Invisible

INVISIBLE TECHNOLOGIES

Beyond text: Why multimodal AI demands a different playbook

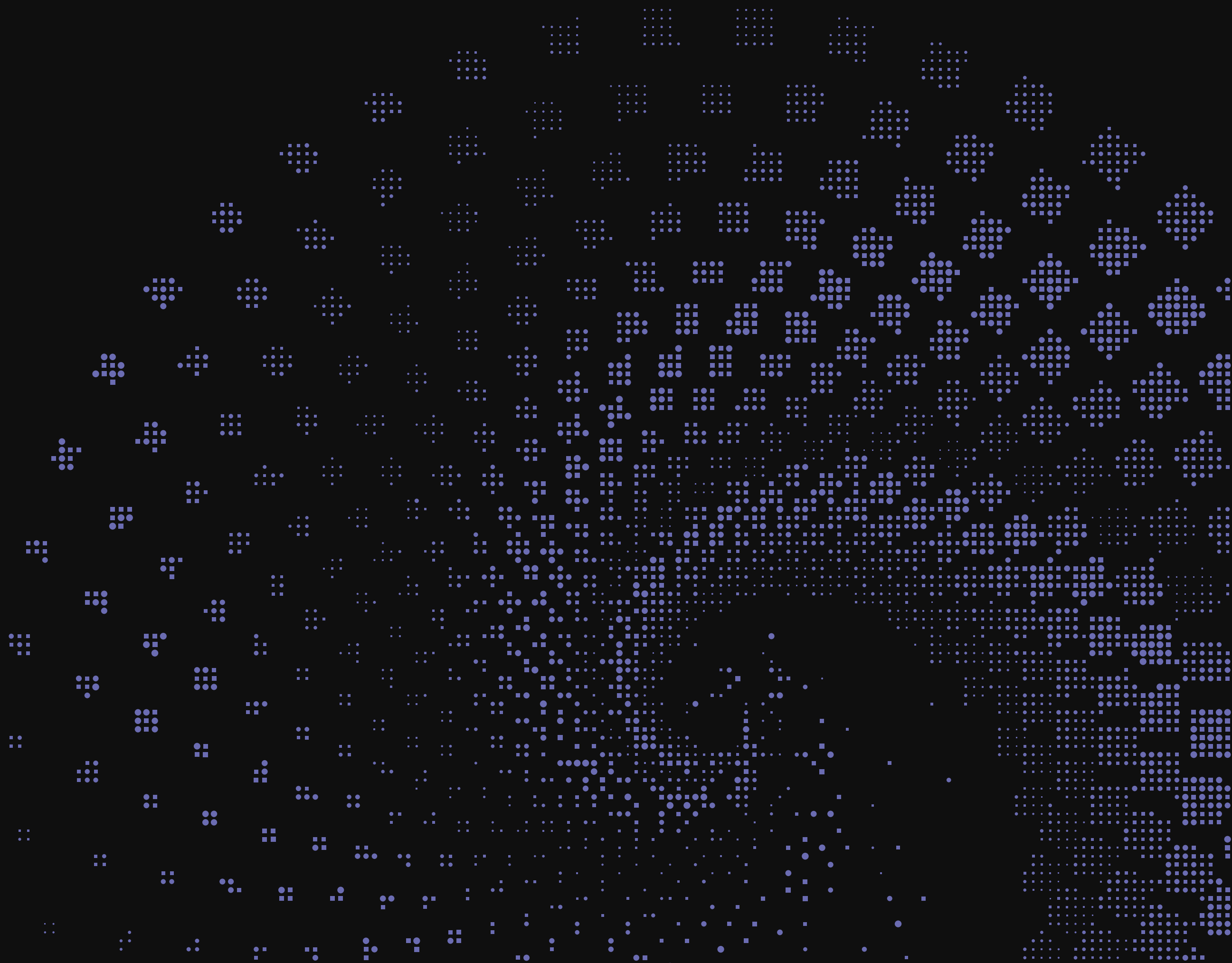


Table of contents

Executive summary

Information about the real world isn't found in text alone, but in sights, sounds, and sensory inputs. That's why multimodal data matters: it opens up possibilities for AI to be able to understand and interpret our world in a deeper, richer way.

