



Multimodal AI

Using text, image, audio, video, and documents together to understand business work.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 04 / EMBER & AMBER

EXECUTIVE WORKING EDITION - 08 OCT 2026

04

AT A GLANCE

Using text, image, audio, video, and documents together to understand business work.

Multimodal AI systems work with more than one kind of input, such as photos, scanned pages, recordings, video and text, and can relate them to one another. A model can read a drawing and a work order together, compare a site photo with a specification, or turn a recorded call into a structured summary. This playbook explains where multimodal systems add real value, where they are unreliable, how to handle sensitive images and voices, and how to test them with realistic evidence rather than clean demonstrations.

WHO IT IS FOR Operations, quality, field, service and technology leaders who work with images, audio, video or complex documents and want to know where AI can reliably help.

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

What you will have at the end

- A shortlist of multimodal use cases with a view of value and risk for each
- Clear rules for capturing, storing and using images, audio and video
- A test set built from real, messy evidence rather than clean samples
- A review design that matches the cost of a missed or false finding
- A 90-day plan from feasibility check to a scale decision

Contents

01	What multimodal AI means for the business	03
02	Where multimodal AI fits and where it does not	04
03	How multimodal AI works, in plain language	06
04	Operating model, roles and ownership	07
05	Risks, controls and human review	09
06	Measuring quality and value	11
07	Implementation guide: from feasibility to a supervised pilot	13
08	Common pitfalls, an illustrative scenario and vendor questions	14
09	A first quarter that produces evidence	16
10	Terms used in this publication	17

EXECUTIVE BRIEF

What multimodal AI means for the business

01



Much of the evidence in a business is not typed text. Multimodal AI lets software read photos, pages, recordings and video, then connect them to the text around them.

Think about where your organization keeps its real information. An inspection is a set of photos with handwritten notes. A customer complaint is a recorded call. A construction record is a drawing, a schedule and a message thread. A claim is a form, a photo and an email. Until recently, software could index the typed parts and had to ignore the rest. Multimodal models can accept several types of input in one request: they can describe an image, read text inside it, transcribe and summarize audio, and reason across a document that mixes tables, diagrams and prose.



This widens what can be automated or assisted, but it does not remove the need for judgment. A model that describes a photo is making an interpretation, and interpretations can be wrong in ways that are hard to notice, such as misjudging scale, missing a small defect, or confusing similar parts. The value is real where AI prepares the evidence for a person: sorting, describing, flagging, comparing and summarizing. It is far less reliable as the sole decision-maker for safety, quality acceptance or legal findings.



- The most dependable uses turn unstructured evidence into structured records that a person can verify
- Image and audio quality set the ceiling: poor lighting, noise and low resolution cause most failures
- Different tasks need different tools: a specialized vision model for a fixed defect may beat a general model
- Privacy and consent matter more than with text, because faces, voices and locations identify people



CORE IDEA

Use multimodal AI to make evidence easier to review, not to make the decision disappear. The question is always what a person sees afterward and how quickly they can confirm it.

WORKING QUESTION

Where does your organization hold important evidence in photos, recordings or scans that nobody has time to review?

WHERE IT FITS

Where multimodal AI fits and where it does not

02

Match the modality to a specific task, then check that the cost of a miss is acceptable and that inputs are consistently usable.

Group candidate uses by what the system produces. Some tasks create text from non-text, such as transcripts, captions and report drafts. Some compare two sources, such as an as-built photo against a specification. Some classify or flag, such as whether a photo shows required labels. Some answer questions over mixed documents, such as asking a question of a drawing set. Each has a different failure profile, and the right control differs accordingly.

USE CASE	FIT	WHY
Summarize recorded service calls into structured notes	Strong	Output is reviewable against the recording; saves wrap-up time
Draft field inspection reports from photos and voice notes	Strong with review	Inspector confirms and corrects before filing

USE CASE	FIT	WHY
Read mixed documents with tables, stamps and handwriting	Moderate to strong	Quality varies by scan; needs validation of key fields
Flag missing labels or PPE in site photos	Moderate	Depends on consistent angles and lighting; false negatives matter
Detect fine defects in manufactured parts	Specialized	Often better served by dedicated inspection systems
Judge safety-critical acceptance from a photo alone	Not suitable	Cannot replace qualified inspection and measurement
Identify individuals from faces or voices	Avoid	High privacy and legal risk; check applicable law

Signals that a use case is ready

Confirm before investing

- Inputs are captured in a reasonably consistent way, or capture can be standardized
- A person can check the output against the original image, audio or page quickly
- The cost of a missed finding is understood and acceptable with review
- Rights to capture and process the material are clear, including consent where needed
- You have at least a few hundred real examples, including poor ones, for testing
- The task is frequent enough to justify setup and monitoring

COMMON MISTAKE

Piloting with the best photos in the archive. Production images come from tired people with dirty lenses in bad light. Test with those.

