2023

MIT-IBM Watson AI Lab

Annual Report

2024

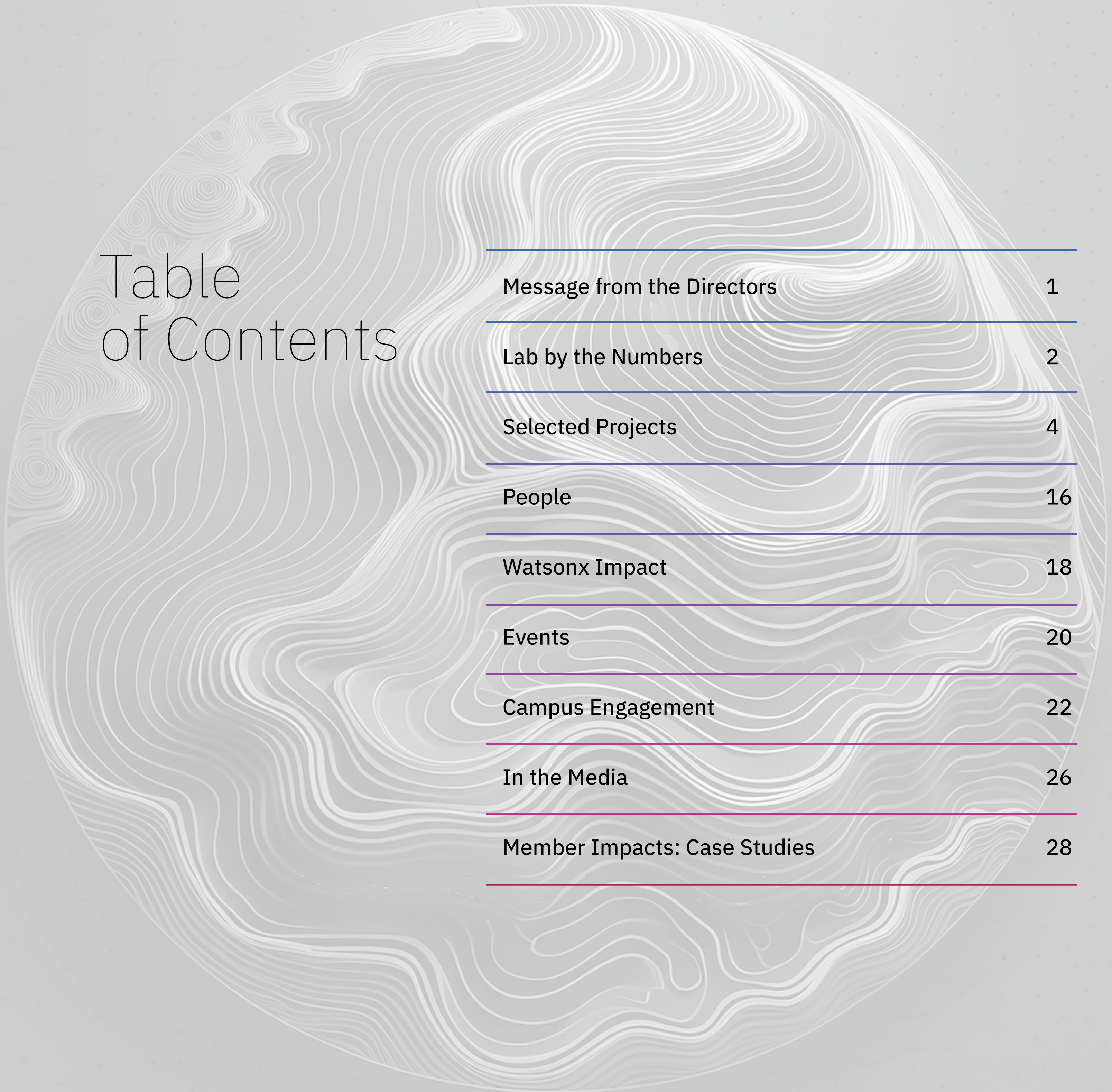


Table of Contents

Message from the Directors	1
Lab by the Numbers	2
Selected Projects	4
People	16
Watsonx Impact	18
Events	20
Campus Engagement	22
In the Media	26
Member Impacts: Case Studies	28

Message from the Directors

Not long ago, whole fictional worlds were built around the promise of AI; now, MIT-IBM Watson AI Lab researchers are bringing it into reality, with an eye toward fairness, reliability, efficiency, and trustworthiness.

Reflecting on the past year, we are incredibly proud of the remarkable progress our collaboration continues to make in advancing impactful solutions for society and business, creating strategic relationships, and preparing students to be AI leaders. Now, with wider public buy in and acknowledgement of AI's benefit, we see that our mission to develop breakthroughs has never been more important or exciting. Together, MIT, IBM, and our corporate members have successfully navigated complex challenges and consistently pushed the boundaries of innovation through novel ideation. From enhancing healthcare and the design process to optimizing business operations and ensuring safety and privacy, our contributions are driving meaningful change and creating value across industries.

Apart from work with our corporate members, a significant portion of the Lab's portfolio has translated into improvements for IBM's open-source Granite foundation models and watsonx applications. Research input to the suite of enterprise-ready models include extending natural language to multiple modalities, more efficient tuning and alignment of LLMs, quantifying reliability for safety and fairness, reducing compute and memory requirements with hardware-aware and hardware-agnostic optimizations, and much more. Additionally, it's the ideas and implementation from students that help make these achievements possible.

We are particularly enthusiastic about our growing commitment to student outcomes. With the addition of our MIT-IBM Watson AI Lab Internship Program launching in 2025, we are further able to get IBM's open-sourced models and tools into the classroom and equip the next generation of AI professionals with the skills and knowledge they seek. In alignment with IBM's open-source technology and open science thrusts, we are seeing that our projects and mentorship are empowering young researchers — helping them to wield and create new AI tools to advance research, facilitate meaningful collaborations, and accelerate scientific discovery.

Our progress would not be possible without the dedication and creativity of our talented team and the support of our corporate members. Looking ahead, we remain focused on expanding our business solutions, aligning our innovations with human values, promoting open-source research and making AI a driving force for positive transformation. Thank you for being an essential part of our journey.

Here, we spotlight research, student, and industry successes from the 2023-2024 year, and look forward to seeing greater progress, creativity, and impact in the future.