Imagine a chatbot that doesn't just respond to written interactions with a customer, but can "read" screen grabs or "listen" to audio messages. Think about security cameras with an AI that analyzes millions of hours' worth of footage in an instant. Text models alone won't get us there.

Combining text with other data types such as images, audio, video and structured data, enables richer context, more accurate decisions, and new classes of enterprise applications beyond what text-only AI can deliver. Multimodal AI systems that process and integrate data from multiple sources including text, audio, images, and sensors are transforming industries from autonomous vehicles to agriculture. From call center transcripts to smart surveillance and natural language assistants, AI's ability to interpret and learn from diverse data streams is unlocking new possibilities.

But with these possibilities come new predicaments. As the complexity of AI models grows, so too do the opportunities and challenges. Along with the benefits, there are higher demands on data, infrastructure, governance and change management, making successful deployment significantly more complex than implementing text-only AI solutions. Training an AI model on a single mode—language—is very different to being trained on other modes like seeing, hearing, or reading.

Executive summary

Simply put, the stakes are higher: we're not just dealing with a chatbot hallucinating a discount code or an automated tool mislabeling a routine billing inquiry, albeit both are critical for customer trust. A multimodal AI problem could result in a factory accident, or an AI model mistakenly identifying a hazardous item as a candy bar at an airport screening. In other words, serious, real-world consequences. Enterprises exploring multimodal applications must confront not only technical hurdles, but also profound ethical and safety considerations. Getting it right calls for a deep understanding of both the technical and ethical dimensions, rigorous data practice, and thoughtful experimentation.

The processes and disciplines that underlie successful multimodal machine learning are markedly different from those used for a single mode like text. And it's why a thoughtful, structured approach is essential to achieve high performance and to ensure sound, unbiased, and efficient machine learning systems.

Alert
High threat detected
Classification: Explosive (IED)
Confidence Score: 94%

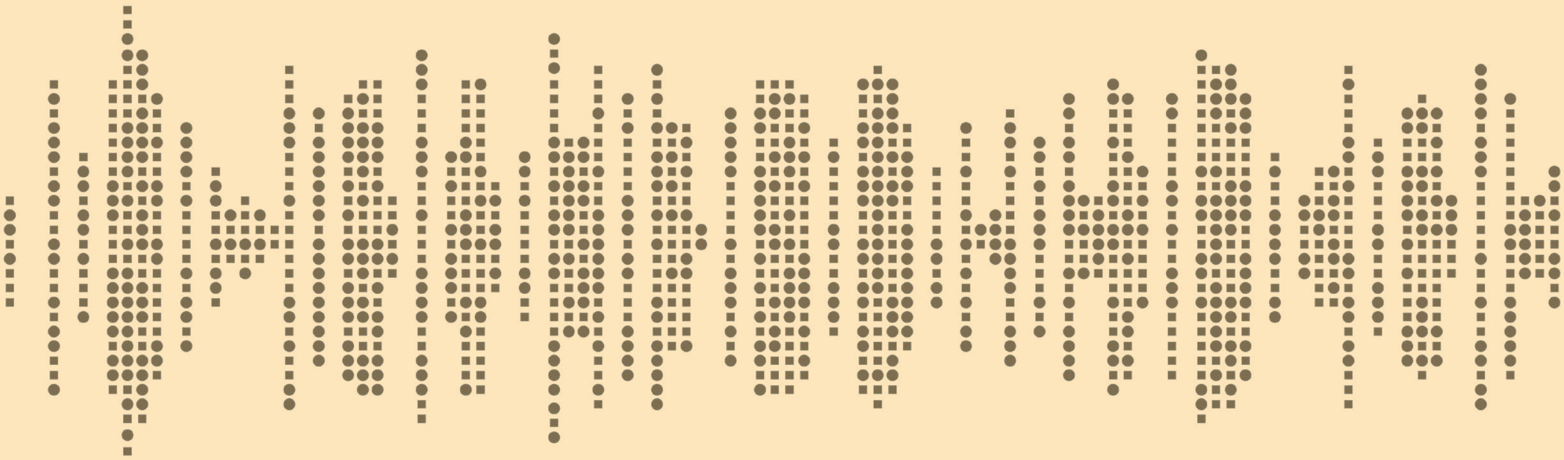
In this paper, we make the case for how to approach multimodal AI intentionally. The possibilities are real. So are the failure modes. The difference between them is design.

