Towards deployment-centric multimodal AI beyond vision and language

Xianyuan Liu[1,32], Jiayang Zhang[1,32], Shuo Zhou[2,32], Thijs L. van der Plas[3], Avish Vijayaraghavan[4], Anastasiia Grishina[5], Mengdie Zhuang[6], Daniel Schofield[7], Christopher Tomlinson[8], Yuhan Wang[9], Ruizhe Li[10], Louisa van Zeeland[3], Sina Tabakhi[2], Cyndie Demeocq[11], Xiang Li[12], Arunav Das[13], Orlando Timmerman[14], Thomas Baldwin-McDonald[15], Jinge Wu[8], Peizhen Bai[2], Zahraa Al Sahili[16], Omnia Alwazzan[17], Thao N. Do[18], Mohammod N.I. Suvon[1], Angeline Wang[19], Lucia Cipolina-Kun[20], Luigi A. Moretti[21], Lucas Farndale[22], Nitisha Jain[13], Natalia Efremova[23], Yan Ge[12], Marta Varela[24], Hak-Keung Lam[9], Oya Celiktutan[9], Ben R. Evans[25], Alejandro Coca-Castro[3], Honghan Wu[26], Zahraa S. Abdallah[12], Chen Chen[2], Valentin Danchev[23], Nataliya Tkachenko[27], Lei Lu[28], Tingting Zhu[29], Gregory G. Slabaugh[17], Roger K. Moore[2], William K. Cheung[30], Peter H. Charlton[31], and Haiping Lu[2,32,*]

[1]Centre for Machine Intelligence, University of Sheffield, Sheffield, UK
[2]School of Computer Science, University of Sheffield, Sheffield, UK
[3]The Alan Turing Institute, London, UK
[4]Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK
[5]Department of Applied AI, Simula Research Laboratory, Oslo, Norway
[6]Information School, University of Sheffield, Sheffield, UK
[7]NHS England, Leeds, UK
[8]Institute of Health Informatics, University College London, London, UK
[9]Department of Engineering, King's College London, London, UK
[10]Department of Computing Science, University of Aberdeen, Aberdeen, UK
[11]School of Informatics, University of Edinburgh, Edinburgh, UK
[12]School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK
[13]Department of Informatics, King's College London, London, UK
[14]Department of Earth Sciences, University of Cambridge, Cambridge, UK
[15]Department of Computer Science, University of Manchester, Manchester, UK
[16]Department of Computer Science, Queen Mary University of London, London, UK
[17]Digital Environment Research Institute, Queen Mary University of London, London, UK
[18]Department of Computer Science, University of Bath, Bath, UK
[19]Department of Classics, King's College London, London, UK
[20]School of Electrical, Electronic and Mechanical Engineering, University of Bristol, Bristol, UK
[21]School of Engineering, University of the West of England, Bristol, UK
[22]Cancer Research UK Scotland Institute, Glasgow, UK
[23]School of Business and Management, Queen Mary University of London, London, UK
[24]City St George's University of London, London, UK
[25]British Antarctic Survey, Cambridge, UK
[26]School of Health and Wellbeing, University of Glasgow, Glasgow, UK
[27]Chief Data & AI Office, Lloyds Banking Group, London, UK
[28]School of Life Course & Population Sciences, King's College London, London, UK
[29]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[30]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[31]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[32]These authors contributed equally: Xianyuan Liu, Jiayang Zhang, Shuo Zhou, and Haiping Lu.
[*]Corresponding author: Haiping Lu (h.lu@sheffield.ac.uk)

ABSTRACT

Multimodal artificial intelligence (AI) integrates diverse types of data via machine learning to improve understanding, prediction, and decision-making across disciplines such as healthcare, science, and engineering. However, most multimodal AI advances focus on models for vision and language data, while their deployability remains a key challenge. We advocate a deployment-centric workflow that incorporates deployment constraints early to reduce the likelihood of undeployable solutions, complementing data-centric and model-centric approaches. We also emphasise deeper integration across multiple levels of multimodality through stakeholder engagement and interdisciplinary collaboration to broaden the research scope beyond vision and language. To facilitate this approach, we identify common multimodal-AI-specific challenges shared across disciplines and examine three real-world use cases: pandemic response, self-driving car design, and climate change adaptation, drawing expertise from healthcare, social science, engineering, science, sustainability, and finance. By fostering interdisciplinary dialogue and open research practices, our community can accelerate deployment-centric development for broad societal impact.

Introduction

Data drives discovery in the 21st century[1]. From satellites monitoring climate change to social media capturing human behaviour, the variety and volume of information have never been greater. Each type of data, or *modality*, offers unique insights, but unimodal approaches often fall short of achieving robust or generalisable performance. Autonomous vehicles relying solely on visual data struggle with object detection in low-light or adverse weather conditions[2]. During the COVID-19 pandemic, relying solely on RT-PCR data for diagnosing SARS-CoV-2 infection led to frequent false negatives, hindering timely interventions[3]. Instead, integrating information from multiple modalities can reveal patterns and solutions that unimodal approaches miss[4].

Multimodal artificial intelligence (AI) leverages multimodal data for better understanding complex systems through machine learning[5–8]. Its promise is evident in multidisciplinary applications such as pandemic response. For example, in healthcare, combining medical imaging, genomic sequencing, and epidemiological data can enhance diagnoses, inform treatment strategies, and support disease prevention[9–11]. In science, fusing genetic and protein structure data holds promise for advancing vaccine development[12]. In engineering, integrating textual specifications with spatial data can improve product design and manufacturing[13], which could extend to optimising ventilator production. Across other disciplines, such as sustainability, finance, and social science, multimodal AI offers the potential to provide deeper insights and actionable strategies[14–17].

While multimodal AI receives increasing attention and holds great promise, research has primarily focused on vision-language models[18,19] (Fig. 1 and Supplementary Fig. 1), leaving other modalities–such as tabular and time-series data–and related disciplines underexplored[20] (Fig. 2). Real-world challenges, such as pandemic response, call for AI capable of integrating diverse types of data through interdisciplinary collaboration, bridging the gap between research and application.

As modality integration broadens, multimodal-AI-specific deployment challenges increase, including missing modality[21,22], cross-modal alignment[8,23], and multimodal privacy risk[24,25]. These issues are compounded by broader barriers such as data limitations, integration complexity, and domain-specific constraints[26]. For example, rural healthcare may face limited compute infrastructure; financial services may face regulatory delays; and real-time applications such as autonomous driving demand strict latency control. Multimodal setups often amplify these challenges due to increased system complexity. Simply combining diverse modalities is not enough; real-world success requires proactive alignment with deployment constraints from the outset.

This Perspective outlines a deployment-centric framework for multimodal AI that incorporates deployment constraints early and addresses challenges specific to multimodal integration across disciplines. We first present a general workflow for developing deployment-ready multimodal AI systems. We then examine three data-intensive, cross-disciplinary use cases, pandemic response, self-driving car design, and climate change adaptation, to illustrate common barriers and actionable

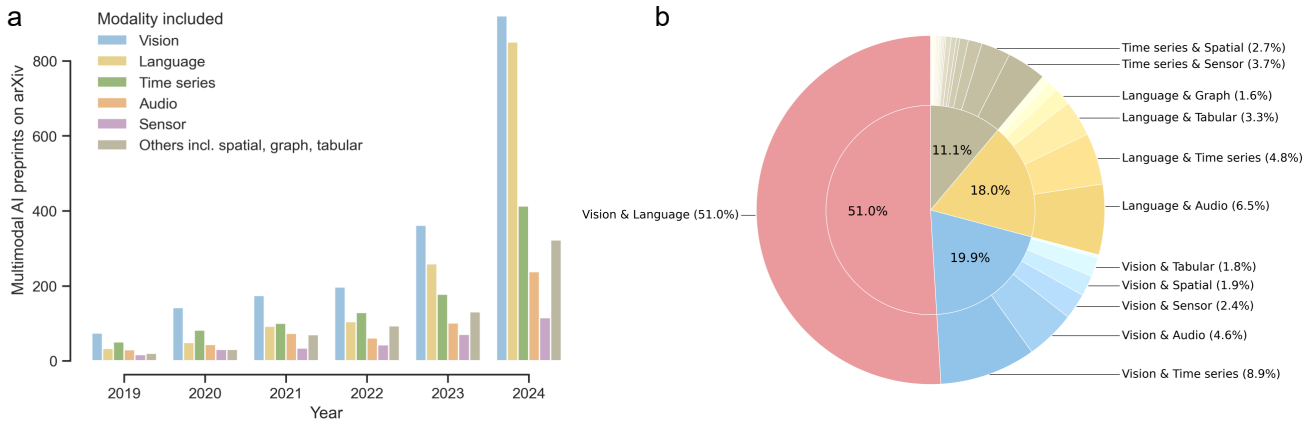


Figure 1. Trends in multimodal AI research (2019–2024) and the dominance of vision and language. a, Yearly growth of multimodal AI preprints on arXiv by modality, showing a steady increase over time and the dominance of vision and language. **b,** Breakdown of modality pairs in multimodal AI preprints on arXiv in 2024, revealing that over half the studies involve vision and language, followed by vision and others (19.9%), language and others (18.0%), and other modality combinations (11.1%). This analysis highlights the most common pairwise modality combinations and shows that those involving vision or language dominate, reaching 88.9%. For clarity and space efficiency, only modality pairs exceeding 1.5% are annotated. See Supplementary Section S2 for methodology and details of the trend analysis. See Supplementary Fig. 1 for a more detailed illustration and analysis of the multimodal AI landscape, including statistics on triple and quadruple modality combinations.

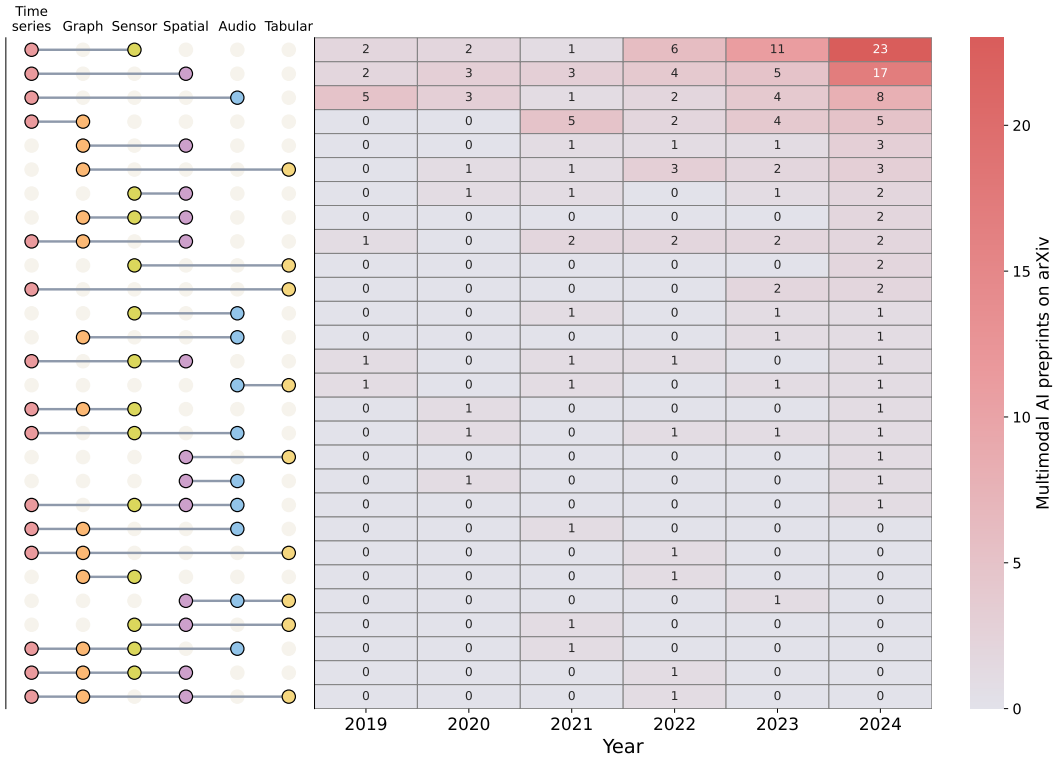


Figure 2. Underexplored modality combinations in multimodal AI (2019–2024). Heatmap of combinations of non-vision, non-language modalities in multimodal AI preprints by year. Each row represents a modality combination, and each column corresponds to a publication year. Darker shades indicate higher counts, with rows ordered by their 2024 totals. The coloured circles on the left identify the modality combinations. For example, in 2023, only two preprints used graph and tabular data (sixth row, fifth column). Time series and sensor data form the most common combination, likely because sensor data are often recorded as time series. Time series and spatial data are the second most common, possibly due to the importance of spatiotemporal modelling. In contrast, combinations involving graph, audio, and tabular data remain sparsely studied. These gaps highlight untapped potential for multimodal AI beyond vision and language. See Supplementary Section S2 for details on data processing and modality extraction.

