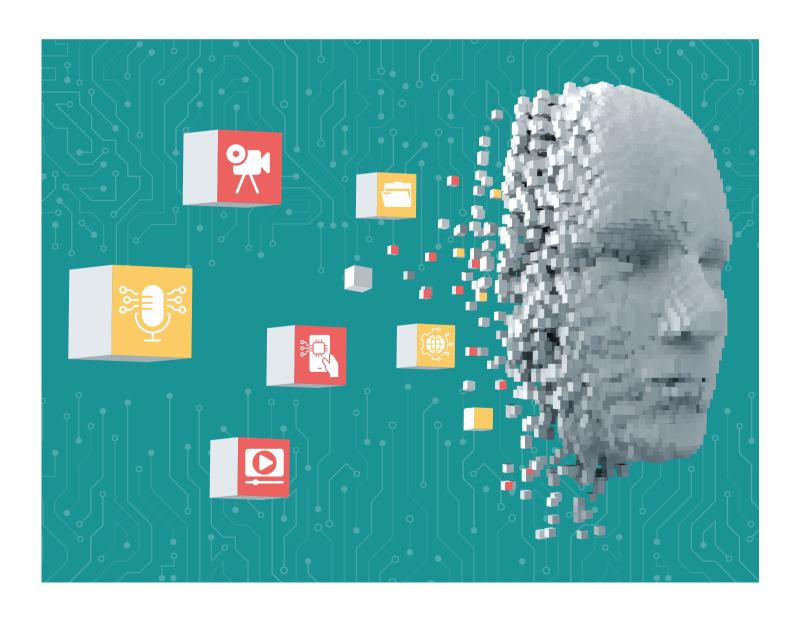


Al models that process multiple types of information at once bring even bigger opportunities, along with more complex challenges, than traditional unimodal Al.

Produced in partnership with



# Multimodal: Al's new frontier



ultimodality is a relatively new term for something extremely old: how people have learned about the world since humanity appeared. Individuals receive information from myriad sources via their senses, including sight, sound, and touch. Human brains combine these different modes of data into a highly nuanced, holistic picture of reality.

"Communication between humans is multimodal," says Jina AI CEO Han Xiao. "They use text, voice, emotions, expressions, and sometimes photos." That's just a few obvious means of sharing information. Given this, he adds, "it is very safe to assume that future communication between human and machine will also be multimodal."

## A technology that sees the world from different angles

We are not there yet. The furthest advances in this direction have occurred in the fledgling field of multimodal Al. The problem is not a lack of vision. While a technology able to translate between modalities would clearly be valuable, Mirella Lapata, a professor at the University of Edinburgh and director of its Laboratory for Integrated Artificial Intelligence, says "it's a lot more complicated" to execute than unimodal Al.

In practice, generative AI tools use different strategies for different types of data when building large data models – the complex neural networks that organize vast amounts of information. For example, those that draw on textual sources segregate individual tokens,

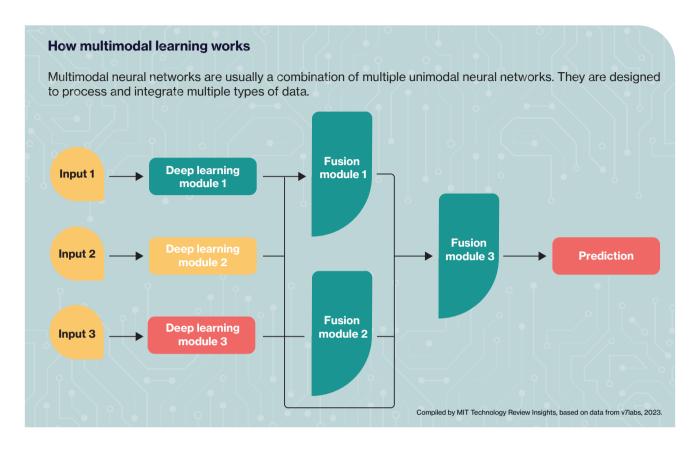
### Key takeaways

- Multimodal AI draws on text, images, audio, and other formats of data to create a nuanced view of the world closer than ever to how a human brain perceives it.
- Development of disruptive multimodal Al-enabled products and services has begun and will grow.
- Multimodal AI faces similar challenges to those of earlier forms of generative AI, but these issues are more complex, multiplying compliance and legal risks from the technology.

usually words. Each token is assigned an "embedding" or "vector": a numerical matrix representing how and where the token is used compared to others. Collectively, the vector creates a mathematical representation of the token's meaning. An image model, on the other hand, might use pixels as its tokens for embedding, and an audio one sound frequencies.