David Cox and Aude Oliva

IBM Research MIT

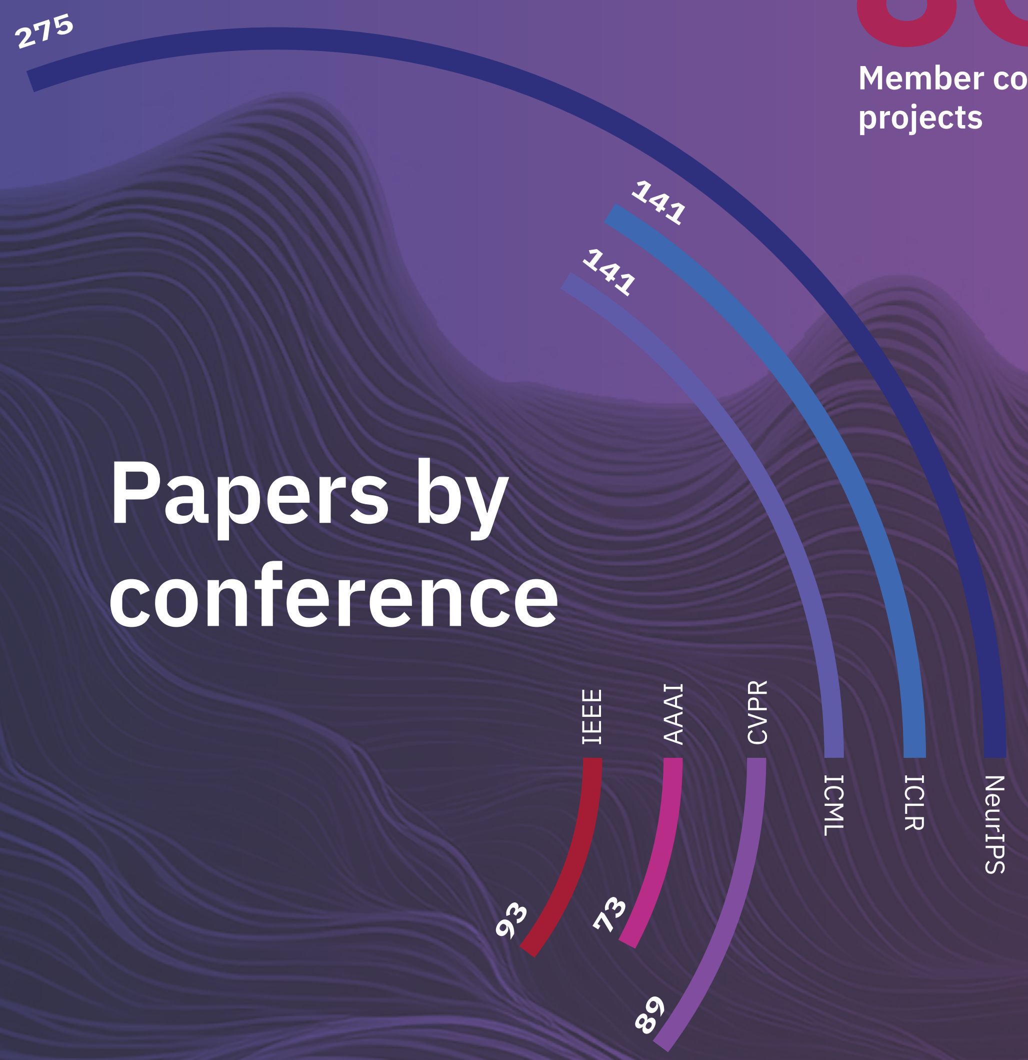
Lab by the Numbers

55
Active projects

Over **500**
Project proposals submitted from MIT and IBM

80+
Member co-funded projects

Papers by conference



$240M
10-year investment to found a joint lab

Over **1,100**
Peer-reviewed publications

1,123
News and media mentions

90,437
Citations

137
H-index

More than **110**
Lab-supported student theses

6
Current member companies

Selected Projects

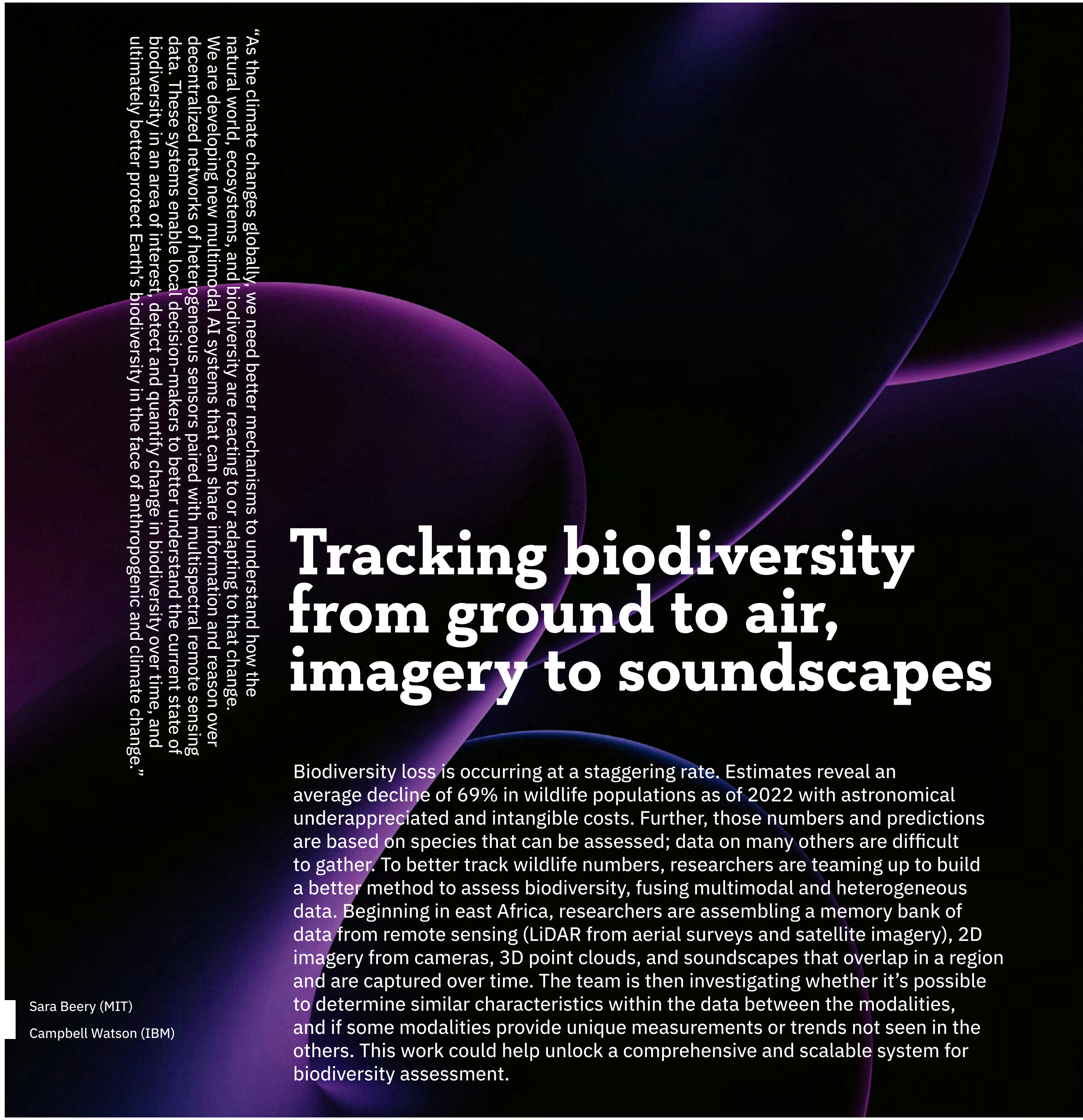
Transforming bold ideas into scalable, high-impact technologies motivates and mobilizes the MIT-IBM Watson AI Lab. As an academic-industry collaboration, we adapt quickly to pursue multiple promising research paths while leveraging significant computational power and multi-domain expertise to dig deeper into fundamental AI questions, upskilling along the way.