What is multimodal AI?

Multimodal AI refers to systems that reason across multiple data types simultaneously—text, images, audio, video, sensor data—rather than text alone. That sounds like an incremental upgrade. In practice it's a different class of problem, with different data requirements.

Cognitive tasks that humans consider complex—playing chess at grandmaster level, solving difficult equations, detecting patterns in financial data—are relatively straightforward for AI. But basic actions people perform unconsciously, like glancing at a sports game and instantly distinguishing players from fans, or picking up a coffee cup without knocking it over, are genuinely hard for machines. We train ourselves through lived experience. We absorb context without noticing we're doing it.

This matters more than it sounds. Because the moment you move beyond text, you're asking AI to operate in that second category, in the domain of perception, context, and judgment that humans have spent a lifetime calibrating.

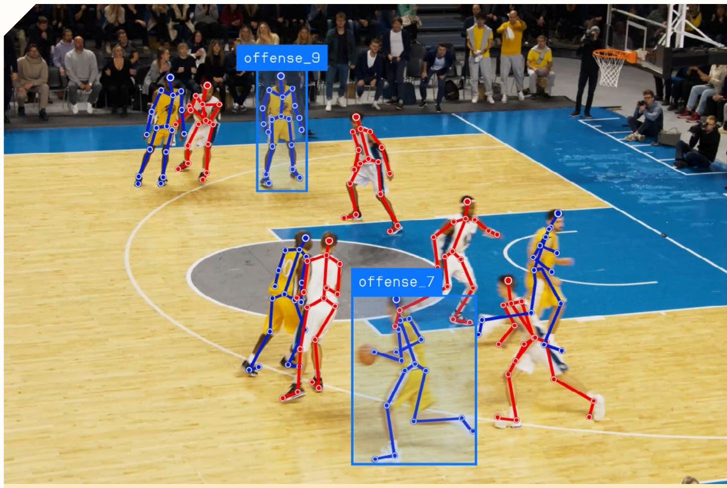


What is multimodal AI?

What multimodal makes possible

The technology isn't theoretical. Multimodal systems are in production across industries, and the use cases that are working share a common characteristic: the teams that built them were precise about what each data type was actually being asked to do.

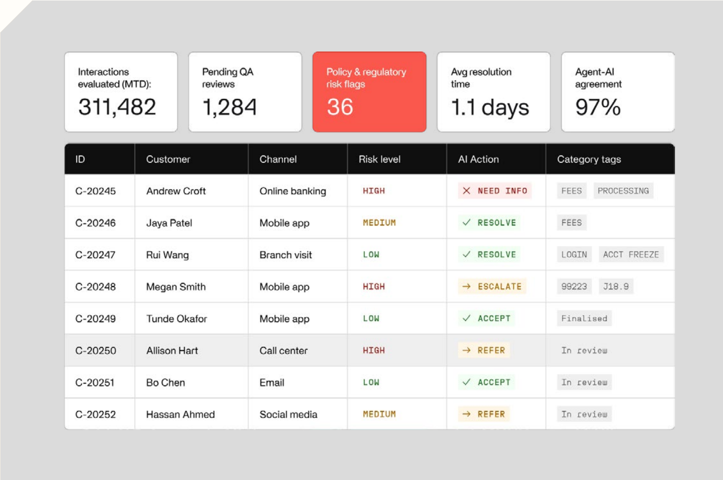
In professional sports, computer vision systems now track motion, interactions, and player behavior across every frame of game footage, at a scale and consistency no analyst team could match. What changes isn't just speed. It's the type of question you can ask: not "what happened in that play" but "what patterns emerge across an entire season, across every player, in every game condition". That kind of insight requires vision, structured data, and reasoning to work together.