strategies. By exploring underrepresented modalities and fostering open, interdisciplinary collaboration, we aim to broaden the impact of multimodal AI beyond vision and language.

Deployment-centric multimodal AI system development

Traditionally, multimodal AI research has been *model-centric*[5–7], focusing on developing new models to outperform existing ones on standardised benchmarks and datasets. The rise of generative AI[28] models such as ChatGPT[29,30] resulted in a shift towards *data-centric* approaches[27,31,32], emphasising data resources and quality for better performance. However, the gap between high expectations and limited real-world impacts[26,33] indicates a pressing need for *deployment-centric* approaches that prioritise real-world applicability, user needs, and ethical considerations, ensuring that AI innovations are both novel and practical, with a positive impact. We advocate a deployment-centric workflow for advancing multimodal AI from research to scalable solutions, built on ideas from machine learning operations lifecycle guidelines[34] and technology readiness levels for machine learning systems[35].

Figure 3a illustrates our deployment-centric workflow in three stages: planning, development, and deployment. Planning focuses on defining the problem, determining the suitability of multimodal AI over unimodal AI, understanding real-world constraints, and formulating AI tasks grounded in realistic assumptions. Development builds the multimodal AI system to learn predictive models from multimodal data, and deployment assesses real-world performance, feeding back insights to improve earlier steps. Guidance from stakeholder engagement and interdisciplinary collaboration (the top right of Fig. 3a) informs the entire workflow, ensuring that the perspectives of domain experts, end-users, and decision-makers are integrated for a robust and socially aligned system[36,37].

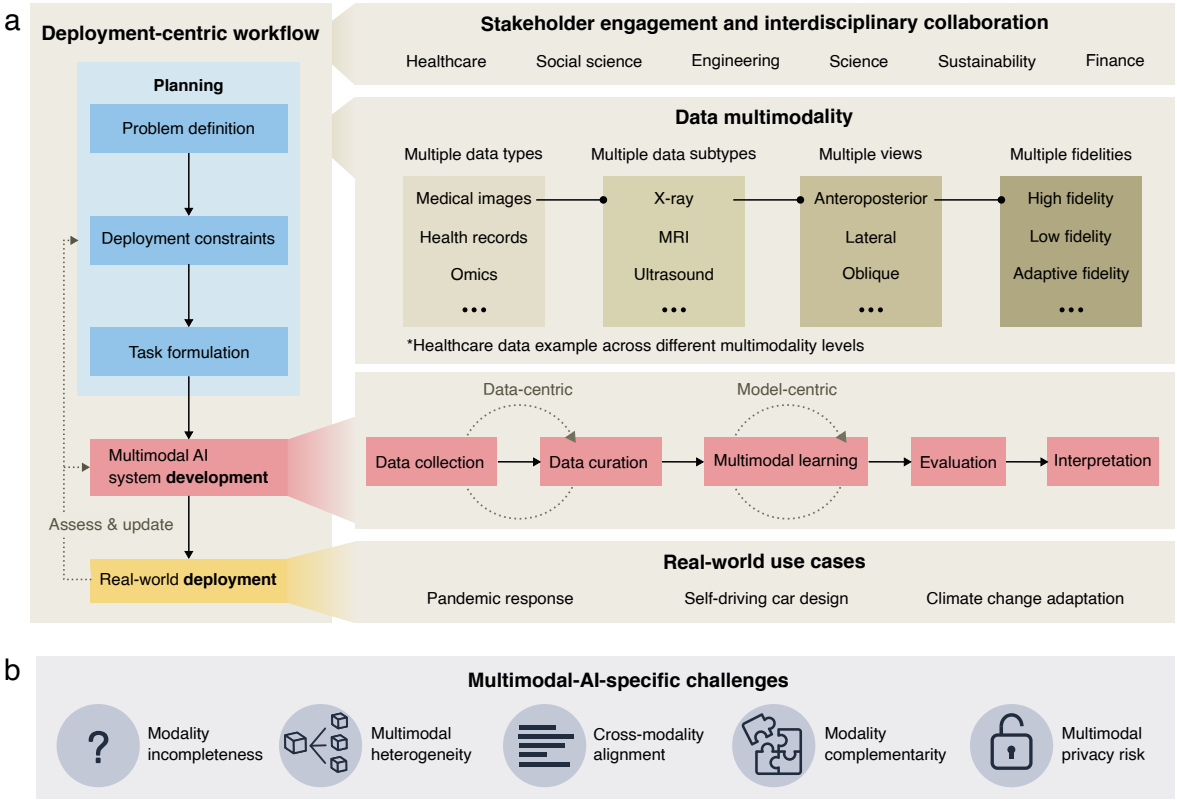


Figure 3. Deployment-centric multimodal AI: workflow and challenges. a, Deployment-centric multimodal AI workflow designed to meet real-world needs through a structured three-stage process covering planning, development, and deployment, with iterative assessments and updates. In particular, the planning stage considers deployment constraints early to ensure alignment with real-world settings and practical needs. This workflow incorporates stakeholder engagement and interdisciplinary collaboration at all stages to ensure that real-world needs and discipline-specific knowledge inform and enhance AI system development. Moreover, we consider multiple levels of data multimodality, from data types and subtypes to views and fidelities, as illustrated with a healthcare example. This broader definition of multimodality offers new perspectives and rich options for leveraging the benefits of multimodality. The system development stage has five steps similar to a standard machine learning pipeline, where the data-centric and model-centric approaches[27] are indicated in the figure to highlight their differences from the deployment-centric approach. **b,** Five multimodal-AI-specific challenges shared across multiple disciplines and real-world applications: modality incompleteness, where one or more modalities are missing at training or deployment; multimodal heterogeneity, reflecting incompatible formats and data structures; cross-modality alignment, which requires synchronising data in time or meaning; modality complementarity, where the goal is to maximise synergy without introducing redundancy; and multimodal privacy risk, where data fusion increases the chance of sensitive re-identification.

Planning for multimodal AI systems

Planning (the three blocks on the top left of Fig. 3a) begins with **problem definition**, which involves clearly articulating the problem's scope and objectives, and preliminarily assessing whether incorporating multimodal data can offer meaningful advantages, such as improved predictive performance or deeper insights, over unimodal approaches[38]. Beyond vision (image, video) and language (text), key modalities include audio, numeric time series (e.g. financial data sequences), sensor signals (e.g. wearable physiological measurements), and spatial (geolocation), tabular (structured data such as clinical records), and graph-based (relationships or networks) data. Furthermore, multimodality exists at multiple levels: multiple data types, subtypes, views, and fidelities[39,40] (the middle right of Fig. 3a and Fig. 5). For example, in healthcare, multimodal data can range from different types, e.g. medical images, electronic health records (EHRs), and omics data[41,42], to subtypes within a data type, e.g. X-ray, magnetic resonance imaging (MRI), and ultrasound within the imaging modality, multiple views of the same data (sub)type, e.g. anteroposterior, lateral, and oblique views of X-ray, and multiple fidelities, e.g. high, low, and adaptive fidelities of anteroposterior X-ray. Under this broad definition, the potential number of modalities can expand substantially as more levels are considered. Understanding these modalities early helps select those suited to the problem's complexity and

capable of offering insights beyond unimodal data.

After establishing the potential benefits of multimodality, we move on to **deployment constraints**, which involves examining the AI system's application context, including user needs, data availability, regulatory compliance[43], ethical considerations, societal impact, and economic trade-offs between high-cost and low-cost data modalities. This step ensures that the chosen modalities and the resulting multimodal AI system are not only technically feasible but also viable within the intended deployment environments. Understanding these constraints early informs task formulation, helping align solutions with real-world requirements and making it more effective and responsible.

The final step, **task formulation**, translates the defined problem and constraints into specific AI tasks, specifying the inputs, outputs, and evaluation metrics required to meet the objectives set during problem definition. This provides a clear development roadmap, ensuring multimodal AI systems are well-positioned to meet both technical and practical needs. Selecting modalities requires balancing utility, acquisition cost, complexity, and deployment feasibility. Fewer well-curated modalities may outperform broader but less practical combinations.

Multimodal AI system development

Developing multimodal AI systems (the lower left of Fig. 3a) parallels standard machine learning system development but introduces added complexities due to the integration of diverse data modalities. We have identified five multimodal-AI-specific challenges (Fig. 3b). **Modality incompleteness** occurs when one or more modalities are missing during training or deployment, necessitating the development of highly flexible or generative models. **Multimodal heterogeneity** deals with varying formats, scales, and data structures, necessitating careful integration strategies to ensure interoperability. **Cross-modality alignment** ensures that the timing or meaning of data from different sources is correctly aligned for coherency and consistency, whether temporal (e.g. matching timestamps) or semantic (e.g. identifying data representing the same concept across modalities). **Modality complementarity** ensures that different modalities contribute complementary information to enhance performance, as more modalities may not improve results, e.g. if they add noise or redundancy. Finally, **multimodal privacy risk**[24,25] arises when independently anonymised datasets become identifiable again (re-identification) through their fusion, thereby revealing sensitive information and necessitating robust privacy-preserving techniques.

The development process consists of five key steps: data collection, data curation, multimodal learning, evaluation, and interpretation (the lower right of Fig. 3a). Each step plays a key role in addressing the multimodal-AI-specific challenges described above.