During the life of the Lab, we've achieved significant milestones — reducing computational demands for computer vision and LLM models, advancing performance and accuracy through new frameworks, and expanding capabilities in both existing and emerging models. This progress underscores our commitment to building efficient, cutting-edge AI systems.

Currently, our research portfolio targets challenges with groundbreaking potential, from operationalizing robust AI and enabling real-time inference on edge devices, to designing explainable algorithms and fashioning hardware-aware computing. In the past year, we've accelerated work in generative AI, synthetic data, and foundation models, unlocking smarter decision-making and revolutionizing enterprise workflows. By developing predictive tools and advanced analytics, we are uncovering insights and trends to enhance assistive agents, novel material design, multimodal learning, and causal decision-making in the workplace.

Anchored by an AI-first ethos, we're delivering actionable, valuable solutions that empower us and our member organizations to lead in rapidly evolving digital landscapes — with tangible results. Boasting a cumulative h-index surpassing 130 and over 90,000 citations, as well as numerous benefits to society, the Lab's impact speaks for itself.

The following Lab projects illustrate the depth and versatility of our team's work to create meaningful change.



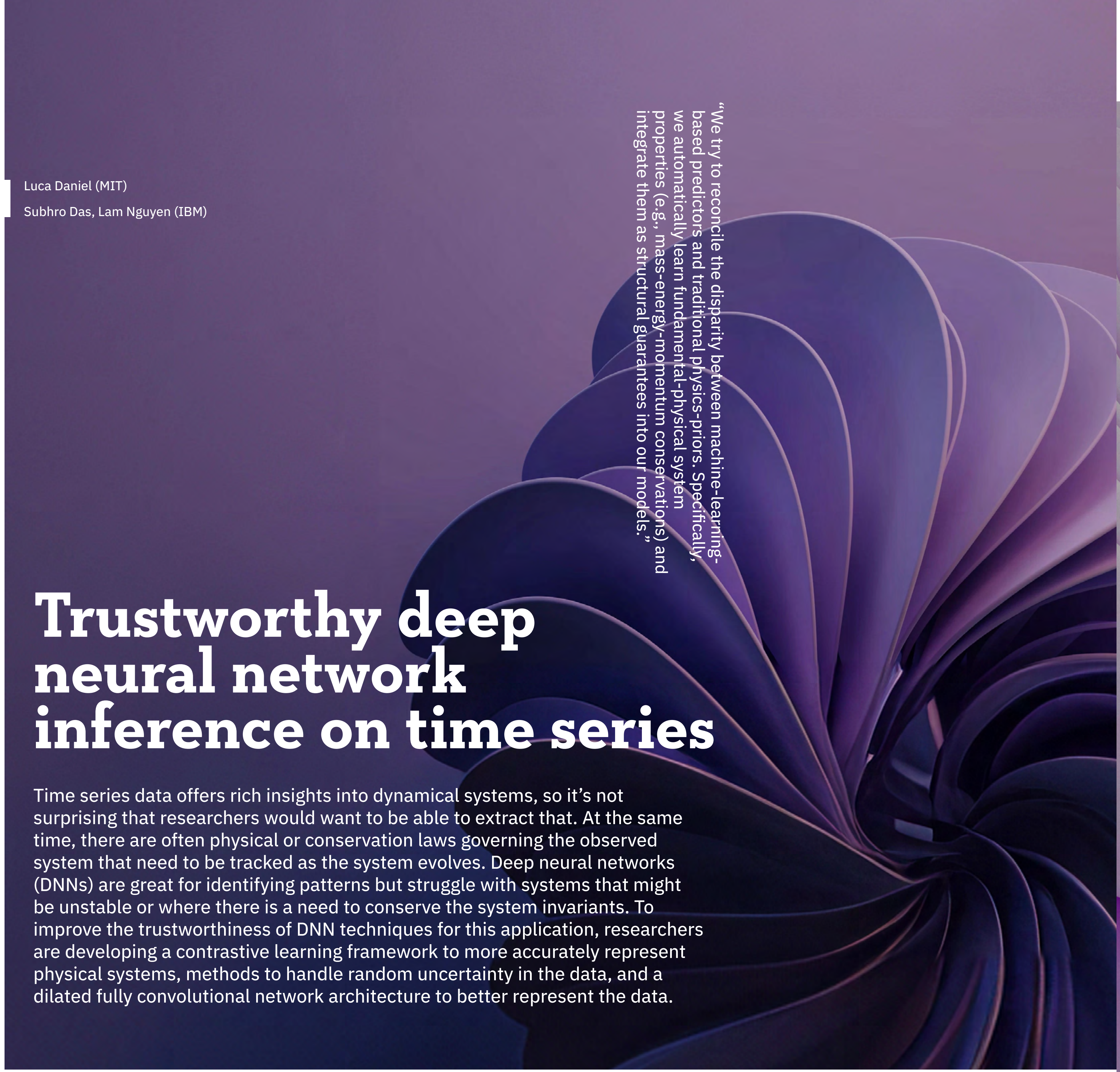
Tracking biodiversity from ground to air, imagery to soundscapes

Biodiversity loss is occurring at a staggering rate. Estimates reveal an average decline of 69% in wildlife populations as of 2022 with astronomical underappreciated and intangible costs. Further, those numbers and predictions are based on species that can be assessed; data on many others are difficult to gather. To better track wildlife numbers, researchers are teaming up to build a better method to assess biodiversity, fusing multimodal and heterogeneous data. Beginning in east Africa, researchers are assembling a memory bank of data from remote sensing (LiDAR from aerial surveys and satellite imagery), 2D imagery from cameras, 3D point clouds, and soundscapes that overlap in a region and are captured over time. The team is then investigating whether it's possible to determine similar characteristics within the data between the modalities, and if some modalities provide unique measurements or trends not seen in the others. This work could help unlock a comprehensive and scalable system for biodiversity assessment.

> "As the climate changes globally, we need better mechanisms to understand how the natural world, ecosystems, and biodiversity are reacting to or adapting to that change. We are developing new multimodal AI systems that can share information and reason over decentralized networks of heterogeneous sensors paired with multispectral remote sensing data. These systems enable local decision-makers to better understand the current state of biodiversity in an area of interest, detect and quantify change in biodiversity over time, and ultimately better protect Earth's biodiversity in the face of anthropogenic and climate change."

Sara Beery (MIT)

Campbell Watson (IBM)



Luca Daniel (MIT)

Subhro Das, Lam Nguyen (IBM)

"We try to reconcile the disparity between machine-learning-based predictors and traditional physics-priors. Specifically, we automatically learn fundamental-physical system properties (e.g., mass-energy-momentum conservations) and integrate them as structural guarantees into our models."