Computer vision systems now track motion, interactions, and player behavior across every frame of game footage.

In manufacturing and industrial operations, the value proposition is similar but the stakes are higher. Vision detects what's happening on the floor, assessing equipment states, proximity events, production anomalies. Structured telemetry tracks machine behavior over time. Operator notes and incident logs add context that neither camera nor sensor can provide alone. Fused together, these inputs give operations teams a picture of what's actually happening across a facility, not just what individual systems report in isolation.

In contact centers, the combination of call transcripts, sentiment signals, and backend system data is turning reactive support into something closer to real-time intelligence, surfacing risk, flagging unusual patterns, and giving managers visibility that previously required hours of manual review.



Interactions evaluated (MTD):	Pending QA reviews	Policy & regulatory risk flags	Avg resolution time	Agent-AI agreement
311,482	1,284	36	1.1 days	97%

ID	Customer	Channel	Risk level	AI Action	Category tags
C-20245	Andrew Croft	Online banking	HIGH	✕ NEED INFO	FEES PROCESSING
C-20246	Jaya Patel	Mobile app	MEDIUM	✓ RESOLVE	FEES
C-20247	Rui Wang	Branch visit	LOW	✓ RESOLVE	LOGIN ACCT FREEZE
C-20248	Megan Smith	Mobile app	HIGH	→ ESCALATE	99223 J18.9
C-20249	Tunde Okafor	Mobile app	LOW	✓ ACCEPT	Finalised
C-20250	Allison Hart	Call center	HIGH	→ REFER	In review
C-20251	Bo Chen	Email	LOW	✓ ACCEPT	In review
C-20252	Hassan Ahmed	Social media	MEDIUM	→ REFER	In review

Contact centers can evaluate 100% of interactions against set policies and quality standards, without relying on sampling.

Across these settings, the common thread isn't the technology. It's the decision to stop treating each data source as a separate input and start designing systems where modalities work together toward a specific operational outcome. That design choice is where the value is created and also where most implementations go wrong.

Making multimodal AI work in the enterprise

For enterprises venturing into multimodal applications, success depends on a deep understanding of both the technical and ethical dimensions, rigorous data practice, and thoughtful experimentation. Multimodal forces enterprises to confront not only technical hurdles, but also profound ethical and safety considerations.

The challenge with multimodal systems isn't adding more inputs. It's deciding where each one should matter and in what order. Video is good at answering "did something happen?", but often bad at answering "was this actually important?" without additional context.

We argue that the hard problems are:

- **Identifying which real-world business problems actually need multimodal AI.** Not every workflow benefits from combining image, audio, and text.
- **Building data systems that reflect how work actually happens,** not just what performs well in lab testing. Deciding which inputs should drive decisions at each stage—when should video lead versus text versus sensor data.
- **Testing the system where it's most likely to break** in the messy, ambiguous scenarios your business will actually encounter.

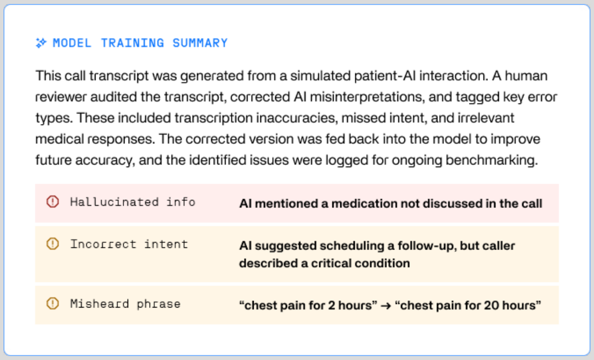
Making multimodal AI work in the enterprise

Multimodal systems rarely fail because the model isn't big enough. They fail because teams try to treat 'vision + audio + sensors' like slightly awkward text, and discover too late that they've built the wrong data, the wrong pipeline, and the wrong evaluation regime for the task they actually care about.