Data collection and **curation** build on the modality choices identified during planning, gathering relevant data from multiple sources and preparing them for AI model training. Diverse and reliable data sources ensure a high-quality, comprehensive representation of the problem space. Synthetic data and weak supervision provide valuable alternatives when labelled data is scarce[44], such as in rare diseases or time-sensitive crises, where expert annotation may be infeasible. Once collected, the data undergoes standard curation, such as wrangling, cleaning, annotation, handling missing values, and quality assurance[26], tailored to multimodal AI. For instance, in healthcare, linking multimodal data from EHRs, imaging, and biosignals poses substantial challenges due to privacy concerns[24] and the need for standardisation. The integration and cross-referencing of these modalities can amplify the risk of re-identification, underscoring the importance of addressing multimodal privacy risk through robust privacy-preserving techniques[45–47]. Moreover, selected modalities should be interoperable and complementary, with their alignment and integration resulting in a cohesive, high-quality dataset suited to technical and deployment requirements.

Multimodal learning integrates diverse data modalities to leverage their unique strengths for improved predictive performance[5]. Traditional fusion strategies, early fusion (combining raw features), intermediate fusion (merging processed features), and late fusion (integrating outputs from independently processed modalities), offer trade-offs between model complexity and when modalities interact. Recent advances, hybrid fusion[4] and knowledge distillation[48], provide greater flexibility, enabling the combination of multiple strategies or the transfer of knowledge from complex to simpler models. Techniques such as co-attention mechanisms[49] can further enhance integration and performance by dynamically adapting to cross-modal interactions. The choice of strategy depends on the specific problem and data characteristics, balancing integration benefits with the added complexity of managing multiple modalities.

Evaluation and **interpretation** support the reliability, effectiveness, and fairness of multimodal AI systems, especially in complex, high-stakes environments. Beyond standard performance metrics such as accuracy, evaluation should assess the contribution of individual modalities, the alignment and synergy between them, and the system's robustness to noise and variability. For example, evaluating safety in multimodal models[50] is key for understanding how biases across modalities may interact or amplify one another[51], potentially undermining system reliability. Interpretation[52,53] ensures that model decisions are understandable and transparent. Techniques such as heat maps, *t*-SNE, and decision trees help visualise and explain how different modalities interact and contribute to outcomes, fostering user trust and supporting error analysis and model refinement. Uncertainty estimation[54], via ensembling, calibration, or Bayesian approaches, supports more reliable decisions in high-stakes applications.

Real-world deployment of multimodal AI systems

Deploying multimodal AI systems (the bottom of Fig. 3a) requires infrastructure for hosting, scaling, and ongoing management, with continuous monitoring to ensure the system remains effective and relevant[26]. Infrastructural and regulatory conditions, including compute availability, connectivity, and legal safeguards, can determine what is practically feasible and where deployment may succeed or fail. Unlike unimodal AI, these systems face unique challenges, such as integrating diverse data modalities in real-time environments and maintaining performance despite modality-specific issues, such as sensor failures or data stream interruptions. Robust monitoring mechanisms enable the system to detect and compensate for failures in one modality by leveraging information from others. Additionally, real-time multimodal data fusion, where multimodal data may arrive asynchronously, requires careful infrastructure design and adaptive algorithms.

The following sections examine three cross-disciplinary use cases that demonstrate the value and applicability of the deployment-centric multimodal AI workflow: pandemic response, self-driving car design, and climate change adaptation.

Use case 1: Pandemic response

Pandemic response (Fig. 4a) presents a complex challenge requiring coordinated efforts across healthcare and other disciplines[55]. Defining the problem involves assessing early how multimodal AI can offer advantages over unimodal approaches by integrating diverse data sources for deeper insights and more accurate predictions. For example, in social science, combining textual content from surveys or social networks with tabular demographic data can help understand mental health impacts and detect changes in online behaviour[56,57]. In sustainability, analysing time-series environmental data, satellite imagery, and sustainability reports can reveal the environmental effects of pandemics[58,59]. In finance, integrating economic indicators, transaction patterns, and market response data with health data can predict socio-economic impacts and inform public health strategies[60,61]. A well-defined problem ensures clear constraints and tasks can be established.

The deployment of multimodal AI in pandemic response faces privacy, technical, and economic constraints. Privacy concerns arise from integrating diverse data sources such as EHRs, social media, and genomic information, which increases the risk of re-identification and amplifies the potential impact of data breaches. Maintaining public trust requires robust privacy-preserving techniques[45–47] and compliance with regulatory frameworks. Technical challenges include real-time monitoring, where multimodal AI systems must integrate and process data streams rapidly for timely interventions. The heterogeneity of health and social data complicates alignment and standardisation, particularly in international or multi-centre collaborations. Achieving minimum predictive accuracy thresholds for outbreak detection and public health decision-making helps build trust in AI-driven systems. Economic constraints, including computational costs and budget limitations, restrict accessibility and scalability. High-cost modalities (e.g. MRI) may be impractical in low-resource settings, so multimodal AI systems should flexibly support lower-cost alternatives (e.g. wearables) to promote equitable deployment.

Task formulation translates the defined problem into specific AI objectives, detailing inputs, outputs, and evaluation metrics for pandemic-related challenges. A primary task is predicting disease outbreaks by integrating multimodal data such as epidemiological records, EHRs, and environmental factors to inform early interventions and resource allocation. Another task is monitoring public adherence to health guidelines using social media data, geospatial information, and biosignals, providing actionable insights to refine policies. Additionally, multimodal AI can support real-time decision-making in healthcare by combining patient data streams, including wearable biosignals, medical images, and clinical notes, for managing patient surges effectively[62,63]. Task formulation ensures that multimodal AI systems align with technical and practical needs, strengthening pandemic response through actionable insights and tailored interventions.

Developing multimodal AI systems for pandemic response follows the standard workflow of the five key steps outlined above. Data collection assembles diverse sources, such as EHRs, genomic data, and biosignals, to create comprehensive datasets for outbreak prediction and patient management. During curation, standardisation and alignment address challenges such as integrating time-sensitive data streams from biosignals and geospatial sources. Multimodal learning can employ self-supervised[64] and cross-domain[65] techniques to capture cross-modal relationships effectively and integrate domain-specific knowledge. Evaluation requires multi-metric, multi-centre validation[66,67], ensuring robustness for tasks such as outbreak prediction and public adherence monitoring. Interpretability and explainability enable healthcare professionals and policymakers to trust and act on multimodal AI insights in high-stakes scenarios[68]. Tailoring these development steps to pandemic-specific needs ensures adaptability, reliability, and ethical compliance in rapidly changing contexts.

Deploying multimodal AI systems in pandemic response translates research into actionable solutions for high-stakes environments. For instance, patient surge management can integrate biosignals, medical images, and clinical notes to optimise intensive care unit bed allocation and staff deployment. Public health decision-making can benefit from multimodal analyses of epidemiological data, EHRs, and social media trends to guide interventions such as lockdowns or resource distribution. Real-time deployment requires robust infrastructure to process asynchronous data streams and address modality-specific interruptions, such as gaps in biosignal or social media data. Adaptive algorithms and reliable systems ensure timely and accurate outbreak prediction and monitoring. Assessment through multi-centre trials and multi-disciplinary benchmarks[69]

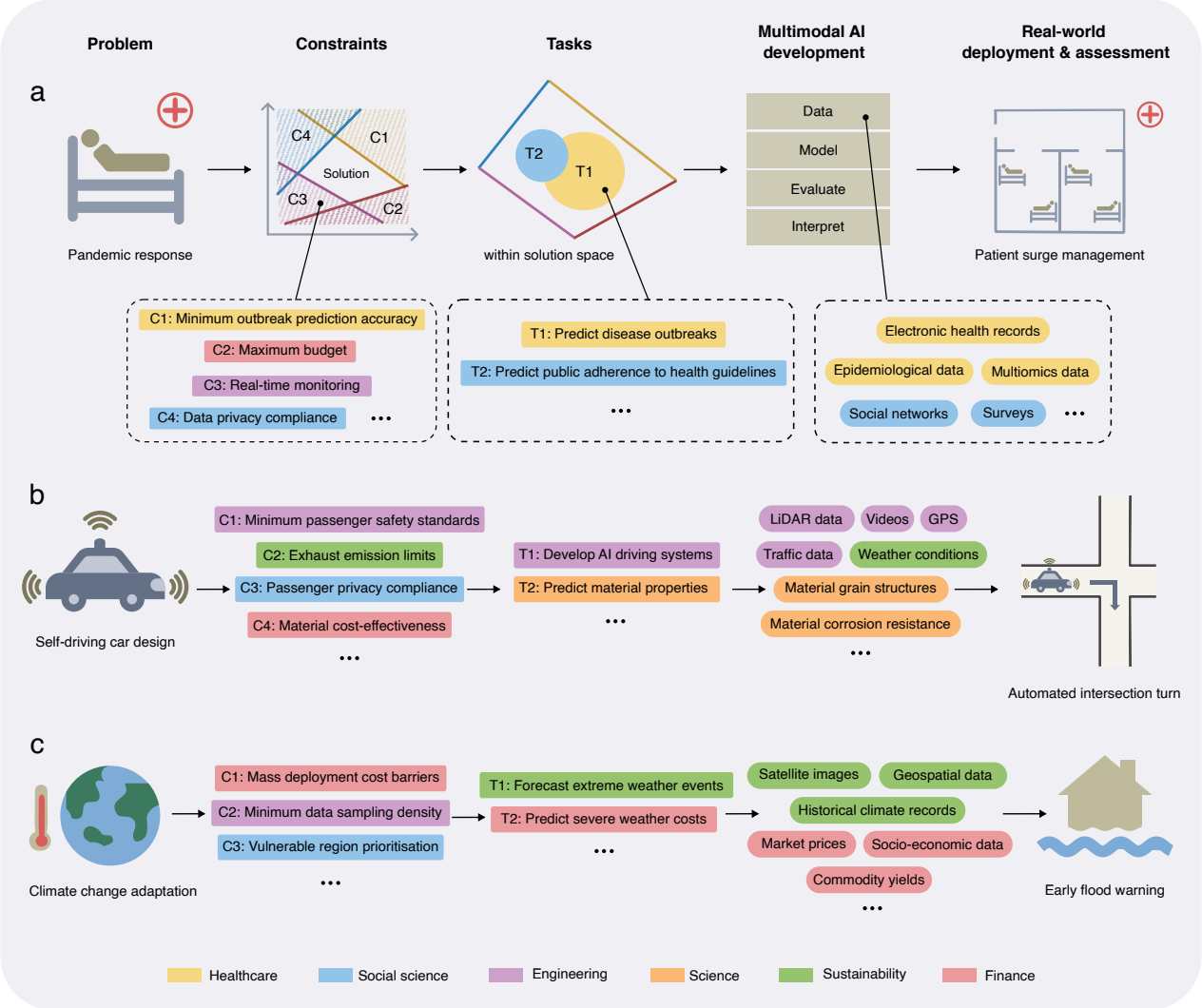


Figure 4. Deployment-centric multimodal AI workflow for three use cases. These examples demonstrate the versatility of the proposed workflow, showing how deployment-centric considerations can be applied across various disciplines to solve complex challenges. **a,** This example illustrates how the proposed deployment-centric multimodal AI workflow can better address real-world challenges in pandemic response, drawing on stakeholder engagement and interdisciplinary collaboration, as indicated by the different colours across stages. The process begins with defining the problem of interest properly and carefully considering the space of various deployment constraints in order to formulate specific tasks. Two tasks are shown in the figure: one with a focus on healthcare and the other on social science. Next, we develop multimodal AI systems using diverse data sources. Finally, we proceed to real-world deployment and assessment, exemplified by the application of patient surge management. **b,** In self-driving car design, the workflow incorporates deployment constraints from social science, engineering, sustainability, and finance disciplines. Two key tasks focus on engineering and science considerations, respectively. Multimodal AI development uses multiple data sources, including remote sensing, weather, and materials data, aiming for real-world deployment exemplified by automated intersection turning. **c,** In climate change adaptation, the workflow considers constraints such as mass deployment cost barriers, minimum data sampling density, and prioritisation of vulnerable regions. Task formulation and multimodal AI development focus on environmental sustainability and financial considerations, reflecting how task and model development are shaped by discipline-specific needs. Early flood warning serves as an example of real-world deployment and assessment.

can validate the added value of multimodal AI over unimodal systems. Measuring the accuracy of outbreak prediction or the effectiveness of intervention informed by multimodal AI can ensure trust and accountability. Further addressing scalability and privacy compliance can support sustainable and ethical deployment.