Trustworthy deep neural network inference on time series

Time series data offers rich insights into dynamical systems, so it's not surprising that researchers would want to be able to extract that. At the same time, there are often physical or conservation laws governing the observed system that need to be tracked as the system evolves. Deep neural networks (DNNs) are great for identifying patterns but struggle with systems that might be unstable or where there is a need to conserve the system invariants. To improve the trustworthiness of DNN techniques for this application, researchers are developing a contrastive learning framework to more accurately represent physical systems, methods to handle random uncertainty in the data, and a dilated fully convolutional network architecture to better represent the data.

Harnessing the power of generative models to improve visual representation learning

Kaiming He (MIT)

Rogerio Feris (IBM)

In order to create the impressive, realistic, and diverse images that we see today, image generators, like diffusion models, need to have a firm grasp on relationships between visual content, class labels, and text prompts. Further, they exploit this learned information to look back and iterate as they produce the final image. In contrast, visual recognition tasks (e.g., object detection and classification) rely on feedforward neural networks, which transform images into abstract representations, and unfortunately, their advancement has recently hit a plateau. Researchers are bridging this gap by transforming generative models into tools for extracting rich, centralized visual embeddings. By reformulating discriminative models as a reverse image generation process, the team is exploring how these generative models can learn from the alignment between image output and labels, ultimately advancing our understanding of visual data and removing a bottleneck in the development of visual representation learning for multimodal systems applications.

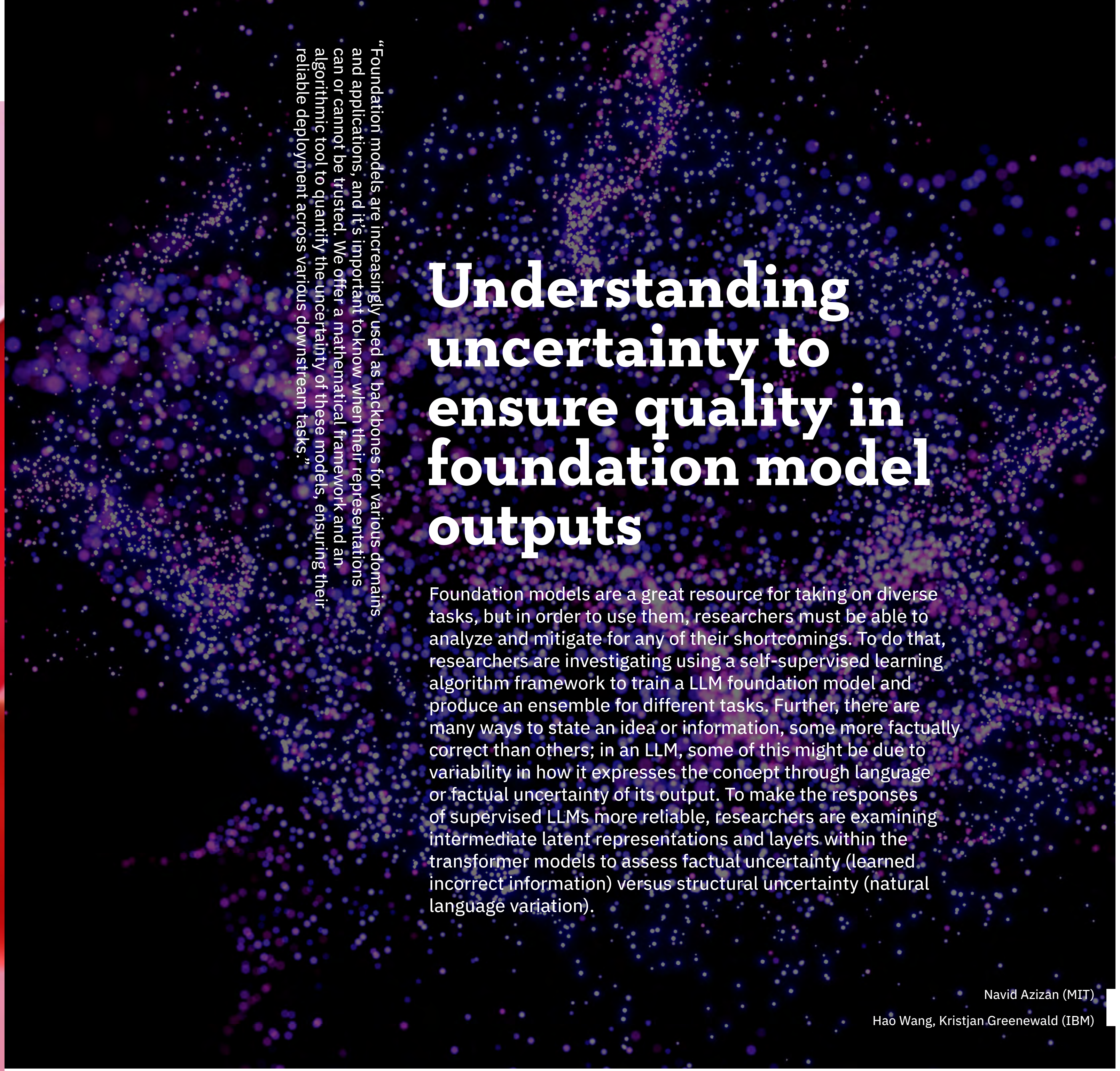
"Our goal is to transform a pre-trained generative model into a powerful visual representation extractor in order to advance and rejuvenate the classical 'analysis-by-synthesis' philosophy, potentially establishing new paradigms and theories for learning visual representations from generative models."

Karen Zheng (MIT)
Markus Ettl (IBM)

Consistent and competitive pricing for marketplace bidding

Quoting the optimal pricing for a product or service in response to a tender depends on many factors, such as the particular customer, the type of market or industry, knowledge of the individual providing the quote, and consistency of pricing decisions. In order to develop competitive pricing strategies for public procurement auctions, researchers are leveraging advanced AI, including counterfactual machine learning and large-scale, pre-trained models, such as transformers, to build a science-based, data-driven, automated pricing recommendation system. Foundation models are working with optimization algorithms to develop dynamic pricing strategies that respond to real-time data. These algorithms can adjust prices based on factors like bid activity, time remaining in the auction, and competitor bids. Additionally, they can assist sales managers by providing data-driven recommendations, enabling a collaborative approach to dynamic pricing.

„Advanced AI, optimization algorithms, and foundation models will work together to develop dynamic pricing strategies that respond to real-time data and enable a collaborative approach to pricing, where AI enhances human expertise."