WORKING QUESTION

Which of your candidate tasks would still work if the input were the worst ten percent of what you collect?

HOW IT WORKS

How multimodal AI works, in plain language

03

The model turns each kind of input into a shared internal form it can reason over, then produces text or structured output. Quality depends on what it can see and hear.

Multimodal models learn from large collections that pair images, audio or video with text. Internally, an image is converted into numerical features and processed alongside the text of your request, so the model can answer questions about the image, read text embedded in it or compare it with another. Audio is usually converted to text by speech recognition first, sometimes with speaker labels and timestamps, and then handled by a language model. Video is generally analyzed as sampled frames plus audio, which means fast events between frames can be missed.

INPUT	TYPICAL PROCESSING	COMMON WEAKNESSES
Photos	Description, text reading, comparison, classification	Small details, scale, glare, unusual angles
Scanned pages	Optical character recognition plus layout understanding	Handwriting, stamps, skewed or faint scans, complex tables
Audio	Speech-to-text with optional speaker separation	Accents, jargon, crosstalk, background noise
Video	Sampled frames plus audio	Missed events, long duration, cost
Mixed documents	Joint reading of text, tables and figures	Figures misread; units and legends lost

01 Capture

Collect the input with enough quality and the right metadata such as time, location, asset and who captured it.

02 Pre-process

Normalize format, rotate, crop or enhance as needed, and remove or mask personal data where possible.

03 Interpret

The model describes, extracts or compares according to a specific, written instruction.

04 Structure

Output is placed into defined fields with confidence indicators and links back to the original evidence. 

05 Review

A person confirms, corrects or rejects, and the corrections are saved for evaluation.

WHY THIS MATTERS

A fluent description is not proof the model saw what matters. Ask it to point to the evidence, such as the region of the image or the time in the recording, so a reviewer can check quickly. 

WORKING QUESTION

For your top use case, can a reviewer jump from each AI statement straight to the evidence behind it?

OPERATING MODEL

Operating model, roles and ownership

04

Multimodal work crosses departments that rarely share ownership: field teams, quality, IT, legal and privacy. Decide who owns capture, interpretation and decisions. 

The common failure is that the people who capture the evidence, the people who build the system and the people who decide on its findings sit in different parts of the organization. Capture standards drift, the model is tuned on old data, and reviewers do not trust the results. Assign an owner for the whole chain, then specific owners for each link. The process owner is

usually the business leader responsible for the outcome, such as quality or service, not the technology team.

ROLE	OWNS	DECIDES
Process owner	Outcome and quality bar	What the system may flag, draft or decide
Capture lead (field or service)	How evidence is gathered	Capture standards: angles, lighting, scripts, devices
Reviewer or inspector	Verification of findings	Accept, correct or reject each output
Data and privacy lead	Consent, retention, masking	What is stored, for how long and who may view it
Technical owner	Pipeline, model configuration, integration	Releases, monitoring, change control
Evaluation lead	Test sets and scoring	Whether a change is good enough to deploy

Capture standards that pay off

- Photo guides: required angles, distance, lighting and a scale reference in frame
- Voice note scripts: state the asset, location, date and finding in a consistent order
- Metadata captured automatically: time, device, location where lawful and appropriate, asset identifier
- Quality gates at capture: reject blurred or dark images immediately, while the person is still on site
- Short training for the people in the field, showing examples of good and bad evidence

Assign someone to maintain the test set as the work changes. New equipment, new forms, new camera models and new accents all alter inputs. The evaluation lead should add fresh examples regularly and retire stale ones, so scores reflect today's reality. Without this routine, performance slides quietly while dashboards still show last quarter's good numbers.

WORKING QUESTION

Who today would notice if the quality of photos arriving from the field slowly got worse?

RISKS AND CONTROLS

Risks, controls and human review

05



Images, voices and video carry personal and commercial information that text often does not. Controls must cover consent, storage and wrong findings.