Use case 2: Self-driving car design

Self-driving car design (Fig. 4b) exemplifies a transformative application of multimodal AI, addressing challenges in safety, efficiency, and sustainability[70,71]. Defining the problem involves determining how multimodal AI can substantially outperform unimodal approaches for navigating complex environments safely and efficiently. For instance, fusing sensors such as LiDAR (light detection and ranging), radar and cameras can enhance perception and decision-making in dynamic settings, especially where a single sensor may fail[72]. Beyond perception, multimodal AI can support physical simulations, such as combining text, images, and videos, to model vehicle dynamics and road interactions[73]. In materials design, integrating structural and chemical data aids in developing lightweight, sustainable materials for vehicle construction[74]. Clearly defining these challenges ensures multimodal AI solutions are effectively tailored to the demands of autonomous driving.

The deployment of multimodal AI in self-driving car design faces safety, privacy, technical, and economic constraints that must be addressed for safe and scalable adoption. Privacy concerns arise from integrating multimodal data, such as LiDAR, cameras and passenger-related information, which increases the potential for data misuse or unauthorised tracking[75]. Mitigating these risks requires robust data anonymisation and adherence to privacy regulations[76]. Technical challenges include real-time data processing in dynamic environments, such as urban intersections or adverse weather. Modality-specific failures, such as LiDAR disruptions in heavy rain, require fallback mechanisms, such as radar to maintain reliability. Self-driving systems must achieve high predictive accuracy for collision avoidance and route planning, and operate reliably across regions with varying regulations and infrastructure. Economic constraints involve balancing the cost of high-resolution sensors with scalability needs. Meeting sustainability goals requires energy-efficient operation, with trade-offs between sensor performance and power consumption. Addressing these constraints holistically enables robust and accessible deployment of self-driving technologies.

In self-driving car design, task formulation maps problems and constraints to concrete AI objectives for perception, navigation, and safety. A key task is developing AI driving systems that integrate multimodal data streams, including LiDAR, radar, GPS, and cameras, for real-time navigation. These systems must dynamically process inputs to handle traffic, obstacles, and weather variability. For example, integrating road and weather data can enhance route planning and safety in adverse conditions[77]. Another task is predicting material properties and performance to support vehicle design. By fusing structural, experimental, and supply chain data, multimodal AI can optimise materials for lightweight, durable, and cost-effective vehicles, aligning with sustainability goals[78]. Additional tasks include passenger-focused objectives, such as integrating biosignals and cabin sensors to enhance comfort and safety. Ensuring privacy compliance remains integral across all tasks.

Developing multimodal AI systems for self-driving cars follows the five-step workflow above, adapted to the demands of autonomous systems. Data collection involves assembling sensor inputs (e.g. LiDAR, radar, cameras) and materials data to build comprehensive datasets. Data curation addresses challenges such as synchronising asynchronous sensor inputs and standardising materials information. Learning can employ fusion strategies to enhance system reliability, while physics-informed neural networks[79,80] can integrate scientific laws to improve realistic simulations of driving scenarios and materials development. Evaluation ensures robustness through multi-metric validation across diverse scenarios, including urban environments and extreme weather conditions. Finally, interpretation in autonomous systems supports real-time explainability and transparency, enabling stakeholders to trust and audit system decisions. By aligning these steps with the unique requirements of autonomous systems, multimodal AI can ensure adaptability, safety, and performance.

Real-world deployment of multimodal AI systems for self-driving cars requires live sensor integration, infrastructure readiness, and operational resilience across diverse environments. Robust in-vehicle computing and inter-vehicle communication enable reliable and coordinated performance at scale. Assessment validates multimodal AI's added value through simulations and multi-centre trials. Testing performance under extreme conditions, such as adverse weather or high-traffic intersections, ensures reliability. Addressing scalability and privacy compliance further supports ethical, sustainable deployment, advancing societal trust in self-driving technologies[81].

Use case 3: Climate change adaptation

Climate change adaptation (Fig. 4c) involves forecasting extreme weather, assessing climate risks, and managing resources to minimise environmental and socio-economic impacts[82–85]. Defining the problem involves assessing whether multimodal AI can surpass unimodal approaches by integrating diverse data sources for deeper insights and actionable strategies. For example, multimodal AI can combine satellite imagery, historical climate records, and geospatial data to model weather patterns and predict extreme events, supporting proactive disaster management[86]. Integrating socio-economic data with environmental observations can enable more comprehensive climate risk assessments[87].

Deploying multimodal AI for climate change adaptation must navigate privacy, technical, and economic constraints. Privacy concerns arise from integrating sensitive data, such as socio-economic records and proprietary satellite imagery. Ensuring compliance with global and regional privacy regulations and the ethical use of data supports responsible deployment. Technical challenges include the heterogeneity and sparsity of environmental and socio-economic data. Economic constraints, such as

the high cost of advanced sensors, necessitate balancing low-cost and low-fidelity data sources with high-cost, high-fidelity alternatives to maintain accessibility and scalability. By addressing these constraints, multimodal AI can deliver equitable and sustainable climate solutions in diverse global contexts.

For climate change adaptation, task formulation aligns complex environmental challenges with actionable AI tasks. A primary task is forecasting extreme weather events by integrating multimodal data such as satellite imagery, historical climate records, and real-time observations. These forecasts can enable timely interventions to mitigate human and economic losses[88], including early warning systems for floods that integrate geospatial, hydrological, and social data streams[89]. Another task is predicting the socio-economic impacts of severe weather by combining spatial, market, and socio-economic data, guiding resource allocation and policy decisions[90]. These tasks align multimodal AI with both technical requirements and societal priorities, advancing climate resilience and equitable adaptation strategies.

Developing multimodal AI systems for climate change adaptation follows a similar five-step workflow. Data collection gathers remote sensing data, near-real-time sensor network outputs, historical climate records, and socio-economic metrics, creating comprehensive datasets for tasks such as extreme weather forecasting and early warning systems. Data curation addresses challenges such as temporal misalignment and data sparsity by standardising inputs and enriching sparse data through interpolation or integration of auxiliary sources. Multimodal learning can benefit from advanced models such as Aurora[91], which fuse satellite imagery and meteorological data to improve atmospheric predictions. Evaluation uses multi-metric benchmarks to test system performance in diverse scenarios. For example, WeatherBench 2 provides a standard platform for assessing atmospheric prediction models[92]. Interpretation ensures outputs are transparent and actionable for stakeholders, supporting evidence-based decision-making. By aligning these steps with climate-specific challenges, multimodal AI can deliver reliable, scalable, and adaptable solutions.

Real-world deployment of multimodal AI for climate change adaptation translates research insights into scalable, operational systems. Early warning platforms and impact forecasting models must synchronise diverse data streams, deliver real-time outputs, and integrate with decision-making infrastructures across sectors and regions[93,94]. Real-time monitoring requires robust infrastructure to synchronise and process diverse modalities. Deployment challenges include latency, data coverage gaps, and variable infrastructure capacity across geographies. Assessment involves stress testing and multi-centre validation to ensure reliability across environmental and socio-economic scenarios. Ethical considerations, such as prioritising vulnerable communities and ensuring equitable access to data and AI tools, help enable sustainable deployment. By overcoming deployment challenges, multimodal AI systems can empower policymakers, industries, and communities to respond proactively to climate challenges, advancing global sustainability efforts.

Outlook

Most existing research on multimodal AI has focused on model-centric development. While attention to data-centric development is increasing, unlocking the full potential of multimodal AI and addressing real-world challenges will ultimately require a shift towards deployment-centric development. This shift will require strategic advancements in data, model, and deployment methods, as well as concerted efforts to address multimodal-AI-specific challenges, namely modality incompleteness, multimodal heterogeneity, cross-modality alignment, modality complementarity, and multimodal privacy risk, through stakeholder engagement, interdisciplinary collaborations, and community-building. Box 1 presents strategic recommendations to advance multimodal AI across disciplines under a deployment-centric perspective. These recommendations draw on insights from the three use cases and the broader cross-disciplinary analysis in the Supplementary Information on multimodal AI beyond these three use cases, providing actionable guidance for future developments.

Deployment-centric development brings challenges around safety, reliability, interpretability, scalability, and ethics, particularly as multimodal AI expands in high-stakes applications such as healthcare and sustainability. Addressing these challenges requires implementing robust human-in-the-loop systems, developing clear standards for safety and transparency, and prioritising scalable, resource-efficient infrastructure to support increasing demands for data integration, processing, and storage.

Robust data-centric development underpins deployment-centric progress by ensuring data availability, diversity, and quality. High-quality multimodal data is often limited, and existing datasets often lack the diversity needed to ensure fairness in AI systems. Clear benchmarks[95,96] and globally accessible datasets (Fig. 5) are needed to enhance trust, consistency, and adaptability across disciplines, facilitating reproducible research. Promoting secure data-sharing frameworks[97], open initiatives such as the European Life Science Infrastructure for Biological Information[98], and globally accessible multimodal data platforms will further strengthen AI's capacity to address diverse global challenges. Knowledge graphs[99] also offer promise for organising varied data formats to unify and contextualise multimodal inputs across disciplines.

Model-centric development for multimodal AI faces unique challenges in effectively and efficiently fusing diverse modalities. While foundation models have proven effective for vision and language tasks, expanding them to other modalities and disciplines[100] could substantially lower development barriers. Guidelines and frameworks for selecting relevant modalities,

comparing multimodal versus unimodal performance, and evaluating the benefits of using many versus few modalities will help maximise model efficiency and utility. As large language models, multimodal foundation models, and generative AI systems grow in prominence [101–104], deployment-centric design plays an increasingly important role in guiding architectural choices, managing inference-time costs, and aligning with data and domain-specific constraints [105,106]. These models also increasingly rely on synthetic or weakly labelled data, or cross-modal supervision to reduce annotation costs [64,107–111], reinforcing the need for deployment-aware data strategies.

Stakeholder engagement and interdisciplinary collaboration are both crucial for progress, as illustrated across the three use cases. Integrating stakeholder input ensures contextual relevance and trust. Standardising data practices, fostering cross-disciplinary knowledge exchange, and developing shared platforms help bridge disciplinary gaps and enable multimodal AI to address complex societal challenges more effectively and holistically [36,37].

Building a dynamic and inclusive multimodal AI community fosters innovation and drives solutions to complex real-world challenges. Collaborative efforts, through regular workshops, forums, and interdisciplinary research initiatives, facilitate knowledge exchange and inspire collective problem-solving across disciplines. Engaging researchers, practitioners, and stakeholders from diverse backgrounds, particularly early-career researchers and underrepresented groups, will not only enrich the AI ecosystem but also broaden the perspectives and expertise shaping it, making it more ‘multimodal’. Community-driven initiatives [112,113], including comprehensive surveys or perspective papers, provide insights that guide future developments, ensuring multimodal AI advances in responsible, impactful directions.