This plays out differently depending on context — in healthcare environments, vision might flag that an interaction occurred, but time-based patterns and human judgment determine whether it matters. In industrial safety monitoring, vision detects proximity to equipment, but additional logic decides whether it's actually unsafe.

Across use cases, the problem isn't model accuracy by itself; it's deciding which input leads the decision, which ones confirm it, and which ones act as safety checks. That's what makes multimodal systems work in production, and also where most enterprises struggle.

Many of the tools that now dominate language modeling were stress-tested in multimodal work first. The new challenge is turning them into robust, end-to-end systems that can handle messy real-world data. Perception-heavy systems like self-driving stacks and industrial robotics were pushing vision, sensor fusion, and control years before large language models emerged, largely independent of NLP. What's changed recently is the coupling: strong language backbones now make it cheap to connect modalities (text↔image, text↔audio, text↔action), so previously separate vision or robotics pipelines can be steered, described, and evaluated through a common linguistic interface.

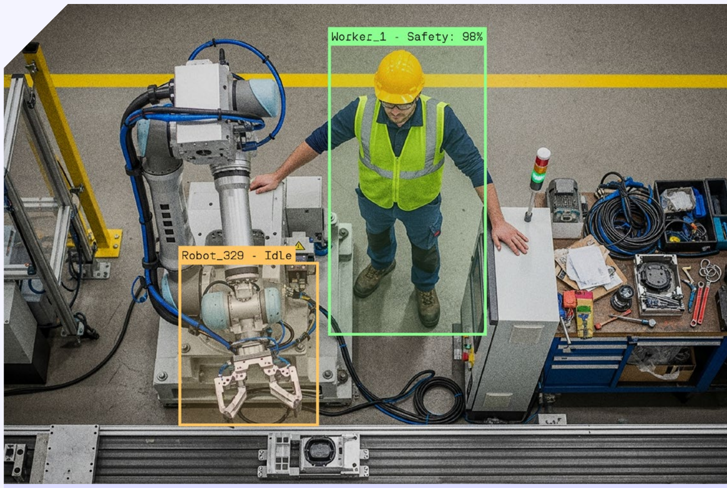


MODEL TRAINING SUMMARY

This call transcript was generated from a simulated patient-AI interaction. A human reviewer audited the transcript, corrected AI misinterpretations, and tagged key error types. These included transcription inaccuracies, missed intent, and irrelevant medical responses. The corrected version was fed back into the model to improve future accuracy, and the identified issues were logged for ongoing benchmarking.

ⓘ Hallucinated info	**AI mentioned a medication not discussed in the call**
ⓘ Incorrect intent	**AI suggested scheduling a follow-up, but caller described a critical condition**
ⓘ Misheard phrase	**"chest pain for 2 hours" → "chest pain for 20 hours"**

In healthcare, real-time insights from medical imagery and live feeds advance diagnostic precision and patient safety.



Proximity data and contextual logic provide intelligence to distinguish safe interactions from industrial hazards.

Making multimodal AI work in the enterprise

No experimentation without ideation

There's a tendency to rush towards testing and building multimodal models without sufficient time spent in ideation: understanding customer needs, surveying available data, and designing their projects accordingly. Early, thorough whiteboarding and inventory of available data assets is crucial. This approach helps ensure that the eventual AI solution addresses genuine business or customer needs. A lot of multimodal projects are optimized for demo performance, rather than real-world adoption. There are so many attributes that are easy to measure, like accuracy, or efficiency, while failing to test for what really matters.

Ask tough questions like:

- Can this model analyze a photo of a job site and cross-reference it with written safety protocols to flag a non-obvious compliance risk?
- Can it look at a damaged component on an assembly line and determine—based on historical maintenance logs—whether to repair it or trigger an immediate part order?