Faces, voices, vehicle plates, screens in the background and location data can all identify people. Rules differ by jurisdiction and context, including data protection law, employment and workplace monitoring rules, and recording consent requirements. The EU AI Act also restricts certain uses of biometric and emotion-related systems. This playbook does not give legal advice. Check current requirements with qualified counsel before capturing or analyzing recordings of people, and apply the NIST AI Risk Management Framework or ISO/IEC 42001 as a structure for documenting risk.



RISK	EXAMPLE	CONTROL
Privacy and consent	Recordings capture staff or customers without notice	Notice and consent where required; masking; avoid unnecessary capture
Biometric inference	Attempts to infer identity or emotion from face or voice	Avoid; seek legal review before any such use
False negative	Defect or hazard present but not flagged	Human inspection for critical items; test miss rates; conservative thresholds
False positive	Unneeded rework triggered by wrong flags	Review queue; track precision; tune thresholds
Misread documents	Wrong amount or date extracted from a scan	Field validation; cross-checks; review of key fields
Sensitive storage	Large media archives with weak access control	Retention limits; access by role; encryption; deletion routine
Manipulated media	Altered photos or generated evidence submitted	Provenance checks; metadata review; escalate suspicious cases



Controls to confirm before go-live

- Lawful basis and notices for capture are confirmed with counsel
- Personal data is minimized, masked or excluded where possible
- Retention periods are set and enforced for media files
- Access to raw media is limited by role and logged
- Critical findings always receive human review
- Reviewer corrections are captured for evaluation
- Suppliers' terms on media storage and model training are reviewed

EXAMPLE (ILLUSTRATIVE)

A building services firm uses AI to pre-sort photos of equipment labels. The system proposes serial numbers and flags photos too blurry to read. Technicians confirm each number on screen. Faces are blurred at upload and raw files are deleted after a fixed period.

WORKING QUESTION

What is the worst thing a missed finding could cause, and which person would catch it?

MEASURING VALUE

Measuring quality and value

06



Measure against ground truth made by experts on realistic evidence, and report the errors that matter, not an average accuracy figure.

Averages hide the costly errors. A system that is correct on most photos but misses the rare serious defect is not safe. For each use case, define the error types and rank their cost. Build a test set from real evidence, labeled by qualified people, with a spread of good, poor and edge cases. Record agreement between experts, because if two inspectors disagree often, a model cannot be judged against a single answer without noting that ambiguity.



METRIC	WHAT IT TELLS YOU	NOTE
Recall (miss rate)	Share of real issues the system flags	Most important for safety and quality checks
Precision	Share of flags that were real issues	Avoids wasted review effort
Field accuracy	Share of extracted values that are correct	Report per field, not only overall
Transcript word error rate	How accurate speech-to-text is	Measure on your accents, jargon and noise
Unusable input rate	Share of inputs too poor to process	Capture-quality signal
Review time per item	Effort left for people	Real savings measure
Escalation accuracy	Share of uncertain items correctly passed to humans	Calibration check



01 Baseline

Time the current process and record error rates and rework from a sample of real cases.





02 Label

Have experts annotate a test set of at least a few hundred items, with a clear guide and double labeling on a subset.

03 Slice

Score by condition: lighting, device, site, language, document type. Weak slices guide fixes.

04 Compare

Run the system, measure against labels, and compare time and quality with the baseline.

05 Set thresholds

Decide the confidence level above which items proceed and below which people review.



ILLUSTRATIVE FIGURES ONLY

If inspectors spend 25 minutes writing up a visit and AI-prepared drafts take 10 minutes to check, that is 15 minutes saved per visit. If one visit in twenty needs a full rewrite, subtract it. Use your measured numbers, not these.

Track drift. A new camera, a changed form layout or a new product line can reduce accuracy without anyone noticing. Re-run the test set monthly and after every system change, and sample live outputs for expert audit.



WORKING QUESTION

Which single error type would force you to switch the system off, and do you measure it?



IMPLEMENTATION GUIDE

Implementation guide: from feasibility to a supervised pilot

07



Prove that the inputs are good enough and the outputs are checkable before building integrations.