Ultimately, the success of multimodal AI lies in its deployment. By embracing a deployment-centric mindset, the community can turn research breakthroughs into real-world impact across disciplines.

Box 1 Strategic recommendations for advancing multimodal AI across disciplines

Deployment-centric development

- *Safety, reliability, and interpretability*

  Challenge: Ensuring reliable, safe deployment across varied conditions and building user trust to facilitate understanding and broader adoption.

  Recommendation: Translate complex data and outputs into accessible formats (images, text, or speech); incorporate domain-specific knowledge; implement human-in-the-loop systems for verification and validation; conduct rigorous usability testing; develop standards for safety, reliability, and interpretability criteria; and enforce such criteria in research and peer review[36].

- *Scalability and resource efficiency*

  Challenge: Scaling multimodal AI to handle vast data volumes while optimising resource consumption, maintaining system availability, and ensuring uninterrupted communication between system components.

  Recommendation: Innovate scalable and resource-efficient multimodal AI solutions; promote collaboration between AI developers and hardware providers for sustainable resource use; ensure stable system availability and seamless inter-component communication; and develop cloud-based solutions and edge computing to power scalability.

- *Ethical compliance and user preparedness*

  Challenge: Addressing privacy, consent, and bias in applications with profound societal impacts.

  Recommendation: Develop clear ethical guidelines; include end-users in the design; involve human oversight and create user feedback loops for continuous solution refinement; and provide comprehensive training for reliable deployment.

Data-centric development

- *Data scarcity and access*

  Challenge: Limited availability of high-quality multimodal datasets spanning a comprehensive range of modalities, complicated further by privacy and ethical constraints and additional heterogeneity when data sharing spans secure data environments, organisations, and regions.

  Recommendation: Promote open data initiatives, such as anonymised data-sharing programmes; invest in cross-discipline efforts for data collection and curation to improve availability; build data infrastructures that ensure secure, privacy-compliant, and ethics-compliant access across domains and regions; establish guidelines to standardise formats and annotations; and develop advanced tools such as knowledge graphs to integrate structured and unstructured data sources for improved contextual understanding, reasoning, and interoperability.

- *Balanced data representation*

  Challenge: Bias due to the limited availability of data representing underrepresented or less-studied categories (e.g. populations and regions).

  Recommendation: Diversify dataset development to enhance model generalisability and reduce bias; and build and enhance multimodal data platforms (e.g. UK Biobank[114], MIMIC[115], Materials Project[116], ERA5[117]) to improve global data accessibility and quality.

Stakeholder engagement and interdisciplinary collaboration

- *Stakeholder inclusion and alignment*

  Challenge: Limited involvement of stakeholders (e.g. domain experts, end-users, and regulators) during early-stage planning can result in solutions that are technically sound but poorly aligned with deployment contexts.

  Recommendation: Integrate stakeholder input across the AI lifecycle through co-design, participatory planning, and regulatory consultation to ensure contextual relevance, trust, and successful real-world adoption.

- *Cross-disciplinary standards and communication*

  Challenge: Discrepancies in multimodal data standards and communication barriers across disciplines.

  Recommendation: Promote cross-disciplinary standards for multimodal data integration; and organise interdisciplinary exchange events to align goals, foster collaboration, and build a thriving ecosystem.

- *Intellectual property (IP) and workflow adaptation*

  Challenge: IP concerns and resistance to change established workflows.

  Recommendation: Develop adaptive strategies and open collaborative platforms to facilitate cross-disciplinary projects.

Model-centric development

- *Multimodal fusion and modality selection*

  Challenge: Managing diverse properties and patterns across modalities, including avoiding unnecessary modality redundancy, to maintain integrity and utility while enhancing integration.

  Recommendation: Assess the value of additional modalities by comparing unimodal, few-modality, and full-modality models; enhance semantic modality alignment, e.g. via large language models; develop guidelines for optimal modality selection aligned with deployment context and cost-performance constraints; and address other multimodal-AI-specific modelling challenges (Fig. 3b).

- *Foundation models (FMs)*

  Challenge: Developing FMs beyond vision and language requires substantial resources, limiting accessibility across disciplines.

  Recommendation: Develop and expand multimodal FMs beyond vision and language as critical reusable infrastructure to lower barriers; invest in downstream research and advance methods to reduce resource costs, e.g. using synthetic or weakly labelled data to ease training demands[44]; and pair underrepresented modalities (e.g. graphs) with more interpretable ones (e.g. language) to improve usability and broaden accessibility for non-expert stakeholders.

Datasets	Year of release	Data origin	Primary data modalities	Access	Scale	Example applications
TCGA	2008	US	images, texts, omics	Open	~11,000 individuals	Discovery of cancer genes and mutations
MIMIC-IV	2023	US	texts, numerical data, time series	Open	~231,000 individuals	Prediction of mortality in lung cancer patients
Trafficking-10k	2017	US, CA	images, texts	Restricted	~10,000 advertisements	Human trafficking detection
*IGDD	2022	US	texts (chat histories, surveys)	Restricted	~26,700 conversations	Detection of adolescent online risks
nuScenes	2019	US, SG	images, radar data, LiDAR data	Open	~1,400,000 images	Prediction of vehicle traffic trajectories
DAIR-V2X	2021	CN	images, LiDAR data	Open	~71,300 LiDAR and camera frames	3D object detection
Materials project	2013	Worldwide	images, texts, graphs	Open	~160,000 materials	Discovery of new inorganic crystals
MathVista	2024	Worldwide	images, texts, symbolic data	Open	~6,000 mathematical problems	Mathematical problem-solving
iNaturalist	2017	Worldwide	images, timestamps, coordinates, species data	Open	~118,000,000 observations	Species recognition
†ERA5	2018	Worldwide	spatiotemporal data (temperature, humidity, wind)	Open	~745,000 latitude-longitude grids	Reconstruction of historical climate extremes
Monetary policy calls	2022	Worldwide	videos, texts, time series	Open	~340 video conference calls	Prediction of gold price movements
China A-shares market	2022	CN	texts, time series, graphs	Open	~5,130,000 news articles	Forecasting of stock price movements

Healthcare | Social science | Engineering | Science | Sustainability | Finance | Open access | Restricted access

*Subtypes of texts are considered as multiple modalities. †Different 'views' of spatiotemporal data are considered as different modalities [46,47].

Figure 5. Examples of multimodal benchmark datasets. We selected two illustrative examples for each of the six disciplines to showcase diverse multimodal data and provide a starting point for multimodal AI exploration. For each dataset, we report six key attributes: year of release, data origin, data modalities, accessibility, scale of the primary modalities, and example applications. In each discipline, datasets are listed in ascending order by year of release. For the healthcare, science, and finance disciplines, we selected one small-scale dataset for quick experimentation and one larger-scale dataset for comprehensive exploration. The datasets include: TCGA (The Cancer Genome Atlas) [118]; MIMIC-IV (The Medical Information Mart for Intensive Care IV database) [119]; Trafficking-10k (human trafficking advertisement dataset) [120]; IGDD project (Instagram data donation project) [121]; nuScenes (autonomous driving scene dataset) [122]; DAIR-V2X (real-scenarios vehicle to everything dataset) [123]; Materials project (material property dataset) [116]; MathVista (mathematical reasoning dataset) [124]; iNaturalist (citizen science platform for biodiversity data) [125]; ERA5 (European Centre for Medium-Range Weather Forecasts Reanalysis v5) [117]; Monetary policy calls (monetary policy call dataset) [126]; China A-shares market (dataset of public companies listed in China A-shares market) [127]. Modality definitions can vary. IGDD and ERA5 are not strictly multimodal in the traditional sense, but we treat them as such under the broader definition presented in this Perspective. The IGDD dataset contains only text data, but we consider its subtypes (chat histories and surveys) as distinct text modalities. Similarly, ERA5 comprises spatiotemporal variables such as temperature, humidity, and wind, which we treat as complementary 'views' offering distinct information streams for climate modelling, as adopted in prior work [128,129].

Data availability

Source data for Fig. 1, Fig. 2, and Supplementary Fig. 1 are available with this paper and at https://github.com/multimodalAI/multimodal-ai-landscape, where they will be updated annually.

References

1. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
2. Kabir, M. M., Jim, J. R. & Istenes, Z. Terrain detection and segmentation for autonomous vehicle navigation: A state-of-the-art systematic review. *Inf. Fusion* **113**, 102644 (2025).
3. Woloshin, S., Patel, N. & Kesselheim, A. S. False negative tests for SARS-CoV-2 infection—challenges and implications. *New Engl. J. Medicine* **383**, e38 (2020).
4. Zhao, F., Zhang, C. & Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **56**, 1 – 36 (2024).

5. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis Mach. Intell.* **41**, 423–443 (2019).

6. Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M. & Zitnik, M. Multimodal learning with graphs. *Nat. Mach. Intell.* **5**, 340–350 (2023).

7. Xu, P., Zhu, X. & Clifton, D. A. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 12113–12132 (2023).

8. Liang, P. P., Zadeh, A. & Morency, L.-P. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.* **56**, 1–42 (2024).

9. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Medicine* **28**, 1773–1784 (2022).

10. Kline, A. *et al.* Multimodal machine learning in precision health: A scoping review. *npj Digit. Medicine* **5**, 171 (2022).

11. Krones, F., Marikkar, U., Parsons, G., Szmul, A. & Mahdi, A. Review of multimodal machine learning approaches in healthcare. *Inf. Fusion* **114**, 102690 (2025).

12. Notin, P., Rollins, N., Gal, Y., Sander, C. & Marks, D. Machine learning for functional protein design. *Nat. Biotechnol.* **42**, 216–228 (2024).

13. Song, B., Zhou, R. & Ahmed, F. Multi-modal machine learning in engineering design: A review and future directions. *J. Comput. Inf. Sci. Eng.* **24**, 010801 (2024).

14. Ofodile, O. C. *et al.* Predictive analytics in climate finance: Assessing risks and opportunities for investors. *GSC Adv. Res. Rev.* **18**, 423–433 (2024).

15. Quatrini, S. Challenges and opportunities to scale up sustainable finance after the COVID-19 crisis: Lessons and promising innovations from science and practice. *Ecosyst. Serv.* **48**, 101240 (2021).

16. Gupta, V. *et al.* An emotion care model using multimodal textual analysis on COVID-19. *Chaos, Solitons & Fractals* **144**, 110708 (2021).

17. Anshul, A., Pranav, G. S., Rehman, M. Z. U. & Kumar, N. A multimodal framework for depression detection during COVID-19 via harvesting social media. *IEEE Transactions on Comput. Soc. Syst.* (2023).

18. Bordes, F. *et al.* An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247* (2024).

19. Zhang, J., Huang, J., Jin, S. & Lu, S. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2024).

20. van Breugel, B. & van der Schaar, M. Position: Why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, 48976–48993 (2024).

21. Ma, M. *et al.* Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2302–2310 (2021).

22. Wu, R., Wang, H., Chen, H.-T. & Carneiro, G. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825* (2024).

23. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V. & Yu, L. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Adv. Neural Inf. Process. Syst.* **35**, 33536–33549 (2022).