A multimodal AI model typically relies on several unimodal ones. As Henry Ajder, founder of AI consultancy Latent Space, puts it, this involves "almost stringing together" the various contributing models. Doing so involves various techniques to align the elements of each unimodal model, in a process called fusion. For example the word "tree", an image of an oak tree, and audio in the form of rustling leaves might be fused in this way. This allows the model to create a multifaceted description of reality.

"It is very safe to assume that future communication between human and machine will also be multimodal."



Fusion, while simple to describe, is very much a technological work in progress. "There has been a lot of experience with aligning human knowledge on the text level, but on the visual level it is quite different, not to mention other modalities such as audio and 3D mesh," Han explains.

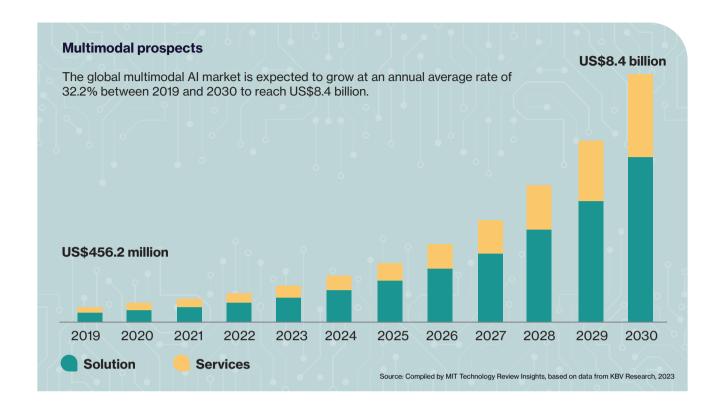
Nevertheless, researchers have made rapid advances. Already, basic multimodal AI models, broadly defined as those able to work across more than one modality, are widely available, explains Lapata. Any number of generative AI services, for example, can create computer code or pictures based on text prompts.

Similarly, notes Ajder, "quite a lot of people have access to some elements of multimodal generative AI, even if they don't necessarily realize it." The latest office productivity suites, for example, have AI-enabled features that let users generate slide shows or other graphical outputs from text inputs. People "just kind of use them," he says. "They often slot seamlessly into existing tool sets and workflows."

Such technology is only a precursor to bigger developments. An ability to draw on distinct modalities promises rapid deployment in new use cases in specific industries, such as media and medicine. More generally across the economy, says Han, the technology will "improve those products that focus on user experience, for example customer service and chatbots." While estimates vary, forecasters agree that the global multimodal AI market will grow rapidly by the end of the decade. One of them, KBV Research, estimates in a 2023 report that the market will reach \$8.4 billion by 2030, marking an average annual growth of 32% in a seven-year period.

"Quite a lot of people have access to some elements of multimodal generative AI, even if they don't necessarily realize it."

Henry Ajder, Founder, Latent Space



# Multimodal innovations already (or very nearly) here

Multimodal Al-enabled tools designed to assist individuals with difficulties in visual perception are already on the market, including applications that help them navigate the web. Han notes that several image-captioning application programming interfaces (APIs), including Jina Al's SceneXplain, can provide webpages with alt-text descriptions of pictures or video scenes in a variety of languages. This enables screen readers for the visually impaired to include descriptions of parts of web pages that were often previously inaccessible.

Meanwhile, general users can now enhance their visual experience with a new class of products: Al-enabled smart glasses with built-in cameras. The most high

profile, the **Ray-Ban Meta Wayfarer**, is available to consumers. It already has numerous competitors in the market or with imminent announced release dates.

Typically, a smart glasses wearer can request audio and text descriptions of images captured by the camera, translate any text, or ask for enriched information, such as where to buy a product seen via the glasses or what dish to make out of ingredients in front of them. The glasses' "ability to take live data ingested visually and translate that into actionable text or recommendations is going to be increasingly exciting," says Ajder. "It is something that people will use." He is not alone: *Computerworld* says "2024 will be 'The year of Al glasses."

"Any problem where you have a lot of images or audio that you want to use to answer questions" will quickly be solved by putting such information into multimodal models.

Mirella Lapata, Professor, University of Edinburgh

### Jina Al's role in the advent of multimodal Al

Multimodal Al looks set to reshape how humans and machines interact. OpenAl's GPT-4, Microsoft's Kosmos-2, and Google Gemini are just some of the new models released to developers in 2023. Each of these can process, analyze, and generate outputs for more than one type of data, including video, images, and audio. The number of such models will only grow.

