



Enterprise generative AI

Securely applying general-purpose generative AI to
everyday knowledge work.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 01 / EMBER & MINT

EXECUTIVE WORKING EDITION - 08 OCT 2026

01

AT A GLANCE

Securely applying general-purpose generative AI to everyday knowledge work.

General-purpose generative AI can draft, summarize, analyze and explain, and most of your people are probably already using it, approved or not. This playbook shows how to move from scattered personal use to a governed, measurable capability. It covers where generative AI fits, how it works in plain terms, who owns what, which controls matter, and how to prove value before you scale. Use it to set a sensible first scope and a clear decision rule for each use case.

WHO IT IS FOR Executives, department heads and IT, security and operations leads who must decide how a whole organization uses general-purpose generative AI for knowledge work.

TIME TO USE 60 minutes to read; one working session to apply

What you will have at the end

- A shortlist of knowledge-work use cases ranked by value and risk
- A simple data-handling rule people can follow without reading a policy
- An ownership model with named roles and review points
- A baseline and a pass mark so value can be measured, not assumed
- A 90-day plan from pilot to a decision to scale, change or stop

Contents

01	What enterprise generative AI is, and what it is for	03
02	Where generative AI fits and where it does not	04
03	How generative AI works, 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: from first pilot to scale	13
08	Common pitfalls and an illustrative scenario	14
09	Buying and vendor questions	16
10	A first quarter that produces evidence	17
11	Terms used in this publication	18

EXECUTIVE BRIEF

What enterprise generative AI is, and what it is for

01



Generative AI is a drafting and reasoning assistant for language-heavy work. Its value comes from fitting it to specific tasks, not from general enthusiasm.

Generative AI systems produce text, summaries, tables, code and analysis from instructions written in ordinary language. For a business, the useful framing is narrow: these tools reduce the time it takes to get from a blank page or a pile of material to a reviewable first version. They do not replace the person who knows the customer, the contract or the consequence of a wrong answer.

Most organizations arrive at this topic from the bottom up. Staff open a public chatbot to rewrite an email, summarize a long report or build a spreadsheet formula, often pasting in material that was never meant to leave the company. Leadership then faces two bad options: ban the tools and lose the benefit, or ignore the behavior and accept the data risk. The better path is a governed default: an approved tool, a clear rule about what data may go in, and a short list of tasks where the benefit is proven.

- Good fit: first drafts, summaries, rewrites, translation of tone, structured extraction from messy text, brainstorming options, explaining unfamiliar material
- Poor fit: anything needing guaranteed accuracy without review, live facts the model cannot see, or decisions that must be made and defended by a named person
- The unit of value is a task done faster or better, measured before and after, not the number of licenses issued
- The unit of risk is the data that enters and the output that leaves, so controls attach to those two points



THE CORE IDEA

Treat generative AI as a capable but unreliable junior colleague. It works fast, writes fluently, and sometimes states wrong things with confidence. You review its work in proportion to what is at stake.

WORKING QUESTION

Which three tasks in your organization consume the most hours between receiving raw material and producing a first draft?

WHERE IT FITS

Where generative AI fits and where it does not

02

Choose use cases by the cost of a wrong output and the volume of repeated work, not by how impressive the demonstration looks.

A useful screen has two axes. The first is frequency and volume: does the task recur often enough that a saved twenty minutes adds up? The second is the cost of error: if the output is wrong and nobody notices, what happens? Tasks that are frequent and low in error cost are the natural starting point. Tasks that are rare and high in error cost should wait, or keep a human as the author with AI as a checker.

USE CASE	FIT	WHY
Summarize a long internal report for a briefing	Strong	Reviewer can compare to source; errors are visible and cheap
Draft a customer reply from a case note	Strong with review	Repeated task; a person sends the final text
Rewrite a policy in plain language	Strong with review	Owner checks meaning once, reuses the result



USE CASE	FIT	WHY
Analyze a spreadsheet Moderate export and propose findings	Moderate	Needs checks on arithmetic and on invented trends
Answer questions about current prices or stock	Weak	Model has no live data unless connected to a system
Decide credit, hiring or discipline outcomes	Not suitable	High impact on people; needs accountable human judgment and legal review

A quick screening test

Confirm before approving a use case

- The task happens at least weekly across the team
- A named person can review the output against a source or their own expertise
- The data involved is classified and allowed under your handling rule
- The cost of an undetected error is tolerable or caught by a later step
- You can measure time or quality for this task today, even roughly

COMMON MISTAKE

Starting with the most visible use case, such as a public-facing chatbot, instead of the most boring one. Internal drafting and summarizing teach you more, with less exposure.

WORKING QUESTION

Which proposed use case fails the screening test, and what would have to change for it to pass?