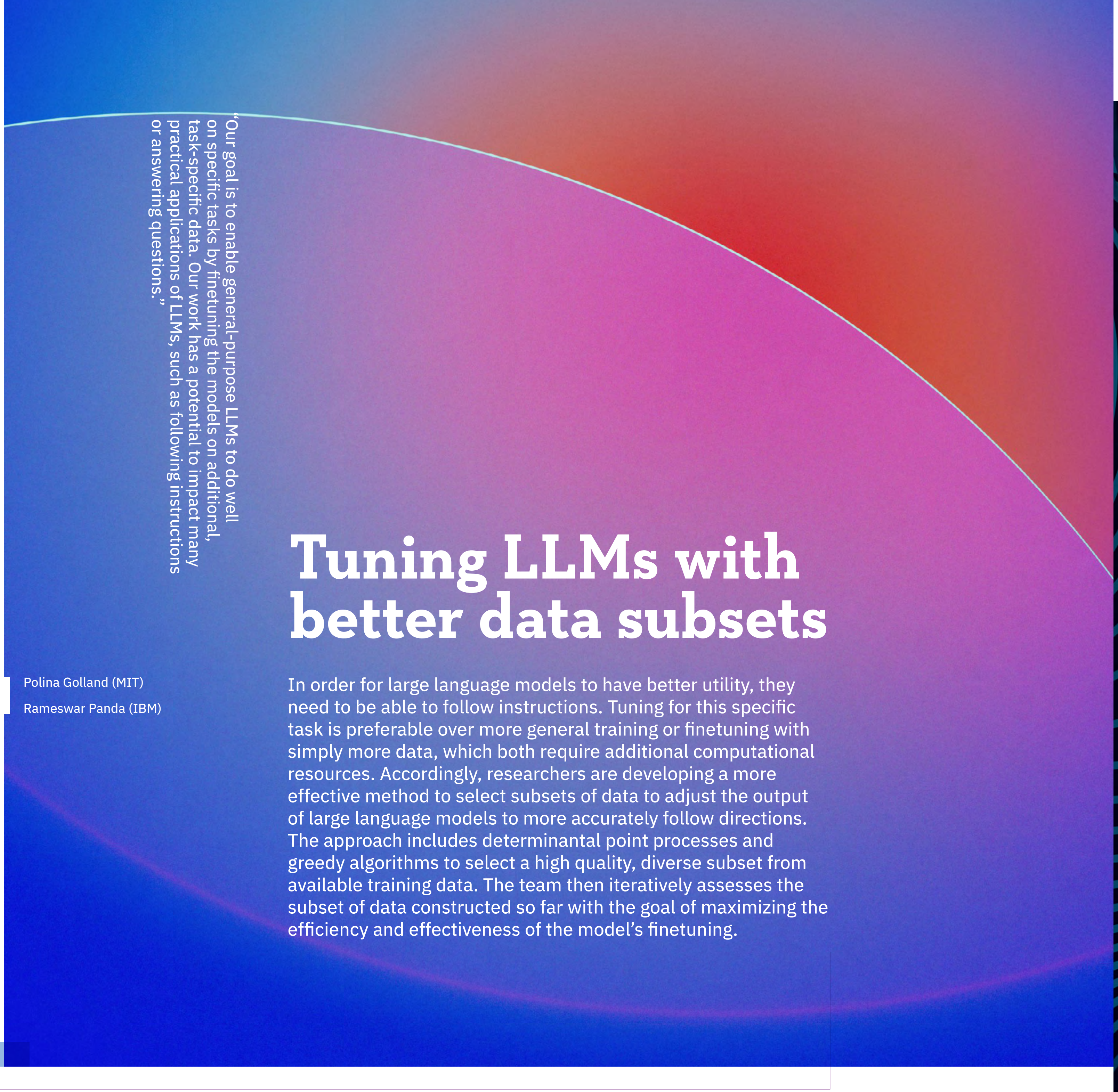


„Foundation models are increasingly used as backbones for various domains and applications, and it's important to know when their representations can or cannot be trusted. We offer a mathematical framework and an algorithmic tool to quantify the uncertainty of these models, ensuring their reliable deployment across various downstream tasks."

Understanding uncertainty to ensure quality in foundation model outputs

Foundation models are a great resource for taking on diverse tasks, but in order to use them, researchers must be able to analyze and mitigate for any of their shortcomings. To do that, researchers are investigating using a self-supervised learning algorithm framework to train a LLM foundation model and produce an ensemble for different tasks. Further, there are many ways to state an idea or information, some more factually correct than others; in an LLM, some of this might be due to variability in how it expresses the concept through language or factual uncertainty of its output. To make the responses of supervised LLMs more reliable, researchers are examining intermediate latent representations and layers within the transformer models to assess factual uncertainty (learned incorrect information) versus structural uncertainty (natural language variation).

Navid Azizan (MIT)
Hao Wang, Kristjan Greenewald (IBM)



“Our goal is to enable general-purpose LLMs to do well on specific tasks by finetuning the models on additional, task-specific data. Our work has a potential to impact many practical applications of LLMs, such as following instructions or answering questions.”

Tuning LLMs with better data subsets

Polina Golland (MIT)

Rameswar Panda (IBM)

In order for large language models to have better utility, they need to be able to follow instructions. Tuning for this specific task is preferable over more general training or finetuning with simply more data, which both require additional computational resources. Accordingly, researchers are developing a more effective method to select subsets of data to adjust the output of large language models to more accurately follow directions. The approach includes determinantal point processes and greedy algorithms to select a high quality, diverse subset from available training data. The team then iteratively assesses the subset of data constructed so far with the goal of maximizing the efficiency and effectiveness of the model’s finetuning.

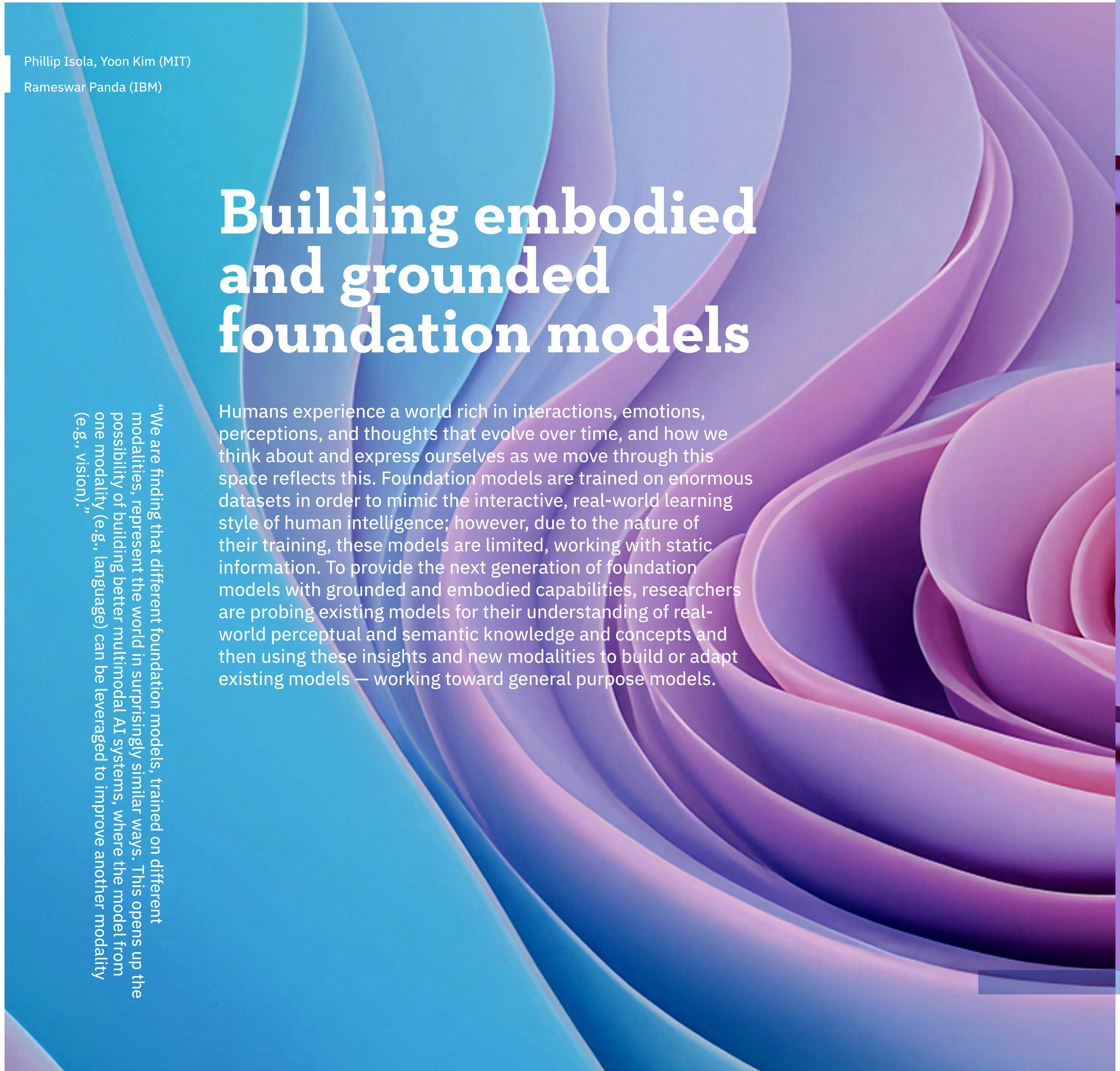
Identifying and designing small molecules for drug discovery