Like all technological innovations, multimodal Al did not simply spring into life. Nor could any single company develop it. These models rely on the evolution of machine learning and AI in general. Research into fields that contributed specifically to multimodal Al goes back decades. For example, the roots of text embedding and natural language processing are typically traced back to the socalled Bag of Words - a text model discussed as early as 1954.

Meanwhile, research on the use of neural networks in audiovisual speech recognition dates to the 1980s. Studies on multimedia indexing and retrieval can be traced back to the 1990s, and those on emotion recognition to the early 2000s. The idea of **transformers**, the key to modern natural language processing (and the "T" in "GPT"), is already seven years old.

The advance and use of multimodal Al relies on the work of a host of researchers and companies engaged in innovation. One of them is Berlin-based Jina Al, which straddles the open-source and corporate world. CEO Han Xiao says his company, founded in 2020, was "one of the pioneers" in working specifically on multimodality. In particular, it has always concentrated exclusively on several key challenges in the field.

The first of these areas is building the capacity for multimodal search or allowing users, with a single prompt, to search across text, audio, video, 3D mesh, and other modalities. Much of Jina Al's work in this area, in the form of generalized

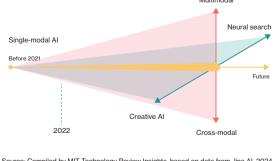
machine-language frameworks, has been made open source. Han says that these tools now allow developers to integrate multimodal searches into their applications "very, very easily."

A related field is embedding. In effect, embedding creates links within a multimodal model between data of different modalities that should be connected. A simple example is the text and accompanying images of a video. Han says Jina Al made "a very significant breakthrough" in this field with the October 2023 release of jina-embeddings-v2. It is the first open-source embedding model with an 8192-token context length, the same as that used by OpenAI for text. The result of this technological achievement is to improve greatly the interoperability of modalities and the capacity of systems to search across them.

The above tools are largely for developers. Jina Al has also taken a lead in the field of prompt engineering, aiming to make life easier for multimodal Al users. The company's commercial **PromptPerfect** service allows nontechnologists to create multimodal prompts combining, for example, audio and video as well as text. This capacity permits users to ask more complex questions.

### Jina Al's spectrum on the future of Al shows the Al industry has entered the multimodal Al era. Multimodal

**Future of Al** 



Source: Compiled by MIT Technology Review Insights, based on data from Jina Al, 2024.

The new technology is also impacting certain industries, notably the entertainment, media, and advertising sectors. These sectors already have substantial multimodal data, such as film and news archives, from which they can derive benefit. Lapata believes that "any problem where you have a lot of images or audio that you want to use to answer questions" will quickly be solved by putting such information into multimodal models.

Another innovation of immediate relevance to these industries comes from Silicon Valley startup **Heygen**. In January 2024, the company introduced a tool that enables an editor to take a video of someone speaking and create another clip of that person saying the same thing, with the same voice, in a different language. The tool also adjusts the video so that it looks natural rather than dubbed. "For entertainment and advertising industries," says Ajder, "this is a complete game changer."

A final innovation indicates how multimodal AI could reshape ways in which humans and personal technology interact. Most multimodal AI tools today recommend actions to human users. In future, however, AI agents that instead carry out instructions — such as "order me a pizza" — are likely to become commonplace. Such products for personal use already exist in the market. For example, Rabbit r1 is a tool that can be prompted in natural language to interface with all the relevant apps on a user's smartphone to do one or more complex tasks.

Tools like Rabbit r1 have potential business applications as well, notes Ajder. Employees will no longer have to juggle different tools to meet a single goal. The cost savings alone of reducing the time used switching between different software and systems is so substantial, he says, that the idea may be the most important innovation of the next few years. It certainly, he says, "holds a lot of promise, if done in a way which is robust and responsible."

The number of use cases and product innovations drawing on multimodal AI will only grow. "There is no doubt that any chief digital transformation officers or chief AI officers worth their salt will be aware of multimodal AI and are going to be thinking very carefully about what it can do for them," says Ajder.

"There is no doubt that any chief digital transformation officers or chief AI officers worth their salt will be aware of multimodal AI and are going to be thinking very carefully about what it can do for them."