HOW IT WORKS

How generative AI works, in plain language

03

You do not need the mathematics. You do need a working model of why the tool is fluent, why it is sometimes wrong, and what you can change.

A large language model is trained on a very large body of text to predict what comes next in a sequence. After further training to follow instructions and behave helpfully, it can respond to requests in natural language. It produces text by choosing likely continuations, one piece at a time. This is why output reads smoothly. It is also why the model can produce a confident statement that is simply untrue: fluency and accuracy are separate properties.

Three practical facts follow. First, the model has a training cutoff and does not know your internal facts unless you provide them in the request or connect it to a source. Second, it works inside a context window, a limit on how much text it can consider at once, so very long inputs may be handled less reliably. Third, results vary: the same prompt can produce different answers, so quality must be judged across many runs, not one.

LEVER	WHAT IT CHANGES	TYPICAL OWNER
Prompt and instructions	Task framing, tone, format, constraints	Team that owns the task
Provided context	Facts the model can use in the answer	Task owner with data steward
Connected sources	Access to documents or systems at with security approval question time	
Model choice	Capability, speed, cost, data terms	Technology lead
Human review step	Whether errors reach a customer or decision	Process owner

01 Instruct

State the role, the task, the audience, the format and what to do when information is missing.

02 Ground

Paste or attach the source material so the model works from facts you supplied rather than from memory.

03 Generate

Ask for a draft, and for sources or quotes where claims matter.

04 Verify

A person checks claims, numbers and tone against the source before use.

05 Save

Keep prompts that work in a shared library so good practice spreads.

WHY THIS MATTERS FOR MANAGERS

If you only evaluate a tool on one impressive answer, you will overestimate it. Judge it on a set of realistic tasks, run more than once, scored against a clear standard.

WORKING QUESTION

Where in your current work would a fluent but wrong answer go unnoticed until it caused harm?

OPERATING MODEL

Operating model, roles and ownership

04

Adoption stalls when everyone is responsible and nobody is accountable. Name a small number of roles and give each a clear decision right.

Enterprise use needs a light structure, not a large committee. The structure answers four recurring questions: which tools are approved, which data may go where, which use cases are allowed, and who responds when something goes wrong. Small organizations can combine roles in one or two people. Larger ones should keep them distinct so that the person pushing for speed is not also the person approving the risk.

ROLE	OWNS	DECIDES
Executive sponsor	Outcomes and funding	Which business goals the program serves
Program lead	Roadmap, adoption, measurement	Sequence of use cases, when to scale or stop
Security and privacy lead	Data handling and vendor terms	Approved tools and permitted data classes
Use-case owner (per team)	Task, quality bar, review step	Whether outputs are good enough for their team
Data steward	Source quality and access	Which documents may be connected
Champions	Peer coaching and feedback	Nothing formal; they surface what works

Decision rights worth writing down

- Who may approve a new tool, and what evidence they need
- Who may approve a new use case, and at what risk level a second reviewer is required
- Who may connect internal documents or systems to an assistant
- Who can pause a use case immediately if quality drops or an incident occurs

- Who reports usage and outcomes to leadership, and how often

Adopt a simple tiering. Tier one covers low-risk personal productivity with public or non-sensitive data. Tier two covers team workflows using internal data, requiring a named owner and a test set. Tier three covers anything that touches customers, regulated data or consequential decisions, requiring formal review. Most requests will be tier one or two, and that is the point: the process should be fast where risk is low.

EXAMPLE (ILLUSTRATIVE)

A regional distributor lets any employee use the approved assistant for tier one tasks on day one. A team wanting to summarize supplier contracts registers a tier two use case with an owner and ten sample contracts. Approval takes days, not months.

WORKING QUESTION

Who in your organization could say stop today if an assistant started giving wrong answers, and do they know it?

RISKS AND CONTROLS

Risks, controls and human review

05

The main risks are data leakage, wrong output, over-reliance and unmanaged tools. Each has a practical control that does not require a large program.

Frameworks such as the NIST AI Risk Management Framework and ISO/IEC 42001 give a vocabulary for governing AI, and the EU AI Act and data protection laws set legal expectations in many places. This playbook does not give legal advice. Check which requirements apply to your sector and locations with qualified counsel, then translate them into controls that people can follow in the course of normal work.