Increasingly, the drug discovery and development pipeline is facing challenges on top of a long timeline to market and rising costs. AI models are now assisting scientists to explore the high-dimensional search space for molecules with the desired chemical properties and molecular structures to fit into protein binding sites. In particular, researchers are augmenting IBM Research’s Biomedical Foundation Model for small molecules (BMFM/SM) with innovative conditional generation methods, by leveraging symmetry and geometry modeling, for scalability and exploiting learned molecular representations to drive property-guided generation. The development of a robust benchmarking suite can then assist researchers to evaluate these generative models; additionally, a view encoder for 3D structured data helps to improve predictive capabilities of BMFM/SM. The team is building these techniques on top of three existing models Symphony, Ophiuchus, and EquiformerV2 — laying the groundwork for scalable, targeted drug discovery.

Tess Smidt (MIT)

Partha Suryanarayanan (IBM)

“We aim to transform generative AI for early-stage drug discovery by pioneering scalable, conditional generation techniques using symmetry primitives and latent representations for precise property and binding structure-guided small molecule generation.”

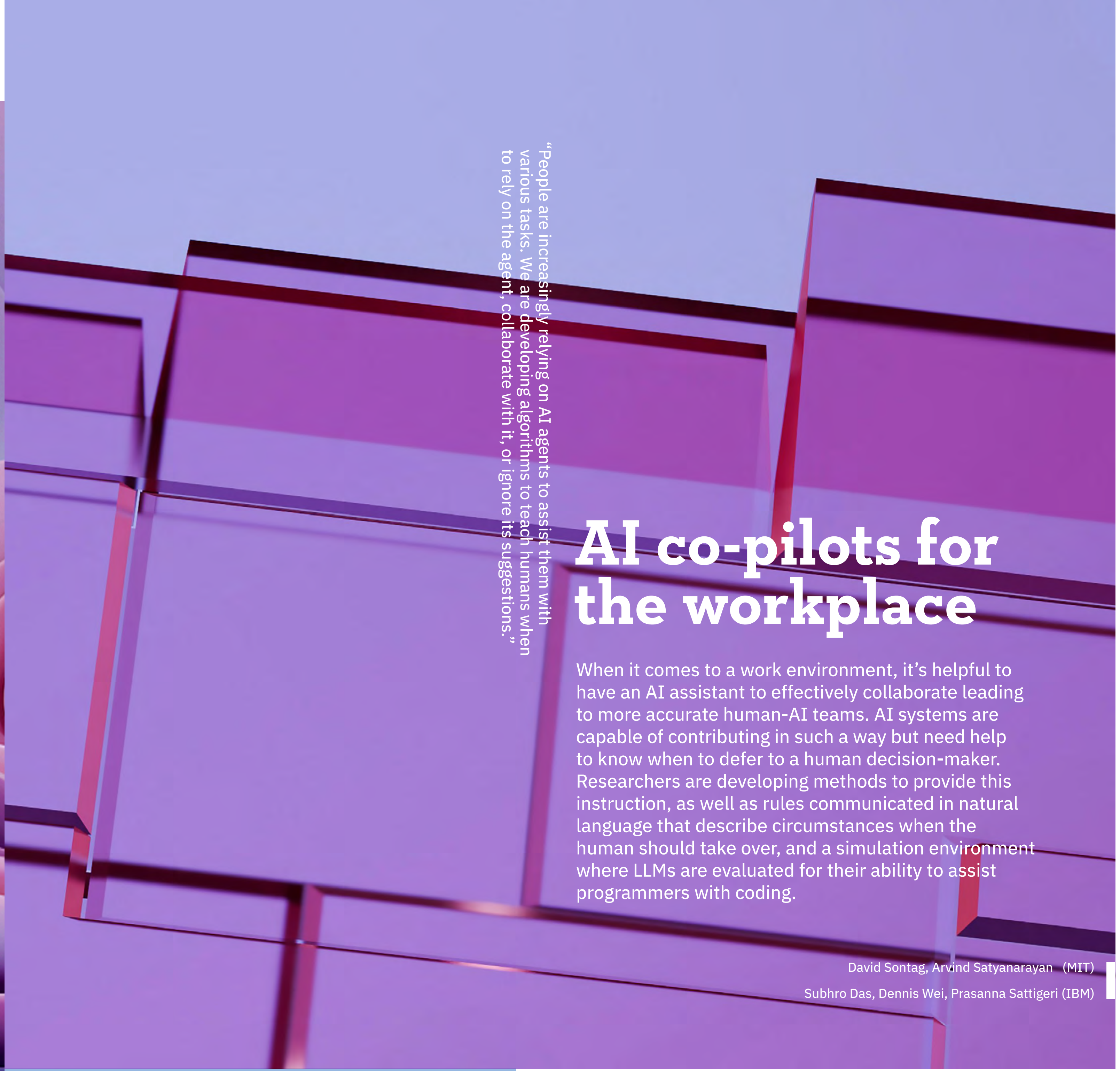


Phillip Isola, Yoon Kim (MIT)
Rameswar Panda (IBM)

Building embodied and grounded foundation models

Humans experience a world rich in interactions, emotions, perceptions, and thoughts that evolve over time, and how we think about and express ourselves as we move through this space reflects this. Foundation models are trained on enormous datasets in order to mimic the interactive, real-world learning style of human intelligence; however, due to the nature of their training, these models are limited, working with static information. To provide the next generation of foundation models with grounded and embodied capabilities, researchers are probing existing models for their understanding of real-world perceptual and semantic knowledge and concepts and then using these insights and new modalities to build or adapt existing models — working toward general purpose models.