- 01 **Pick one task and one modality**
For example, drafting reports from photos and voice notes for one inspection type. Avoid combining everything at first. 
- 02 **Collect real evidence**
Gather a few hundred real inputs from different sites, devices and conditions, with permission.
- 03 **Run a feasibility test**
Try two or three approaches on the sample and compare against expert labels. Include a specialized option if one exists.
- 04 **Standardize capture**
Fix guides, scripts and quality gates. Many gains come from better inputs, not better models. 
- 05 **Design the review screen**
Show output beside the original evidence, with fields editable and confidence visible.
- 06 **Pilot and learn**
Run with a small group, collect corrections, and review failures weekly.
- 07 **Decide**
Compare to baseline and thresholds. Scale, change or stop.

Feasibility test

- At least 200 real examples across normal and difficult conditions
- Expert labels with documented guidance
- Results sliced by condition, not just averaged
- Costs per item estimated, including storage and processing



- Privacy review completed for the sample data
- Clear definition of what will be reviewed by a person



Keep the review interface simple. Reviewers will judge the system by how fast they can confirm or fix its output. Place the original photo, page or audio clip next to the proposed text, highlight the region or timestamp the model used and make correction a single action. Every correction is training and test material for the next version, so store it with the original input and the reviewer's identity and date.

COMMON MISTAKE

Integrating deeply with core systems before proving accuracy on real evidence. Prove value in a standalone review tool first, then integrate.



WORKING QUESTION

What is the smallest real-world pilot that would show whether your inputs are good enough?



PITFALLS AND VENDORS

Common pitfalls, an illustrative scenario and vendor questions

08

Demonstrations on clean samples mislead. Insist on your own evidence, your own conditions and your own error costs.



- Judging a system on vendor demos rather than your messy inputs
- Reporting overall accuracy while ignoring the rare, expensive miss
- Capturing recordings of people without a clear purpose, notice or retention rule
- Using a general model where a specialized inspection tool is more reliable
- Skipping capture standards and hoping the model compensates
- Letting AI-written summaries replace the original evidence in records

ILLUSTRATIVE SCENARIO (HYPOTHETICAL)

A food packaging plant tests AI on label photos supplied by the vendor and sees excellent results. On the line, glare and motion blur cut accuracy sharply. The plant adds a fixed camera position, lighting and an unreadable-image alert, then retests on two weeks of its own images. Results improve, and humans still verify allergen declarations.



VENDOR QUESTION	WHAT TO LOOK FOR
Can we test on our own images, audio and documents?	Yes, before contract, with results broken down by condition
How do you handle accents, noise, handwriting and poor scans?	Documented performance and configuration options
Where is media stored, for how long, and is it used for training?	Clear written terms, regional options, deletion controls
Can outputs link to the exact region or timestamp used?	Evidence references that support quick review
What happens when the model or version changes?	Notice, version control and a way to re-test
What does cost scale with?	Per image, per minute of audio or video, storage and review tooling



WORKING QUESTION

If your vendor's accuracy claim does not hold on your site photos, how would you find out before rollout?



30 / 60 / 90

A first quarter that produces evidence

09

Establish that your real inputs are good enough, then build the review process around them.

First 30 days: ground it

- Choose one task and one modality with a named process owner
- Collect at least 200 real examples with permission and note conditions
- Draft capture guides and a privacy and retention rule
- Time the current process to set a baseline

Days 31 to 60: test it

- Have experts label a test set and measure agreement
- Run a feasibility comparison of two or three approaches, sliced by condition
- Fix capture quality gates based on the weakest slices
- Build the side-by-side review screen

Days 61 to 90: decide

- Pilot with a small group and capture corrections
- Track recall, precision, review time and unusable-input rate weekly
- Re-test monthly and after any change
- Decide to scale, change or stop; plan the next modality or use case

GLOSSARY

Terms used in this publication

10

TERM	MEANING
Multimodal	Able to work with more than one kind of input, such as text, images and audio.
Optical character recognition	Software that reads printed or handwritten text from images and scans.
Speech-to-text	Conversion of spoken audio into written transcript.
Word error rate	The share of words in a transcript that are wrong, missing or added.
Recall	The share of real issues that the system successfully flags.
Precision	The share of flagged items that turn out to be real issues.
Ground truth	Expert-verified labels used as the standard for scoring.
Confidence threshold	The score above which an output proceeds without human review.
Provenance	The recorded origin and history of a file, used to assess authenticity.
Data minimization	Collecting and keeping only the data needed for the purpose.
Drift	A gradual change in inputs or behavior that reduces accuracy over time.
Biometric data	Data about physical or behavioral traits used to identify a person.

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.