24. Zhao, T., Zhang, L., Ma, Y. & Cheng, L. A survey on safe multi-modal learning systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6655–6665 (2024).

25. Pranjal, R. *et al.* Toward privacy-enhancing ambulatory-based well-being monitoring: Investigating user re-identification risk in multimodal data. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5 (2023).

26. Paleyes, A., Urma, R.-G. & Lawrence, N. D. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.* **55**, 1–29 (2022).

27. Seedat, N., Imrie, F. & van der Schaar, M. Navigating data-centric artificial intelligence with DC-Check: Advances, challenges, and opportunities. *IEEE Transactions on Artif. Intell.* (2023).

28. Jo, A. The promise and peril of generative AI. *Nature* **614**, 214–216 (2023).

29. Brown, T. *et al.* Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901 (2020).

30. Achiam, J. *et al.* GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

31. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 34892–34916 (2023).

32. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 28541–28564 (2023).

33. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).

34. Kreuzberger, D., Kühl, N. & Hirschl, S. Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access* **11**, 31866–31879 (2023).

35. Lavin, A. *et al.* Technology readiness levels for machine learning systems. *Nat. Commun.* **13**, 6039 (2022).

36. Nielsen, M. W. *et al.* Intersectional analysis for science and technology. *Nature* **640**, 329–337 (2025).

37. Lekadir, K. *et al.* FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388** (2025).

38. Huang, Y. *et al.* What makes multi-modal learning better than single (provably). In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 10944–10956 (2021).

39. Meng, X., Babaee, H. & Karniadakis, G. E. Multi-fidelity Bayesian neural networks: Algorithms and applications. *J. Comput. Phys.* **438**, 110361 (2021).

40. Penwarden, M., Zhe, S., Narayan, A. & Kirby, R. M. Multifidelity modeling for physics-informed neural networks (PINNs). *J. Comput. Phys.* **451**, 110844 (2022).

41. Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* **50**, 71–91 (2019).

42. Lunke, S. *et al.* Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nat. Medicine* 1–11 (2023).

43. Wu, E. *et al.* How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nat. Medicine* **27**, 582–584 (2021).

44. van Breugel, B., Liu, T., Oglic, D. & van der Schaar, M. Synthetic data in biomedicine via generative artificial intelligence. *Nat. Rev. Bioeng.* **2**, 991–1004 (2024).

45. Al-Rubaie, M. & Chang, J. M. Privacy-preserving machine learning: Threats and solutions. *IEEE Secur. & Priv.* **17**, 49–58 (2019).

46. Wendland, P. *et al.* Generation of realistic synthetic data using multimodal neural ordinary differential equations. *npj Digit. Medicine* **5**, 122 (2022).

47. Che, L., Wang, J., Zhou, Y. & Ma, F. Multimodal federated learning: A survey. *Sensors* **23** (2023).

48. Wang, Q., Zhan, L., Thompson, P. M. & Zhou, J. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1828–1838 (2020).

49. Yu, Z., Yu, J., Fan, J. & Tao, D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 1839–1848 (2017).

50. Weidinger, L. *et al.* Holistic safety and responsibility evaluations of advanced AI models. *arXiv preprint arXiv:2404.14068* (2024).

51. Luccioni, S., Akiki, C., Mitchell, M. & Jernite, Y. Stable bias: Evaluating societal representations in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 56338–56351 (2023).

52. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **116**, 22071–22080 (2019).

53. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

54. Gawlikowski, J. *et al.* A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023).

55. Khojasteh, D., Davani, E., Shamsipour, A., Haghani, M. & Glamore, W. Climate change and COVID-19: Interdisciplinary perspectives from two global crises. *Sci. Total. Environ.* **844**, 157142 (2022).

56. Elmer, T., Mepham, K. & Stadtfeld, C. Students under lockdown: Comparisons of students' social networks and mental health before and during the COVID-19 crisis in switzerland. *Plos One* **15**, e0236337 (2020).

57. Valdez, D., Ten Thij, M., Bathina, K., Rutter, L. A. & Bollen, J. Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of twitter data. *J. Med. Internet Res.* **22**, e21418 (2020).

58. Liu, Z. *et al.* Near-real-time monitoring of global CO2 emissions reveals the effects of the COVID-19 pandemic. *Nat. Commun.* **11**, 5172 (2020).

59. Zheng, B. *et al.* Satellite-based estimates of decline and rebound in China's CO2 emissions during COVID-19 pandemic. *Sci. Adv.* **6**, eabd4998 (2020).

60. Mosser, P. C. Central bank responses to COVID-19. *Bus. Econ.* **55**, 191–201 (2020).

61. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int. J. Surg.* **78**, 185–193 (2020).

62. Ding, X. *et al.* Wearable sensing and telehealth technology with potential applications in the coronavirus pandemic. *IEEE Rev. Biomed. Eng.* **14**, 48–70 (2020).

63. Charlton, P. H. *et al.* Wearable photoplethysmography for cardiovascular monitoring. *Proc. IEEE* **110**, 355–381 (2022).

64. Zong, Y., Mac Aodha, O. & Hospedales, T. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis Mach. Intell.* 1–20 (2024).

65. Yang, X., Zhang, T. & Xu, C. Cross-domain feature learning in multimedia. *IEEE Transactions on Multimed.* **17**, 64–78 (2014).

66. Han, R. *et al.* Randomised controlled trials evaluating artificial intelligence in clinical practice: A scoping review. *The Lancet Digit. Heal.* **6**, e367–e373 (2024).

67. Plana, D. *et al.* Randomized clinical trials of machine learning interventions in health care: A systematic review. *JAMA Netw. Open* **5**, e2233946–e2233946 (2022).

68. Imrie, F., Davis, R. & van der Schaar, M. Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. *Nat. Mach. Intell.* **5**, 824–829 (2023).

69. Mincu, D. & Roy, S. Developing robust benchmarks for driving forward AI innovation in healthcare. *Nat. Mach. Intell.* **4**, 916–921 (2022).

70. Soni, A. *et al.* Design of a machine learning-based self-driving car. *Mach. Learn. for Robotics Appl.* 139–151 (2021).

71. Badue, C. *et al.* Self-driving cars: A survey. *Expert. Syst. with Appl.* **165**, 113816 (2021).

72. Yeong, D. J., Velasco-Hernandez, G., Barry, J. & Walsh, J. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors* **21**, 2140 (2021).

73. Yang, J. *et al.* Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14662–14672 (2024).

74. Zhang, W. & Xu, J. Advanced lightweight materials for automobiles: A review. *Mater. & Des.* **221**, 110994 (2022).

75. Hansson, S. O., Belin, M.-Å. & Lundgren, B. Self-driving vehicles—an ethical overview. *Philos. & Technol.* **34**, 1383–1408 (2021).

76. Chowdhury, A., Karmakar, G., Kamruzzaman, J., Jolfaei, A. & Das, R. Attacks on self-driving cars and their countermeasures: A survey. *IEEE Access* **8**, 207308–207342 (2020).

77. Dey, K. C., Mishra, A. & Chowdhury, M. Potential of intelligent transportation systems in mitigating adverse weather impacts on road mobility: A review. *IEEE Transactions on Intell. Transp. Syst.* **16**, 1107–1119 (2014).

78. Kamran, S. S. *et al.* Artificial intelligence and advanced materials in automotive industry: Potential applications and perspectives. *Mater. Today: Proc.* **62**, 4207–4214 (2022).

79. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).

80. Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).

81. Liu, P., Yang, R. & Xu, Z. Public acceptance of fully automated driving: Effects of social trust and risk/benefit perceptions. *Risk Analysis* **39**, 326–341 (2019).

82. Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).

83. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).

84. Mathiesen, K. Rating climate risks to credit worthiness. *Nat. Clim. Chang.* **8**, 454–456 (2018).

85. Breitenstein, M., Ciummo, S. & Walch, F. Disclosure of climate change risk in credit ratings. *ECB Occas. Pap. Ser.* (2022).

86. Imran, M., Ofli, F., Caragea, D. & Torralba, A. Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. *Inf. Process. Manag.* **57**, 102261 (2020).

87. Tanir, T., Yildirim, E., Ferreira, C. M. & Demir, I. Social vulnerability and climate risk assessment for agricultural communities in the United States. *Sci. The Total. Environ.* **908**, 168346 (2024).

88. Kumar, P. *et al.* An overview of monitoring methods for assessing the performance of nature-based solutions against natural hazards. *Earth-Science Rev.* **217**, 103603 (2021).

89. Tkachenko, N., Jarvis, S. & Procter, R. Predicting floods with Flickr tags. *PloS One* **12**, e0172870 (2017).

90. Thulke, D. *et al.* ClimateGPT: Towards AI synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646* (2024).

91. Bodnar, C. *et al.* Aurora: A foundation model of the atmosphere. Tech. Rep. MSR-TR-2024-16, Microsoft Research AI for Science (2024).

92. Rasp, S. *et al.* WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024).

93. Morshed, S. R. *et al.* Decoding seasonal variability of air pollutants with climate factors: A geostatistical approach using multimodal regression models for informed climate change mitigation. *Environ. Pollut.* **345**, 123463 (2024).

94. Reichstein, M. *et al.* Early warning of complex climate risk with integrated artificial intelligence. *Nat. Commun.* **16 1**, 2564 (2025).

95. Liang, P. P. *et al.* Multibench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 1–20 (2021).

96. de la Fuente, J. *et al.* Towards a more inductive world for drug repurposing approaches. *Nat. Mach. Intell.* 1–14 (2025).

97. Torabi, F. *et al.* The common governance model: A way to avoid data segregation between existing trusted research environment. *Int. J. Popul. Data Sci.* **8** (2023).

98. Crosswell, L. C. & Thornton, J. M. ELIXIR: A distributed infrastructure for european biological data. *Trends Biotechnol.* **30**, 241–242 (2012).

99. Liang, K. *et al.* A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2024).

100. Archit, A. *et al.* Segment anything for microscopy. *Nat. Methods* 1–13 (2025).

101. Li, C. *et al.* Multimodal foundation models: From specialists to general-purpose assistants. *Foundations Trends Comput. Graph. Vis.* **16**, 1–214 (2024).

102. Fei, N. *et al.* Towards artificial general intelligence via a multimodal foundation model. *Nat. Commun.* **13**, 3094 (2022).

103. Narayanswamy, G. *et al.* Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations* (2025).

104. Cui, H. *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623–633 (2025).

105. Anthropic. Introducing the model context protocol. https://www.anthropic.com/index/model-context-protocol (2024). Accessed June 2025.

106. Weisz, J. D. *et al.* Design principles for generative ai applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–22 (2024).

107. Li, J., Li, D., Xiong, C. & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900 (PMLR, 2022).

108. Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742 (PMLR, 2023).

109. Driess, D. *et al.* Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 8469–8488 (PMLR, 2023).

110. Tan, Z. *et al.* Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 930–957 (2024).

111. Ding, B. *et al.* Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, 1679–1705 (2024).

112. Chan, A.-W. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Medicine* **26**, 1364 – 1374 (2020).

113. de Hond, A. A. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *npj Digit. Medicine* **5**, 2 (2022).

114. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

115. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).

116. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1** (2013).

117. Hersbach, H. *et al.* The ERA5 global reanalysis. *Q. J. Royal Meteorol. Soc.* **146**, 1999–2049 (2020).

118. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol. Onkologia* **2015**, 68–77 (2015).