The true success of a model is: 'does it do the job well for the end user', not 'did it get good grades'? We don't tell people our SAT scores 10 years after graduating high school.

Computer vision for infrastructure management analyzes images from drones to detect cracks, corrosion, or infrastructure risks without manual surveys.

Making multimodal AI work in the enterprise

Carefully curate your data

Multimodal AI relies on large, high-quality, well-labeled datasets. Data curation is more difficult with heterogeneous sources.

Enterprises should:

Employ human domain experts for labeling complex data types (e.g. medical images, audio clips) for accuracy.

Be strategy agnostic—willing to clean, augment, or redesign their data capture pipeline based on the specifics of each application.

Choose data partners capable of supporting this flexibility, rather than being tied to a single labeling or augmentation technique.

Data infrastructure and pipeline design should be considered foundational. Building robust systems for ingesting, cleaning, and storing multimodal data from the onset prevents much more expensive problems downstream.

You need to recognize when your dataset is unbalanced: say, overrepresenting one demographic or object class. Careful weighting, targeted sampling, and continuous statistical monitoring are necessary to avoid encoding bias into your models.

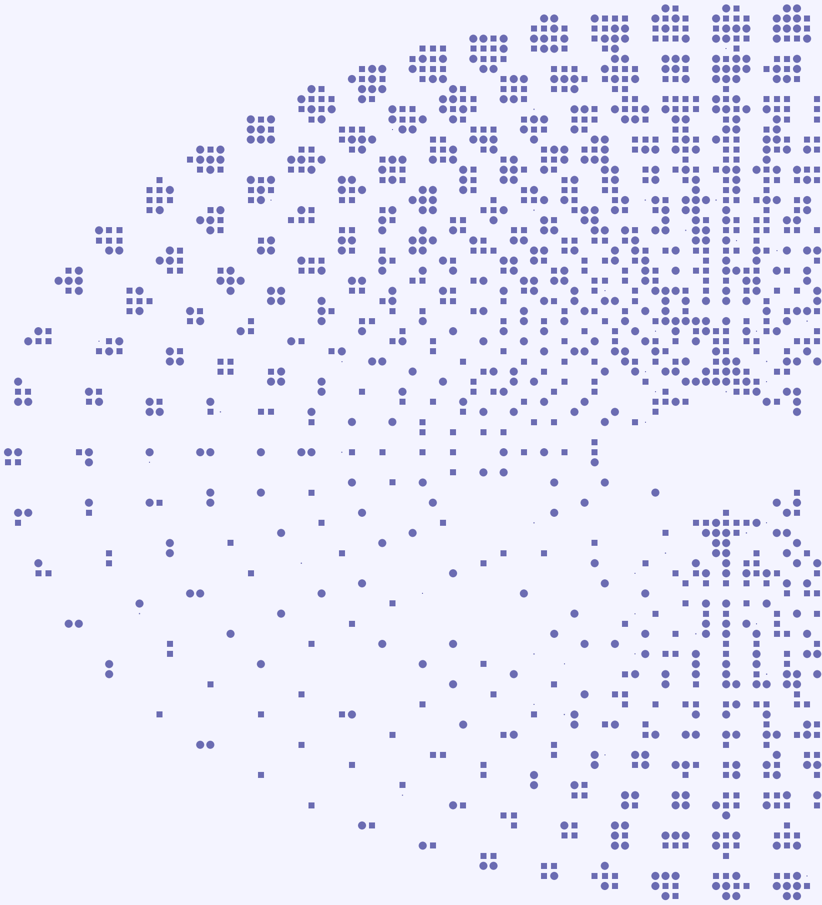
If you’re coming from the text-only world, you need to challenge a lot of your instincts. In text-only land, you might get away with ‘just add more data’ and cheap augmentation (although we don’t suggest this).

In multimodal settings, every extra sample is expensive (video, sensors, expert labels), and most of the value comes from how well modalities are aligned and annotated, not how many raw hours you’ve scraped.