> "We are finding that different foundation models, trained on different modalities, represent the world in surprisingly similar ways. This opens up the possibility of building better multimodal AI systems, where the model from one modality (e.g., language) can be leveraged to improve another modality (e.g., vision)."



> "People are increasingly relying on AI agents to assist them with various tasks. We are developing algorithms to teach humans when to rely on the agent, collaborate with it, or ignore its suggestions."

AI co-pilots for the workplace

When it comes to a work environment, it's helpful to have an AI assistant to effectively collaborate leading to more accurate human-AI teams. AI systems are capable of contributing in such a way but need help to know when to defer to a human decision-maker. Researchers are developing methods to provide this instruction, as well as rules communicated in natural language that describe circumstances when the human should take over, and a simulation environment where LLMs are evaluated for their ability to assist programmers with coding.

David Sontag, Arvind Satyanarayan (MIT)
Subhro Das, Dennis Wei, Prasanna Sattigeri (IBM)

Designing AI agents that can reflect, learn, and reason in open world environments

When it comes to complex decision-making, humans employ executive functions to digest and tackle problems: assessing goals and ideal outcomes, considering restrictions and safety concerns, planning and weighing options, and executing tasks to address different issues or challenges associated with it to achieve the desired result. To mimic the ability to rapidly adapt and deliberate in dynamic environments, researchers are building an integrated system with multiple modules to devise optimal solutions, as well as consider prediction accuracy verses compute time. These modules, which include LLMs or visual-language models and planning algorithms (some that are domain-specific), leverage prior and hierarchical knowledge, observe the state of a problem, and consider ideal outcomes. They then compete for developing the best path forward, with an executive module that will choose the best option, with the added capacity for human oversight.

“We aim to design and build intelligent systems that combine multiple AI techniques to address problem domains that may present a wide variety of challenges. Our systems will learn fast reactions to common situations, retain the ability to deliberate and plan when complex multi-step solutions are required, use LLMs to fill knowledge gaps when encountering new situations, and even ask a human when it is in serious doubt about what to do. Such systems will be safer and more interpretable while still providing high performance in domains such as maintaining a large factory, where an enormous variety of potential problems could arise.”

Leslie Kaelbling (MIT)

Dan Gutfreund (IBM)

James Glass (MIT)

Hilde Kuehne, Rogerio Feris, Leonid Karlinsky, Samuel Thomas (IBM)

Uniting natural language with multimodal and perception learning

The ability to understand representations and reason about objects or concepts across natural language, audio, speech, and visual modalities is inherently human; however, by fusing these capabilities, AI models are extending their utility for real-world situations. To advance this effort, researchers are developing a combination of self-supervised and instruction tuning curriculum methods to learn cross-modal correspondences between multimodal inputs and leveraging advances in LLMs to combine audio-visual perceptual models with the world knowledge, reasoning, and natural language capabilities found in LLMs.

“The emergence of AI models that can sense and perceive their environment, reason using world knowledge and communicate via natural language will have widespread applications across society, including education, healthcare, business, and entertainment.”

People

Aude Oliva

MIT Director, MIT-IBM Watson AI Lab

Director of Strategic Industry Engagement, MIT Schwarzman College of Computing

David Cox

IBM Director, MIT-IBM Watson AI Lab

Vice President for AI Models, IBM Research

Anantha Chandrakasan

MIT Chair, MIT-IBM Watson AI Lab

Chief Innovation and Strategy Officer, MIT

Dean, MIT School of Engineering

Vannevar Bush Professor of Electrical Engineering and Computer Science

Dario Gil

IBM Chair, MIT-IBM Watson AI Lab

Senior Vice President and Director of Research, IBM

Daniel Huttenlocher

MIT Co-Chair, MIT-IBM Watson AI Lab

Dean, MIT Schwarzman College of Computing

Henry Ellis Warren Professor of Electrical Engineering and Computer Science

Powering and employing large language models for scaled deployment

As titans of the AI world, language models have demonstrated their powerful abilities for numerous practical applications; however, more stands to be gained as their efficiency, accuracy, and capabilities improve, and they're incorporated into agents and systems to assist people and businesses with a greater variety of tasks. A significant portion of the MIT-IBM Watson AI Lab's research portfolio aims to bolster this growth and remove friction points through various projects, ranging from shrinking AI models and designing hardware-specific algorithms to enabling autonomous robots that can understand natural language commands and integrating multimodalities. Lab researchers are incorporating many of the resulting innovations into IBM's Granite, a series of LLMs and accompanying AI tools, to facilitate enterprise utilization at scale, as well as tailor to specific use cases. At the same time, Lab teams are leveraging IBM foundation models for geospatial observations and biomedical insights and related tools like InstructLab, for synthetic data generation and alignment. A handful of the Lab's innovative projects exemplify how our researchers are forging ahead in the language and generative AI space.

The immense computational demand of LLMs limits their application to top-tier hardware, locking out smaller businesses and edge devices, but a reduction in runtime memory usage could tip the scales. In doing so, Lab researchers Song Han and Chuang Gan are also enhancing the efficiency and cost-effectiveness to run LLMs, like those in IBM's watsonx data and AI platform, with less compute. A couple initiatives from the team have resulted in techniques like SmoothQuant and AWQ that smooth the data outliers and compress the model with their 4-bit quantization algorithm. In one case, they squeezed the memory usage by 4 and tripled the inference speed while preserving performance. When combined with the researchers' TinyChat, an efficient inference engine, the system's speed is supercharged. One such application converts Cobol, an older programming language widely used by banks and governments, into Java by way of a Granite code model, with a lower cost. In addition, another method called StreamingLLM allows a chatbot to hold onto only the initial sink tokens and recent tokens in the key-value (KV) cache rather than storing all the tokens, enabling long conversations with fixed memory.

"[Going] from a large model to a smaller model will bridge the gap between this supply versus the demand of computing. Since we can make it more efficient, we can also make it more environmentally friendly," says Han. Further, "we can deploy these large language models locally on their [businesses'] devices, which can protect the data's privacy and protect the safety of our enterprise data."

From a higher vantage point, hardware availability and utilization also concern Lab researchers Jonathan Ragan-Kelley and Rameswar Panda. "We're trying to get architectures that, at inference time, are better matched to fully exploit all the resources of the hardware that we have. But that requires not just thinking from systems perspective, but also a model architecture perspective, changing the workload that we're running," says Ragan-Kelley. Through model architecture redesign and programming language development, the team is utilizing Exo to leverage specialized hardware, like IBM's NorthPole chip for more efficient computing. One of their transformer architectures called "Intenseformer" exploits 4-bit quantization and puts "free FLOPS" (idle compute power) to work, increasing the number of calculations, while reducing data movement and the memory footprint. This addresses a common and critical memory bottleneck of transformer models during inference — the KV cache (the attention mechanism), which they also modified in another project to save some activations on a subset of layers and approximate the missing ones in later layers at inference. Consequently, pre-training speeds doubled, and the memory usage was halved, allowing them to double the context length.