119. Johnson, A. E. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

120. Tong, E., Zadeh, A., Jones, C. & Morency, L.-P. Combating human trafficking with deep multimodal models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1547–1556 (2017).

121. Razi, A. *et al.* Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–9 (2022).

122. Caesar, H. *et al.* nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631 (2020).

123. Yu, H. *et al.* Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21361–21370 (2022).

124. Lu, P. *et al.* Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations* (2024).

125. iNaturalist contributors. iNaturalist research-grade observations. https://doi.org/10.15468/ab3s5x (2025). Accessed via GBIF.org on 7 July 2025.

126. Mathur, P. *et al.* Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2276–2285 (2022).

127. Cheng, D., Yang, F., Xiang, S. & Liu, J. Financial time series forecasting with multi-modality graph neural network. *Pattern Recognit.* **121**, 108218 (2022).

128. Chen, K. *et al.* Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948* (2023).

129. Jin, Q. *et al.* Spatiotemporal inference network for precipitation nowcasting with multimodal fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **17**, 1299–1314 (2024).

Acknowledgements

This work was enabled and supported by the Alan Turing Institute. We thank T. Chakraborty and C. Li for inspiring this work, D. A. Clifton for his invaluable support, and T. Dunstan for contributing to the climate change adaptation section. J.Z. is supported by donations from D. Naik and S. Naik. S.Z. is supported by EPSRC (grant no. EP/Y017544/1). T.L.vdP. was supported by EPSRC (grant no. EP/Y028880/1). A.V. is supported by UKRI CDT in AI for Healthcare (grant no. EP/S023283/1). A.G. is supported by the Research Council of Norway (secureIT project, no. 288787). M.Z. is supported by EPSRC (grant no. EP/X031276/1). C.T. is supported by UKRI CDT in AI-enabled Healthcare (grant no. EP/S021612/1). R.L. is supported by the Royal Society (grant no. IEC\NSFC\233558). L.V.Z. is supported by NERC (grant no. NE/W004747/1). O.T. is supported by UKRI CDT in Application of Artificial Intelligence to the study of Environmental Risks (grant no. EP/S022961/1). Z.S. is supported by Google DeepMind. O.A. is supported by NIHR Barts BRC (grant no. NIHR203330). T.N.D. is supported by UKRI CDT in Accountable, Responsible and Transparent AI (grant no. EP/S023437/1). L.F. is supported by MRC (grant no. MR/W006804/1). N.J. is supported by the EU's co-funded HE project MuseIT (grant no. 101061441). M.V. is supported by St George's Hospital Charity. A.C-C. is supported by EPSRC (grant no. EP/Y028880/1). H.W. is supported by MRC (grant no. MR/X030075/1). C.C. is supported by the Royal Society (grant no. GS\R2\242355). T.Z. was supported by the Royal Academy of Engineering (grant no. RF\201819\18\109). G.G.S. is supported by EPSRC (grant no. EP/Y009800/1). P.H.C. is supported by BHF (grant no. FS/20/20/34626). H.L. is supported by EPSRC (grant no. UKRI396). The views expressed in this material are those of the authors and do not necessarily represent the views of their affiliated institutions or funders.

Author contributions

X.L., J.Z., S.Z. and H.L. contributed equally. X.L., J.Z., S.Z., T.L.vdP., S.T., T.Z., W.K.C., P.H.C., and H.L. conceptualised the manuscript. X.L., J.Z., S.Z., A.G., M.Z., L.vZ., O.T., and H.L. designed the figures. X.L. and H.L. coordinated the entire project. All authors contributed to writing, resources, or editing.

Competing interests

G.G.S. is a scientific advisory board member at BioAIHealth. P.H.C. provides consulting services to Cambridge University Technical Services for wearable manufacturers. The remaining authors declare no competing interests.

Supplementary information
Towards deployment-centric multimodal AI beyond vision and language

S1. Multimodal AI beyond the three use cases

This section highlights discipline-specific advancements and challenges in multimodal AI that extend beyond the three primary use cases of pandemic response, self-driving car design, and climate change adaptation. By exploring advancements and challenges in healthcare, social science, engineering, science, sustainability, and finance, we showcase the versatility of multimodal AI across diverse domains. These insights underline the importance of interdisciplinary collaboration in addressing complex global challenges while paving the way for future advancements.

Healthcare

Beyond its critical role in pandemic response, multimodal AI holds transformative potential in healthcare by advancing diagnostics, personalising treatment, and improving patient care. By integrating diverse data sources such as electronic health records (EHRs), imaging, biosignals, and multiomics, multimodal AI systems can provide comprehensive insights into patient health[1]. For example, integrating wearable biosignals with EHRs can enable continuous monitoring, supporting the early detection of chronic conditions such as arrhythmias and heart failure[2]. In oncology, the fusion of imaging modalities and multiomics data can enhance cancer diagnosis and treatment planning[3], offering a more nuanced understanding of disease progression. Similarly, in neurology, combining imaging, sensor data, and clinical observations can facilitate the early detection of neurodegenerative diseases[4].

Wider adoption of multimodal AI is hindered by challenges such as privacy concerns, the heterogeneity of healthcare data, and the scarcity of high-quality datasets. Addressing these challenges requires standardised methodologies, privacy-preserving techniques, and interdisciplinary collaborations. Building datasets that reflect the complexities of patient pathways and addressing bottlenecks in data integration are crucial steps towards integrating multimodal AI into routine clinical practice. Techniques such as federated learning, differential privacy, and privacy-preserving generative AI for synthetic multimodal data generation can help mitigate privacy risks, enabling the development of robust and trustworthy AI systems. Integrating multimodal data, such as multiomics[5], into routine clinical workflows will require rigorous validation and improved standardisation. The limited use of randomised controlled trials (RCTs) to validate AI in clinical settings[6] has prompted the development of community-based guidelines[7] to improve reliability and transparency.

Social science

In social science, multimodal AI can drive innovations in behavioural analysis, public policy decision-making, and societal impact assessments. Beyond its applications in pandemic response, multimodal AI can help monitor societal behaviours and safeguard vulnerable populations, such as protecting children online, by integrating text, image and metadata from social media[8]. Multimodal AI systems can also help detect online criminal activity[9] and analyse societal trends to inform public policy[10]. Additionally, acoustic data can add depth to communication analysis by capturing emotional nuances[11]. Combining economic transaction data with geospatial information can reveal trends in societal decision-making[12].

Persistent challenges include restricted access to sensitive data such as social media or mental health records, privacy concerns, and the complexity of annotating subjective human behaviours that differ across cultures. Moreover, specialised evaluation metrics are needed to assess the effectiveness of these systems in capturing nuanced social trends. To advance multimodal AI in social science, efforts should focus on expanding data diversity, addressing privacy challenges, developing robust annotation and evaluation methods, and fostering robust ethical frameworks and cross-disciplinary expertise. These efforts will ensure that the insights generated are fair, transparent, and aligned with societal goals.

Engineering

Beyond self-driving cars, multimodal AI supports advanced autonomous systems, including robotics, precision manufacturing, and medical applications. Robots equipped with multimodal AI can integrate vision, haptic feedback, and sensor data to perform delicate tasks, such as in surgical procedures[13] or complex manipulation[14], where precision is paramount. Additionally, multimodal AI can enable seamless human-robot interaction by integrating speech, vision, and environmental understanding[15]. Speech technology and natural language processing can enhance user interfaces, improving interaction while minimising distractions for operators in robotic applications.

Despite the potential of multimodal AI, challenges remain in achieving effective data fusion, system interoperability, and robust model generalisation. Furthermore, the lack of diverse datasets, particularly for edge cases or atypical deployment environments, limits the scalability of these systems. Establishing open platforms for sharing data and toolkits can accelerate innovation and ensure interoperability across engineering applications. Integrating insights from legal studies and social sciences can enhance compliance with evolving regulations and improve user-centred design, boosting societal acceptance of

autonomous systems. Environmental science can contribute to the development of energy-efficient and sustainable designs, particularly for robotics and manufacturing processes. Addressing these challenges requires interdisciplinary collaborations to improve data diversity and develop scalable, interoperable systems.

Science

Multimodal AI is transforming scientific discovery by integrating diverse data sources to model complex phenomena. Beyond its role in vehicle dynamics simulation, as seen in self-driving car design, multimodal AI is accelerating advancements in materials science. For instance, integrating experimental and computational data enables the development of improved batteries, fuel cells and supercapacitors[16]. We can incorporate domain knowledge via techniques such as physics-informed neural networks[17,18] and neurosymbolic AI[19] to improve the accuracy of scientific simulations and enhance model trustworthiness and predictive power across applications[20,21].

Remaining challenges of multimodal AI for scientific discovery lie in integrating heterogeneous data, ranging from atomic-level properties to macroscopic observations, incorporating synthetic data, and addressing the scarcity of high-fidelity datasets amidst an abundance of low-fidelity data. Advancing scientific discovery with multimodal AI requires comprehensive datasets, rigorous data standards, and interdisciplinary research environments. These efforts will ensure that models are aligned with real-world dynamics, unlocking new possibilities in materials science, physics, and beyond[22,23].

Sustainability

Multimodal AI can help address sustainability challenges, such as biodiversity conservation and environmental monitoring[24,25]. Beyond its applications in climate change adaptation, multimodal AI can combine geospatial data, bioacoustic recordings, and environmental DNA to track species and ecosystems[26], supporting conservation planning and enhancing understanding of ecosystem dynamics. In terrestrial environments, ecoacoustic and LiDAR data can be combined to model biodiversity variation across complex landscapes[27]. In marine settings, integrating remote sensing with in situ sensors enables long-term monitoring of ecosystem health and debris pathways[28].

Geographic imbalances in data availability, particularly in underrepresented regions, limit model fairness and accuracy. Marine environments pose unique sensing challenges[29], including high turbidity and limited optical visibility, which limit the effectiveness of conventional imaging approaches. This necessitates innovative solutions such as autonomous underwater vehicles and acoustic sensors to access deeper waters, remote habitats, and ecologically sensitive regions[30]. Future advancements will depend on building diverse, scalable datasets, de-biasing models, and leveraging digital twins[31] for real-time monitoring and actionable insights. By addressing these challenges, multimodal AI can make meaningful contributions to global sustainability efforts[32].