RISK	WHAT IT LOOKS LIKE	PRACTICAL CONTROL
Data leakage	Staff paste contracts or personal data into unapproved tools	Approved tool with suitable data terms; plain data rule; access logging
Wrong or invented output	Fabricated figures, citations or policy statements	Review step scaled to stakes; ask for quotes from source
Over-reliance	Reviewers skim and approve without checking	Spot audits; sample checks; train reviewers on typical errors
Bias or unfair tone	Skewed language in people-related drafts	Exclude consequential people decisions; human judgment required
Intellectual property	Unclear ownership or reuse of generated content	Check vendor terms; policy on external publication
Shadow use	Personal accounts used because approved tool is slow or restricted	Make the approved option easy and good

Minimum controls before any team goes live

- Written data rule: what may and may not be entered, in one page
- Approved tool with reviewed terms on data retention and training use
- Named reviewer for outputs that leave the team
- Test set of realistic tasks and a recorded pass mark
- Route to report a bad output or a data mistake without blame
- Logging or usage visibility proportionate to the risk

Calibrating human review

- Low stakes, internal use: author reviews their own draft before use
- Medium stakes, shared or external use: second person checks facts and tone
- High stakes, regulated or consequential: formal review by an accountable expert, with the review recorded
- Always: no output is sent to a customer or filed as a record without a person taking responsibility for it

COMMON MISTAKE

Writing a long policy nobody reads. A one-page rule with three examples of allowed and forbidden inputs does more than forty pages of principles.

WORKING QUESTION

What is the single most likely way sensitive data leaves your organization through AI today, and what stops it?


MEASURING VALUE

Measuring quality and value

06



If you cannot show the before and after, you have a feeling, not a result. Set the baseline first and score outputs against a test set.

Measure at the task level. Before rollout, time a sample of real tasks done without AI and rate the outputs against a simple rubric. After rollout, repeat with the same kind of tasks. Self-reported time savings are useful but biased, so combine them with observed timings and quality ratings from a reviewer who does not know which version used AI. Avoid claiming organization-wide savings from a handful of enthusiastic users.



METRIC	HOW TO MEASURE	USE
Task time	Timed sample before and after, same task type	Value estimate
Output quality score	Rubric of 3 to 5 criteria, blind-rated by a reviewer	Quality gate
Rework rate	Share of drafts needing major correction	Reviewer burden
Error rate by type	Count factual, tone, format and missing-content errors in a sample	Risk and training
Active use	Weekly users as share of licensed users	Adoption health
Escalations and incidents	Count and severity of data or output issues	Governance signal



Building a test set

01 Collect real tasks

Gather twenty to fifty examples of the work, with realistic messiness, including awkward and incomplete inputs.

02 Define good

Write the rubric and agree on what counts as acceptable, with two or three worked examples.

03 Record the baseline

Score human-only results and note time taken.

04 Run and score

Run the assistant on the same tasks, several times, and score blind.

05 Set the pass mark

Agree in advance the threshold for go, fix or stop, and write it down before seeing results.

ILLUSTRATIVE NUMBERS ONLY

Suppose a team drafts weekly summaries in 40 minutes and the assisted version takes 15 including review. That is 25 minutes saved per summary. Multiply by volume, then discount for tasks where review found major errors. Use your own timings, not these.

Value also includes things that do not show up as hours: faster response to customers, fewer inconsistencies between documents, and better use of expert time. State these as hypotheses with a way to check them, such as response time or revision count, and be willing to report that a use case did not pay off.

WORKING QUESTION

For your top use case, what would you measure this week to establish a baseline, and who would do it?

IMPLEMENTATION GUIDE

Implementation guide: from first pilot to scale

07



Start small, learn with real work, and expand only when the numbers and the controls both hold.

01 Set the frame

Name the sponsor and program lead, choose the approved tool and publish the one-page data rule.

**02 Select three to five use cases**

Use the screening test from earlier. Favor frequent, reviewable, low-risk tasks across at least two departments.

03 Train by role

Teach prompting, grounding and review with the team's own examples, not generic demonstrations.

04 Pilot for four to six weeks

Run with a test set, a weekly feedback call and a shared prompt library.

**05 Review the evidence**

Compare to baseline and pass mark. Decide to scale, change or stop each use case.

06 Scale deliberately

Add teams in waves, with champions and the same measurement approach.

What good training covers

- How the tool works and why it errs, shown with failures from your own domain
- Writing clear instructions: role, task, audience, format, constraints, what to do if unsure
- Supplying source material and asking for quotes so claims can be checked
- A realistic review routine: what to check first, what to sample, when to escalate
- The data rule, with examples of permitted and forbidden inputs

**Pilot readiness**



- Use cases have named owners and a baseline
- Data rule is published and understood
- Tool access is provisioned through your normal identity system
- Test set and pass mark are written down
- Feedback channel is live and monitored
- Exit criteria exist, including the option to stop

Plan for the second-order effects. When drafting becomes faster, the bottleneck moves to review and approval. When more people can produce analysis, the quality of the question matters more. Expect job content to shift and talk about it openly with the people affected. Workforce training should be treated as part of the implementation, not an afterthought, because adoption depends more on confidence and habit than on features.



WORKING QUESTION

What would a six-week pilot need to show for you to commit budget to a wider rollout?