Henry Aider, Founder, Latent Space

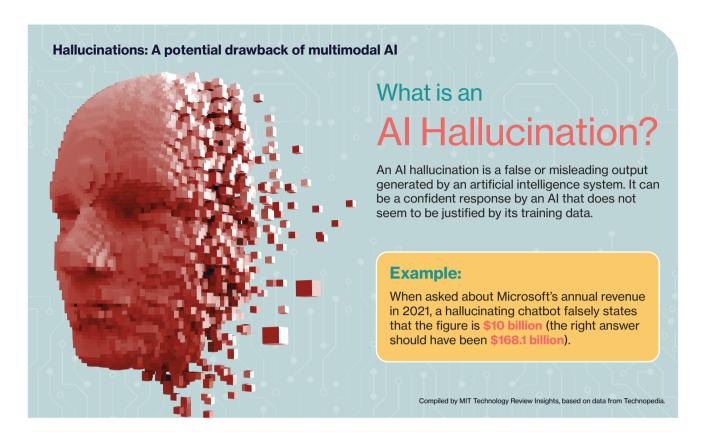
For now, the need for enterprises to learn how to apply the technology overshadows visible changes. "It will take at least half a year for willing businesses to build new revenue streams on multimodal technology," says Han. That leaves limited time before potentially highly disruptive innovations hit markets.

### The complexities of multimodality

As companies deploying unimodal AI have found, implementation brings a range of challenges. "The leap to multimodal adds another level of complexity," says Ajder. In Lapata's view, multimodal models "inherit all the problems that text-only models have, but everything is more complicated."

The complex barriers begin with data. Unimodal generative AI models already require huge volumes of it; combining several in the same model multiplies that need. In some cases, existing technology is the problem. Lapata notes, for example, that current models cannot absorb enough data to process a two-hour video.

Meanwhile, even for organizations with the necessary data, Han explains that "alignment [between different modalities in models] remains a very big problem." Together, he adds, inadequate data and lack of alignment mean that few companies can build foundational multimodal models. The exceptions tend to be those that already hold substantial amounts of largely aligned information, such as the text, image, and video archives of media conglomerates.



An obvious solution to data paucity is to download an open-source or commercial foundation model, which are increasingly available. Independently, however, these have limited utility. Lapata compares them to "perennial B students" who can do many things but fail to excel in specific fields.

Most companies, therefore, will build bespoke models by adding their own data to foundation ones, Han predicts. While this approach may be essential for some business to consider using multimodal AI, Ajder notes that making these additions still entails substantial cost and talent requirements. Moreover, Lapata warns that existing foundation models are not transparent to those who adopt them – even with models "that are called 'open', we have no idea what they have seen and how they're trained."

Deploying multimodal AI can also be far more complicated than using unimodal models. "The more modalities, the more things you may need to put in the prompt," says Ajder, referring to the queries users pose to AI platforms to get the output they want.

Multimodal prompting is more art than science. "There will be a skill element and certainly a learning curve around kind of juggling different modalities within one prompt," Ajder adds. "There will be a language to learn." At the very least, multimodal models will need the capacity to input multimodal prompts, which requires specialized software (see sidebar "Jina Al's role in the advent of multimodal Al").

Finally, multimodal AI has more complex versions of the output, bias, and security issues that beset unimodal AI. One of the best-known drawbacks of generative AI is its propensity to produce hallucinations: outputs that the tools say are accurate but are entirely fictional. The shortcoming is inherent in the design and **limits** of large language models, say researchers at the National University of Singapore.

"Alignment [between different modalities in models] remains a very big problem."

Han Xiao, CEO, Jina Al

# "There is no guarantee that someone crazy won't download available models and train them to do something malicious."

#### Mirella Lapata, Professor, University of Edinburgh

Multimodality exacerbates the challenge of dealing with hallucinations in at least two ways. First, such errors are harder to detect in images and other modes than in simple text, says Lapata. "I don't think we have a good grasp on how to deal with it right now," she adds. Ajder warns the consequences can be greater: "If you're using an LLM to power, let's say, image generation, there may be a higher-risk chance that the hallucinations become much more visceral and much more costly," he explains, referring to potential reputational damage or legal liability.

Bias and unethical use of data constitute another acknowledged challenge with any form of generative Al. Relevant issues begin with the models' need for representative data that does not violate intellectual property. As Lapata noted previously, though, without knowing how foundation models were trained, it is unclear if this is the case.

