUE MEASURE
Product content	Factual match to specification; clarity; violations per hundred items; edit rate	Content throughput; time to publish; conversion on tested pages
Service drafting	Policy correctness; tone; acceptance rate without edits	Handle time; first-contact resolution; satisfaction scores
Customer-facing assistant	Correct answer rate on test set; correct handoff rate	Contacts resolved without an agent, checked against repeat contacts and satisfaction
Search improvement	Relevance judgments on sample queries	Search-to-purchase rate and zero-result rate in a controlled test
Returns analysis	Accuracy of reason categories	Return rate for items where content or sizing was corrected

How to run the evaluation

01 Record the baseline

Capture current handle time, edit time, conversion, return rate and contact rate for the scope you will change.

02 Build a test set

Collect fifty to one hundred real products or conversations, including awkward ones: missing specifications, multiple languages, angry customers, unusual policies.

03 Score by consequence

Wrong safety or material claims count more than a clumsy phrase. Set stricter pass marks for riskier items.

04 Test in the market

Where volume allows, use a holdout group or a time-split test so changes in demand, promotions and seasonality are not mistaken for AI effects.

05 Track continuously

Sample weekly after launch. Product ranges, policies and seasons change, and quality changes with them.

ILLUSTRATIVE FIGURES ONLY

If an editor spends 15 minutes drafting a description and 5 minutes checking an AI draft, the saving is 10 minutes per product. For a thousand products that is about 167 hours before subtracting corrections. Use your own measured times, and do not assume that saved hours turn into cash.



Take care with attribution. Conversion moves with price, stock, traffic mix and promotions, so a rise after an AI change does not prove cause. Use controlled comparisons where possible and report ranges honestly. Include the cost of review, tools, data work and ongoing maintenance in the calculation.

WORKING QUESTION

Which metric would convince your finance lead that the pilot worked, and how will you separate the AI effect from season and promotion?


IMPLEMENTATION GUIDE

Implementation guide: start behind the storefront

07



Begin with content and agent assistance, strengthen the product and policy data, then extend toward customers as evidence builds.

01
Pick the slice

Choose one category or one service queue with enough volume and clear specifications or policies.


02
Fix the sources

Clean the attributes, confirm the style guide, write down the claims rules and update the policy text that the system will use.

03 Write the instructions and examples

Capture how your best editor or agent works: required sections, tone, forbidden claims, escalation triggers, with approved examples.

04 Set up the test and baseline

Label a test set and record current performance for the same scope.

05 Configure and tune

Run the tool, inspect failures, update instructions and data, and repeat until pass marks are met.

06 Pilot with human approval

Release to a small team with every output reviewed or sampled, an edit log and a weekly review.

07 Decide and extend

Compare with the baseline. Expand to more categories or to a customer-facing step only after controls prove themselves.

Data and content readiness

- Product attributes are structured, with agreed names and allowed values
- Supplier specifications are available and current
- A style guide with approved examples exists
- Restricted and banned claims are listed with owners
- Policies are written in plain language with effective dates
- Customer data used by the tool is limited to what the task needs

Do not skip the people work. Editors and agents need to learn how to review quickly without rubber-stamping, how to give feedback that improves the instructions and how to spot typical errors. Be explicit about what the system is for: removing repetitive effort so staff can spend more time on judgment, exceptions and customers. Share early results openly, including the failures, since trust in the process matters as much as trust in the output.

STAGE	FOCUS	EXIT TEST
Days 1 to 30	Scope, sources, claims rules, baseline, test set	Test set labelled; owners named; data rules set
Days 31 to 60	Configuration and evaluation	Pass marks met; controls tested; reviewers trained
Days 61 to 90	Supervised pilot	Measured effect with acceptable error and reviewer confidence

STAGE	FOCUS	EXIT TEST
After day 90	Scale, adjust or stop	Decision based on observed results and total cost

WORKING QUESTION

Which category or queue has the cleanest specifications or policy text, and so would make the best first slice?

PITFALLS AND SCENARIO

Common pitfalls, an illustrative scenario and vendor questions

08

Most failures come from weak product data, over-eager customer-facing launches and success claims that were never tested.

- Generating content from titles alone, so the model fills gaps with invention
- Launching a customer chat assistant before policy text and handoff paths are ready
- Treating edits as noise instead of the best source of improvement
- Measuring clicks or speed without checking accuracy or returns
- Using customer data in tools whose terms and retention are unknown
- Forgetting languages, regions and regulated categories in testing
- Letting tone drift until every product sounds the same

ILLUSTRATIVE SCENARIO (HYPOTHETICAL)

A fashion retailer asks a model to write descriptions from product titles. Fabric claims are wrong on a number of items and returns citing 'not as described' rise. The team switches to generating from structured fabric and care fields, blocks unverified claims by rule, and has editors sample each batch. Returns from description mismatch are tracked per category to see whether they fall.

Buying questions for vendors

QUESTION	WHAT TO LOOK FOR
Can the tool generate only from fields we supply and show which fields it used?	Source traceability and a refusal path when data is missing
How are banned claims and brand rules enforced?	Configurable rule checks, not only prompt wording
What customer data does it use and for how long?	Written terms on retention, location and training use process, where, and for how long?
How does the assistant hand off to a person and what context is passed?	Smooth transfer with conversation and order context
Can we test on our own catalog and conversations before buying?	A pilot with agreed metrics and our labelled test set
How do we monitor quality and roll back a change?	Dashboards, logs, versioning and a pause switch
What are the total costs at our volumes, including review and integration?	A model at pilot and production scale with assumptions stated

Build or buy

- Your commerce or service platform already offers the feature and can be tested
- You need custom controls for regulated categories or unusual data
- You can staff evaluation, content operations and monitoring after launch
- Integration with catalog, order and policy systems is understood
- An exit plan exists if quality, cost or terms change

WORKING QUESTION

If a vendor demonstrated fluent product copy on ten items, what would you ask to learn how it would behave on the other ten thousand?



30 / 60 / 90

A first quarter that produces evidence

09



Start with supervised content and agent assistance on clean data, then widen to customers once the controls have proven themselves.

First 30 days: ground it

- Choose one category or service queue and name a business owner
- Define verified sources, claims rules and policy text for the slice
- Record the baseline for edit time, handle time, returns and contacts
- Label a test set from real products or conversations
- Publish rules for which tools may see product and customer data

**Days 31 to 60: test it**

- Configure generation from structured fields with rule checks
- Evaluate against the test set and tune until pass marks are met
- Test handoff, logging, pause switch and policy-change handling
- Train editors and agents to review quickly and log corrections



Days 61 to 90: decide

- Run a supervised pilot with weekly review of edits and errors
- Compare results with the baseline, using a holdout where volume allows
- Review cost including review time and maintenance
- Decide to scale to more categories, add a customer-facing step or stop

GLOSSARY

Terms used in this publication

10

TERM	MEANING
Catalog enrichment	Adding or correcting product attributes, categories and tags so items can be found and compared.
Product information management	A system of record for product data, specifications and content.
Attribute	A structured product fact such as color, material or size.
Style guide	Written standards for tone, terminology and claims in brand content.
Grounding	Writing or answering from supplied facts rather than the model's general memory.
Retrieval	Finding the relevant policy or record and supplying it to the model with the question.
Handoff	Transferring a conversation from an assistant to a person with context.
Semantic search	Search that matches meaning rather than only exact words.
Holdout group	A comparison group that does not receive a change, used to measure its effect.
Edit rate	The share of drafts that reviewers change before approval.
Banned claim list	Words and claims that content may not use without verified support.
Return reason code	A category recorded for why an item was sent back.

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.