PITFALLS

Common pitfalls and an illustrative scenario

08

Most failures are predictable and organizational, not technical. Recognizing them early is cheaper than recovering.



- Buying licenses before choosing use cases, leading to low usage and unclear value
- Banning tools without offering a usable alternative, which drives shadow use
- Measuring enthusiasm instead of task outcomes
- Skipping the review step once early outputs look good
- Letting every team write its own rules, producing inconsistent risk positions
- Treating the pilot as the end rather than the start of a managed capability

ILLUSTRATIVE SCENARIO (HYPOTHETICAL)

A mid-sized professional services firm issues licenses to all staff with a one-hour demo. Three months later, only a fifth use the tool weekly, a partner has pasted a client file into a personal account, and no one can say what it saved. The firm resets: one approved tool, a one-page data rule, four use cases with owners, and a monthly review. Usage becomes purposeful and measurable.

WARNING SIGN	LIKELY CAUSE	RESPONSE
Usage drops after month one	Tool not tied to real tasks	Rework use cases with teams; share working prompts
Reviewers approve everything	Review is a formality	Introduce sampling and seeded errors
Complaints about wrong answers	No grounding or test set	Supply sources; define quality bar
Staff use personal tools	Approved option is slow or limited	Fix access and speed; restate the data rule
Savings claims cannot be verified	No baseline	Measure a sample now and restate claims

The pattern behind these pitfalls is the same: technology was introduced without changing how work is defined, reviewed and measured. Generative AI is a change in working practice. Treat it with the same discipline you would give a new process or system, including ownership, standards and the willingness to stop what does not work.

WORKING QUESTION

Which of these pitfalls is already visible in your organization, and who can act on it this month?

BUYING AND VENDORS

Buying and vendor questions

09

Tool choice matters less than terms, controls and fit with how your people work. Ask for answers in writing and verify the important ones.

Do not choose on the strength of a demonstration. Run your own tasks through the shortlisted tools using your test set. Involve security and privacy early, because contract terms about data use and retention are hard to change after rollout. Check current vendor documentation and certifications yourself rather than relying on sales claims, and keep exit options open by avoiding deep dependence on one vendor's proprietary features until value is proven.

QUESTION	WHAT A GOOD ANSWER LOOKS LIKE
Is our data used to train models?	Clear written terms that business data is not used for training by default
Where is data processed and stored, and for how long?	Specified regions, retention settings you can configure
How is access controlled?	Single sign-on, role-based access, audit logs
What admin visibility do we get?	Usage reporting and policy controls at organization level
How are incidents and changes communicated?	Defined notification process and change notices for model updates
How do we leave?	Data export and deletion on exit, documented

Evaluation checklist

- Your test set run on each shortlisted option, scored blind
- Security and privacy review completed and recorded
- Total cost modeled at pilot and at full volume, including review time
- Integration with identity and existing systems confirmed
- Training and support included or priced

- Plan for model changes that alter behavior after you have tested

WATCH FOR

Pricing that looks cheap per user but ignores usage-based charges, and features marketed as available that are still on a roadmap. Ask what is generally available today.

**WORKING QUESTION**

Which contract term, if it turned out to be unfavorable, would force you to change course after rollout?



30 / 60 / 90

A first quarter that produces evidence

10

Move from scattered personal use to a governed, measured capability in ninety days.

**First 30 days: ground it**

- Name the sponsor, program lead and security owner
- Publish the one-page data rule and approve one tool
- Select three to five use cases and record baselines for each
- Draft the test set and pass mark for the top two use cases



Days 31 to 60: test it

- Train pilot teams with their own examples
- Run the pilot with weekly feedback and a shared prompt library
- Sample outputs for errors and record rework rates
- Fix access and speed problems that drive shadow use

Days 61 to 90: decide

- Compare results to baseline and pass marks
- Decide scale, change or stop for each use case
- Plan the next wave with champions and a measurement routine
- Report outcomes and incidents to leadership

GLOSSARY

Terms used in this publication

11

TERM	MEANING
Generative AI	AI that produces new content such as text, images or code from instructions.
Large language model	A model trained on large amounts of text to predict and generate language.
Prompt	The instruction and context given to the model for a task.
Grounding	Supplying source material so the model answers from facts you provide.
Hallucination	A fluent output that is false or unsupported by any source.
Context window	The limit on how much text the model can consider in one request.
Test set	A fixed group of realistic tasks used to score quality consistently.
Human in the loop	A design where a person reviews or approves outputs before they are used.
Shadow AI	Use of AI tools by staff outside approved channels and controls.

TERM	MEANING
Data classification	Labeling data by sensitivity so handling rules can be applied.
Baseline	A measurement of current performance taken before introducing a change.
Prompt library	A shared collection of tested prompts for recurring tasks.



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.