The bottleneck shifts from “do I have enough tokens?” to “do I have the right cross-modal examples, at the right granularity, to support the task I actually care about?” Recent fine-tuning work by Jian et al.[1] shows that even modest fractions of subtly incorrect data (on the order of 10-25%) can significantly degrade domain performance and induce misalignment, with a fairly sharp threshold on how much ‘bad’ data a model can tolerate in practice.

For multimodal regimes, where each labeled example is far more expensive, that tolerance is effectively lower still. You can’t rely on scale to wash the noise out.

It’s also a good idea to balance your training data via sub-sampling, up-sampling, and weighting strategies. This is where “understand your data” comes to the fore: you cannot correct for what you haven’t measured or are unwilling to see. Compare error rates, calibration, and abstention behavior across demographics, geographies, environments, and device types. In multimodal work, that often means conditioning on both who is in the scene and how the scene was captured (camera, mic, channel). If your intended system is to be used in higher-stakes settings like finance, medical, navigation, or security, you should know where it is systematically worse, and decide in advance what happens there: abstain, fall back to a simpler model, or route to a human. Treat those choices as part of the system design.



[1] Ouyang, J., et al., “How Much of Your Data Can Suck? Thresholds for Domain Performance and Emergent Misalignment in LLMs,” 2025.

Making multimodal AI work in the enterprise

Watch for ethical, legal, and safety concerns

Multimodal AI almost always processes personal or biometric data (like voice or images), so during the problem definition and data design stage, enterprises need to ask questions like:

- **What exactly are we allowed to represent?**
  Are we storing raw faces, raw voices, or only derived embeddings? Are we stripping EXIF, GPS, or background audio that could re-identify people or locations? These choices determine what your latent space can encode.
- **Where does consent live in the pipeline?**
  Is consent tied to the raw media, the derived features, or the downstream tasks? Can you handle a withdrawal of consent through your storage, training, and evaluation stack, or would you have to rebuild everything from scratch?
- **Which jurisdictions and regimes apply at each stage?**
  A single sample can cross GDPR, CCPA, HIPAA, or sector-specific rules as it moves from collection → labeling → training → deployment. Your data flows and retention policies need to be drawn with these boundaries in mind, not patched in after an incident.
- **What can be inferred by fusing modalities?**
  Combining innocuous signals (room audio + low-res video + metadata) can reconstruct identity, health status, or location in ways no single modality could. Model and data design should assume cross-modal inference.

Crucially, these considerations must move further upstream. In other words, think about and plan for these issues at the start, not as an afterthought. It will likely incur additional cost, but it will save money downstream on costly redesigns or even legal jeopardy after significant investment.

Technical challenges: feature extraction, model design, and compute

Architectures must be designed to not only process diverse data sources but also create a unified context that links related information across formats: between, say, a photo, a PDF, and a name field.

Compute and memory constraints are a barrier—processing multimodal data is far more resource-intensive than text alone. Capacity planning should account not just for initial experiments, but for continual improvement and deployment, understanding that model efficiency will improve over time.

Making multimodal AI work in the enterprise

Other considerations

Data complexity and readiness

Multimodal projects require high-quality, well-labeled data across multiple formats. Image, audio, and video data are often harder to collect, standardize, annotate, and govern than text, increasing time-to-value.

Talent and organizational maturity

Multimodal AI initiatives often require cross-functional expertise spanning data engineering, machine learning, domain knowledge, and user experience design—capabilities that many organizations are still developing.

Higher technical and infrastructure costs

Multimodal models are typically larger and more resource-intensive, requiring greater compute capacity, specialized hardware, and more sophisticated MLOps pipelines.

Risk, governance, and explainability

As systems become more complex, explaining decisions, managing bias, ensuring compliance, and maintaining trust become more challenging, particularly in regulated industries.

Integration and system design complexity

Successfully combining multiple modalities demands careful system architecture, synchronization of inputs, and robust error handling. Failures in one modality can degrade overall system performance.