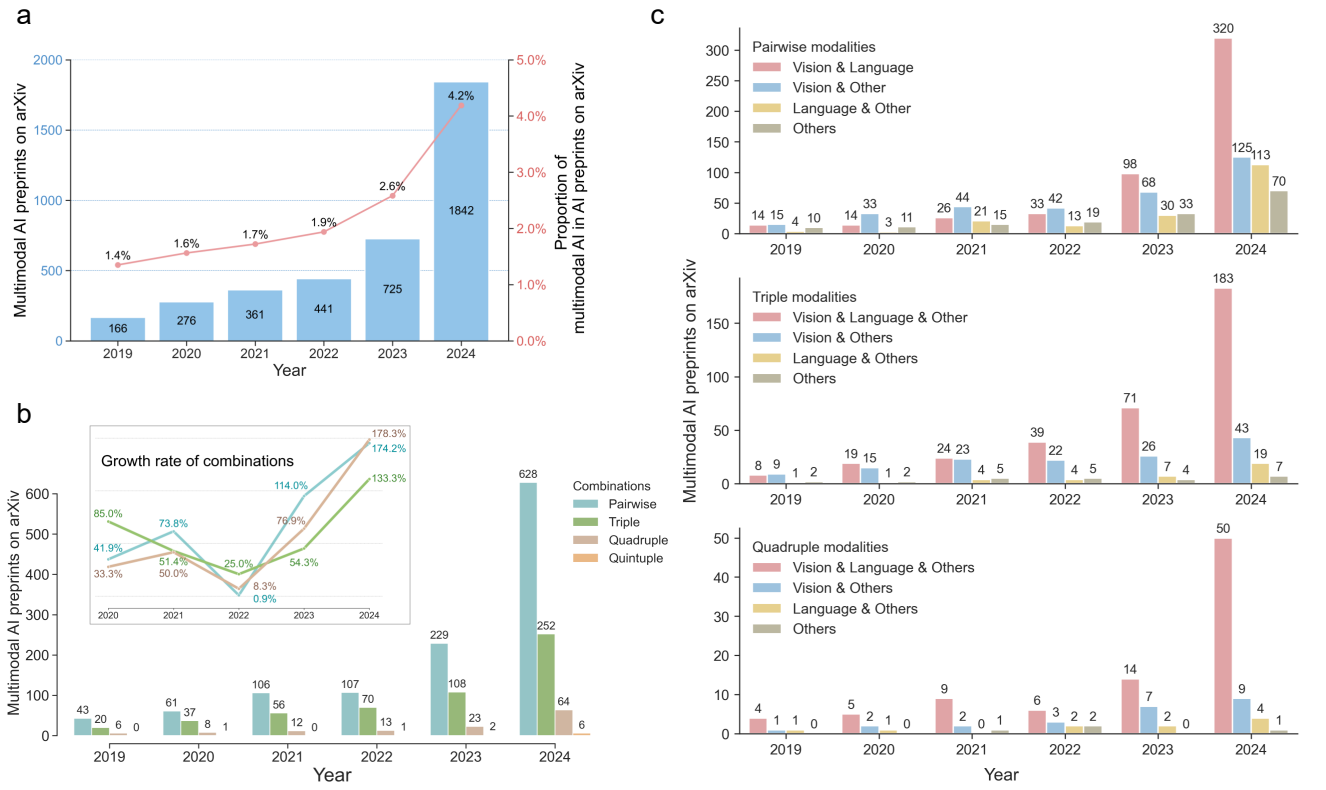
Finance

Multimodal AI can advance finance by integrating financial, environmental, and social data to improve risk assessments, market forecasting, and sustainable investment strategies. In particular, it can enhance ESG (environmental, social, and governance) investing by helping stakeholders make more responsible financial decisions[33]. Integrating financial data and environmental data, such as satellite imagery, can enable a more comprehensive approach to risk management and resilience[34]. Additionally, graph-based multimodal models can monitor systemic risks by capturing complex relationships within financial ecosystems[33], supporting the early detection of market disruptions. Traditional financial models often struggle to respond to economic shocks such as natural disasters. Integrating real-time, multimodal data can improve model adaptability and accuracy under volatile conditions, ultimately enhancing financial stability[35].

Challenges such as data privacy, regulatory complexity, and the need for explainability continue to be substantial barriers. Addressing these issues requires comprehensive multimodal datasets and transparent evaluation frameworks to build trust and ensure compliance with regulations. Interdisciplinary collaborations can help refine multimodal AI systems in the finance sector. Environmental science informs environmental risk assessments, enabling more resilient investment strategies and credit evaluations in the face of severe adverse events. Social sciences provide socio-political insights that promote transparency, equity, and ethical compliance, ensuring that financial practices align with societal goals. These collaborations align financial systems with ethical and sustainable practices while equipping stakeholders to make resilient decisions in an evolving financial landscape.

S2. Methods and extended analysis of the multimodal AI landscape

All arXiv preprints[36] from 2019 to 2024 were analysed to examine the latest trends in multimodal AI research. ArXiv was chosen for its open access, broad adoption within the AI community, and ability to reflect emerging developments more rapidly than peer-reviewed platforms. Raw data were extracted from the publicly available arXiv metadata dump[37] hosted on Kaggle and filtered using queries for six general AI terms “AI”, “A.I.”, “artificial intelligence”, “machine learning”, “deep learning”,



Supplementary Figure 1. Landscape of multimodal AI research (2019–2024). a, Yearly growth in the number of multimodal AI preprints on arXiv (bars) and their proportion among all AI preprints (line), both showing an accelerating upward trend since 2022, reflecting the field's rapid expansion, likely driven by the large language model (LLM) revolution. **b,** Distribution of multimodal AI preprints by the number of combined modalities. The inset line plot shows the growth rate of each combination. While pairwise combinations remain the most prevalent, combinations involving more modalities are steadily gaining attention, reflecting increasing interest in richer data fusion for tasks requiring diverse information sources. **c,** Detailed breakdown of pairwise, triple, and quadruple modality combinations, highlighting trends across four modality clusters: vision and language, vision and other(s), language and other(s), and others. Combinations involving vision and language remain dominant. In particular, the number of pairwise combinations involving language has increased substantially from 2023 to 2024, likely due to the LLM-driven surge in multimodal research.

and "neural network" appearing in either the title or the abstract to identify relevant preprints. The list was further refined to identify multimodal AI preprints by searching for "multimodal" and "multi-modal", and specific modalities were identified through targeted queries: "vision", "image", "video", and "visual" for vision; "text", "language", and "textual" for language; and similar sets of terms for the other six modalities: time series, graph, audio, spatial, sensor, and tabular data (see Data availability). As these queries are inherently approximate, conclusions are confined to overall trends supported by strong evidence rather than specific numerical details.

Supplementary Figure 1 provides more detailed analyses of the multimodal AI landscape, including statistics on triple and quadruple modality combinations.

References

1. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Medicine* **28**, 1773–1784 (2022).

2. Krittanawong, C. *et al.* Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. *Nat. Rev. Cardiol.* **18**, 75–91 (2021).

3. Steyaert, S. *et al.* Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. Mach. Intell.* **5**, 351–362 (2023).

4. Huang, G., Li, R., Bai, Q. & Alty, J. Multimodal learning of clinically accessible tests to aid diagnosis of neurodegenerative disorders: a scoping review. *Heal. Inf. Sci. Syst.* **11**, 32 (2023).

5. Lunke, S. *et al.* Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nat. Medicine* 1–11 (2023).

6. Han, R. *et al.* Randomised controlled trials evaluating artificial intelligence in clinical practice: A scoping review. *The Lancet Digit. Heal.* **6**, e367–e373 (2024).

7. Chan, A.-W. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Medicine* **26**, 1364 – 1374 (2020).

8. Ali, S. *et al.* Getting meta: a multimodal approach for detecting unsafe conversations within instagram direct messages of youth. *Proc. ACM on Human-Computer Interact.* **7**, 1–30 (2023).

9. Goyal, B. *et al.* Detection of fake accounts on social media using multimodal data with deep learning. *IEEE Transactions on Comput. Soc. Syst.* (2023).

10. Androutsopoulou, A. & Charalabidis, Y. A framework for evidence based policy making combining big data, dynamic modelling and machine intelligence. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, 575–583 (2018).

11. Zhu, X. *et al.* A review of key technologies for emotion analysis using multimodal information. *Cogn. Comput.* **16**, 1504–1530 (2024).

12. Huang, X., Wang, S., Lu, T., Liu, Y. & Serrano-Estrada, L. Crowdsourced geospatial data is reshaping urban sciences. *Int. J. Appl. Earth Obs. Geoinformation* **127**, 103687 (2024).

13. He, L. *et al.* Robotic simulators for tissue examination training with multimodal sensory feedback. *IEEE Rev. Biomed. Eng.* **16**, 514–529 (2022).

14. Wang, T., Zheng, P., Li, S. & Wang, L. Multimodal human–robot interaction for human-centric smart manufacturing: a survey. *Adv. Intell. Syst.* **6**, 2300359 (2024).

15. Duncan, J. A., Alambeigi, F. & Pryor, M. W. A survey of multimodal perception methods for human–robot interaction in social environments. *ACM Transactions on Human-Robot Interact.* **13**, 1–50 (2024).

16. Liu, X. *et al.* Recent advances in artificial intelligence boosting materials design for electrochemical energy storage. *Chem. Eng. J.* 151625 (2024).

17. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).

18. Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).

19. Garcez, A. d. & Lamb, L. C. Neurosymbolic AI: The 3rd wave. *Artif. Intell. Rev.* 1–20 (2023).

20. Yang, L., Liu, S., Meng, T. & Osher, S. J. In-context operator learning with data prompts for differential equation problems. *Proc. Natl. Acad. Sci.* **120**, e2310142120 (2023).

21. Rezaei-Shoshtari, S., Hogan, F. R., Jenkin, M., Meger, D. & Dudek, G. Learning intuitive physics with multimodal generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 6110–6118 (2021).

22. Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **625**, 476–482 (2024).

23. Raayoni, G. *et al.* Generating conjectures on fundamental constants with the ramanujan machine. *Nature* **590**, 67–73 (2021).

24. Gui, S., Song, S., Qin, R. & Tang, Y. Remote sensing object detection in the deep learning era—a review. *Remote. Sens.* **16**, 327 (2024).

25. Van der Plas, T. L., Alexander, D. G. & Pocock, M. J. Monitoring protected areas by integrating machine learning, remote sensing and citizen science. *Ecol. Solutions Evid.* **6**, e70040 (2025).

26. Pollock, L. J. *et al.* Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nat. Rev. Biodivers.* 1–17 (2025).

27. Rappaport, D. I., Royle, J. A. & Morton, D. C. Acoustic space occupancy: Combining ecoacoustics and lidar to model biodiversity variation and detection bias across heterogeneous landscapes. *Ecol. Indic.* **113**, 106172 (2020).

28. Maximenko, N. *et al.* An integrated observing system for monitoring marine debris and biodiversity. *Oceanography* **34**, 52–59 (2021).

29. Briciu-Burghina, C., Power, S., Delgado, A. & Regan, F. Sensors for coastal and ocean monitoring. *Annu. Rev. Anal. Chem.* **16**, 451–469 (2023).

30. Whitt, C. *et al.* Future vision for autonomous ocean observations. *Front. Mar. Sci.* **7**, 697 (2020).

31. Li, X. *et al.* Big data in earth system science and progress towards a digital twin. *Nat. Rev. Earth & Environ.* **4**, 319–332 (2023).

32. Zhao, T. *et al.* Artificial intelligence for geoscience: Progress, challenges and perspectives. *The Innov.* (2024).

33. Ang, G. & Lim, E.-P. Learning dynamic multimodal network slot concepts from the web for forecasting environmental, social and governance ratings. *ACM Transactions on Web* **18**, 1–32 (2024).

34. Leng, M., Li, Z., Dai, W. & Shi, B. The power of satellite imagery in credit scoring: a spatial analysis of rural loans. *Annals Oper. Res.* 1–38 (2024).

35. Cao, L. AI and data science for smart emergency, crisis and disaster resilience. *Int. J. Data Sci. Anal.* **15**, 231–246 (2023).

36. Ginsparg, P. Arxiv at 20. *Nature* **476**, 145–147 (2011).

37. arXiv.org submitters. arxiv dataset (2024).