Inappropriate use is another potential problem, Lapata adds. "There is no guarantee that someone crazy won't download available models and train them to do something malicious," she says. One way to address this, notes Han, are the ethical and copyright filters applied on many models. These, however, can bring their own unintended risks. In February 2024, Google was left red-faced after its Al tool, Gemini, created ahistorical images, which were blamed on badly applied ethical filters.

"Multimodal has clear benefits and advantages, but there's no sugarcoating the complexities it brings."

Henry Aider, Founder, Latent Space

Security is a fourth generative AI issue that becomes more complicated in multimodal models. Even in unimodal AI, it is difficult to rule out that company data may be vulnerable to hacking. With the greater volume of data and complexity required for multimodal AI, the number of possible attack vectors increases. To illustrate this point, Ajder asks, "how are you going to make sure that there aren't vulnerabilities that someone could exploit almost in the gaps" between, for example, the audio and visual models?

Despite the uncertainties, however, many companies will likely have no choice but to face the potential risks of multimodal AI, given the opportunities it could herald. Ajder sums up succinctly what is at stake: "Multimodal has clear benefits and advantages, but there's no sugarcoating the complexities it brings."



"Multimodal: Al's new frontier" is an executive briefing paper by MIT Technology Review Insights. We would like to thank all participants as well as the sponsor, Jina Al. MIT Technology Review Insights has collected and reported on all findings contained in this paper independently, regardless of participation or sponsorship. KweeChuan Yeo was the editor of this report, and Nicola Crepaldi was the publisher.

### About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of *MIT Technology Review*, the world's longest-running technology magazine, backed by the world's foremost technology institution—producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the U.S. and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. And through its growing MIT Technology Review **Global Insights Panel**, Insights has unparalleled access to senior-level executives, innovators, and entrepreneurs worldwide for surveys and in-depth interviews.

### From the sponsor

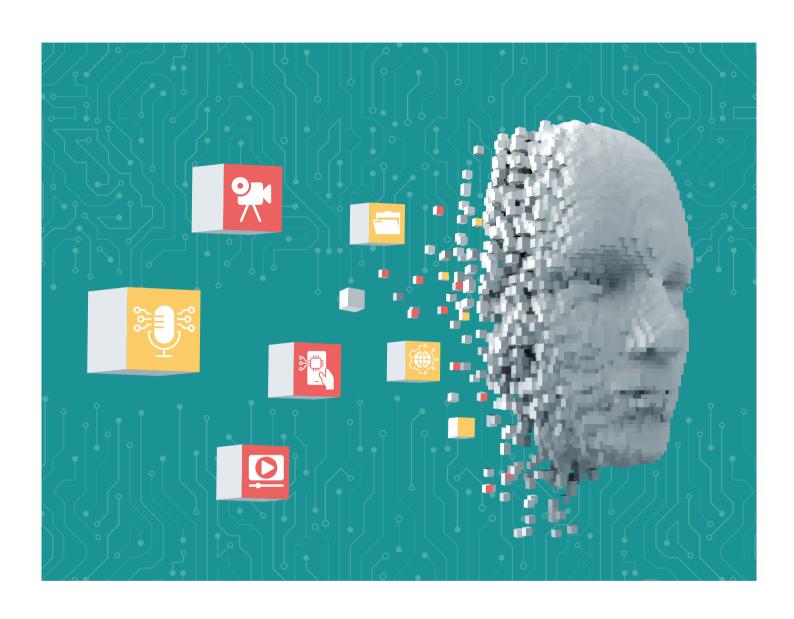
Jina AI pioneers the innovative Search Foundation, a transformative paradigm in neural information retrieval. Our approach seamlessly integrates **embeddings**, **rerankers**, prompt optimizers, and core infrastructure, revolutionizing data utilization. Renowned for our suite of tools, including industry-leading embeddings and rerankers, we excel in escalating search relevancy and RAG accuracy. Additionally, our offerings extend to sophisticated solutions like **PromptPerfect** for prompt engineering and **SceneXplain** for image captions, catering to the diverse needs of enterprises. Trusted by global businesses, our commitment to advancing the AI landscape is underscored by our active involvement in inventive research initiatives.



#### Illustrations

Illustrations assembled by Tim Huxford with elements from Adobe Shutterstock and Adobe Stock...

While every effort has been taken to verify the accuracy of this information, MIT Technology Review Insights cannot accept any responsibility or liability for reliance by any person on this report or any of the information, opinions, or conclusions set out in this report.



## MIT Technology Review Insights

www.technologyreview.com insights@technologyreview.com