Making multimodal AI work in the enterprise

In our experience from working with foundation model labs, robotics and perception teams, and enterprises trying to deploy multimodal systems under real constraints, here's the approach we've found useful in practice:

01

Start from the **task and error costs**

02

Inventory and align your data before you chase a benchmark

03

Treat pipelines as engineered systems with explicit fusion and failure points

04

Use evaluation as a three-legged structure of benchmarks, targeted red teaming, and fine-tuning loops that answer five concrete questions about model choice, safety, readiness, deployment, and monitoring

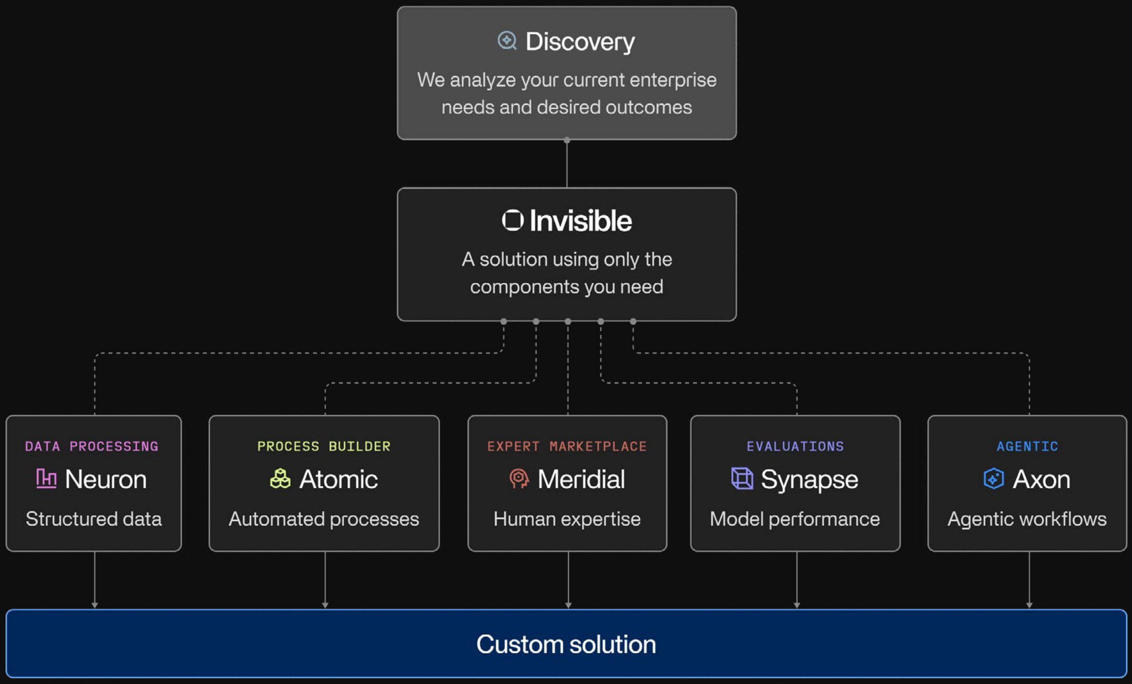
Conclusion

A model for making multimodal AI work

Multimodal AI represents a powerful next step in enterprise AI adoption, unlocking deeper insights, broader automation, and differentiated experiences that text-only AI can't achieve by itself.

However, the increased value comes with increased complexity. Building successful multimodal AI systems is about strategic planning, rigorous evaluation, ethical fortitude, and continual adaptation. **Enterprises that take the time to assess needs, plan for data and ethics, and invest in robust measurement and improvement practices, will be best positioned to both innovate and deliver solutions that are safe, effective, and widely adopted in the real world.**

The organizations that get multimodal right won't be the ones who moved fastest. They'll be the ones who asked the right questions before they built anything. If you're at that stage, we should talk.



Invisible

We’ve trained over 80% of the world’s top AI models.

From people to process to platform, we solve the thing behind the thing.

Call it services-to-software. Or just call it done.

↗ REQUEST A DEMO