"This is very important in the business context, because in watsonx, we mainly look for enterprise applications," says Panda. For other models, analyzing and summarizing SEC filings requires reading in portions of the report at a time, affecting performance; or "you can exploit the long context capability using our techniques, so that you can do processing and analysis on the whole document."

"My sense is these large language models are very powerful and useful, but to truly democratize them — to widen the scope of settings in which they can be applied — requires more efficient and more aligned models," says Lab researcher Yoon Kim. Alongside Rameswar Panda, Yikang Shen, Yang Zhang, Zaizhi Qian, and Akash Srivastava, Kim is addressing this in several ways including quantization, efficient finetuning, and recurrent architectures. By analyzing model input distributions, the team strategically adjusted the weights and activations of an LLM down to 4 bits per parameter. "We find that when you regularize these models such that their activations have certain properties, it is possible to quantize them after training to a lower precision format such that, when these models are deployed, the inference of these models is faster and also consumes less energy," says Kim, who notes that this is hardware-dependent but found that their performance almost matched that of a less-efficient 16-bit model. Other Lab work with Polina Golland, Kim, and Panda created a technique for more accurate instruction-following tasks by LLMs. The group identified and extracted a subset of diverse, high-quality data from the original training data for more efficient finetuning.

The team also eyed the attention mechanism in transformer architectures — particularly ways to reduce its mathematical complexity. Since it "performs nonlinear pairwise comparisons between all possible inputs," attention incurs complexity that is quadratic in input length, meaning that scaling to millions of tokens balloons the cost. However, sub-quadratic methods could offer savings for sequence modeling. "If you slightly modify the attention in transformers, they can be reformulated as recurrent neural networks, which have linear complexity in the input length," says Kim. To exploit this, the team developed variants of linear attention transformers, along with model training algorithms tailored to the available hardware, that rivaled transformer performance but with computational benefits.

The utility of LLMs extends beyond screens. "We have looked at how we incorporate large language models into robot task and motion planning, how we allow multiple robots to work together using large models, [and] how we use large language models as external tools," says Lab researcher Chuchu Fan.

For agents like robots to seamlessly execute tasks, they must grasp commands, interpret their surroundings, and possess an understanding of semantic relationships. "As of now, we can see that large language models are not good at solving a lot of very complicated problems, especially multi-step sequential decision-making in constraint environments," says Fan. Her Lab team with Yang Zhang decided to address this from multiple angles: They've applied LLMs to help translate natural language instructions into an intermediate temporal logic language that robots can comprehend. The group leveraged a planning algorithm with a built-in feedback loop to self-reflect on task and motion planning recommendations. Further, the team has explored frameworks for how LLM-powered agents could communicate and collaborate, finding that a hybrid centralized/decentralized system offers scalability and accuracy.

"The way large language models do inference is fundamentally different from solving all these problems," says Fan, "so, what we are doing is combining the strength of large models, but, teaching them to be able to use external tools and reason over the results given to help it solve all these problems automatically, or even help people — providing assistance much easier. That can have a lot of different use cases depending on the problem we are solving."

Events

At the MIT-IBM Watson AI Lab, we've cultivated a dynamic and deeply engaged community for students through experienced professionals that thrives on collaboration. Each touchpoint — from virtual seminars, client and research meetings, and mentorship, to networking events and competition — serves to advance our initiatives for AI progress for all, based on open-sourcing technology and open science. What follows is an overview of the Lab's ongoing contributions to the AI ecosystem.

MIT-IBM Watson AI Lab Open House

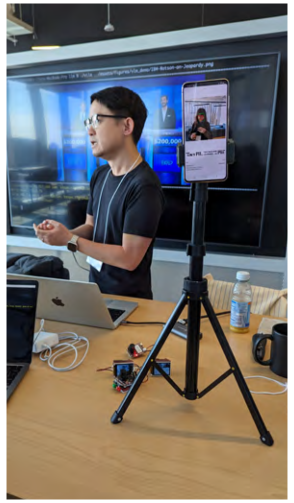
Before IBM THINK, the MIT-IBM Watson AI Lab, along with IBM Quantum and IBM Storage, invited 200 industry and academic attendees to network and explore some of the technology being developed in the Lab.

Ask the Experts Series:

To share ideas and help our corporate members to keep ahead of AI developments, MIT-IBM Watson AI Lab co-director Aude Oliva organized and moderated a series of virtual presentations and discussions throughout the year from Lab researchers on pertinent AI topics. This year, they included how machine learning and intuitive interfaces can assist with design and multi-material fabrication; how program searching, paired with compiler design and automatic GPU scheduling algorithms can lead to efficient, hardware-aware computing; how advancements in language and audio-visual models are allowing multimodal agents to perceive, understand, and communicate about their environment; and how machine learning, generative AI, graphics, and 3D geometry processing can be combined to design and fabricate materials, robots, and shapes that can be optimized for multiple objectives. These opportunities allow our members to build their knowledge base, as well as explore other avenues for AI inclusion in business solutions.

MIT-IBM Watson AI Lab Demo Day

The MIT Corporation spent an afternoon exploring AI demonstrations highlighting cutting-edge research and applications developed in the MIT-IBM Watson AI Lab. MIT students and researchers presented seven projects: Dribblebot, a soccer playing, quadrupedal robot; an algorithm for generating new molecules and predicting their properties; a generative model for designing complex mechanical systems; an audio-large language model capable of recognizing, understanding, and conversing about speech or audio it hears; language-guided manipulation of a robotic arm; the use of causal inference in business decisions; and TinyChat, an efficient model for running LLMs on an edge device.



MIT Ignite: Generative AI Entrepreneurship Competition

Over 100 teams of students and postdocs from MIT and beyond participated in the inaugural MIT Ignite: Generative AI Entrepreneurship Competition, led by the MIT-IBM Watson AI Lab and the Martin Trust Center for MIT Entrepreneurship, with support from MIT's School of Engineering and the MIT Sloan School of Management. Teams proposed generative AI-powered startup ideas addressing challenges in human health, climate change, education, and workforce dynamics. Twelve finalists received coaching before a pitching live to expert judges, who evaluated innovation, feasibility, real-world impact potential, and presentation quality.

Flagship prizes went to eMote, LeGT.ai, Sunona, UltraNeuro, and UrsaTech, with Alikorn earning the first-year undergraduate team prize. Runner-up awards were given to Autonomous Cyber, Gen EGM, Mattr AI, Neuroscreen, The Data Provenance Initiative, and Theia.

"MIT has a responsibility to help shape a future of AI innovation that is broadly beneficial," said MIT President Sally Kornbluth, emphasizing the role of young researchers in that picture. "Entrepreneurship is an essential element for our goal of organizing for positive impact."

Campus Engagement

Executing against a people-centric model for real-world impact, the MIT-IBM Watson AI Lab develops and connects drivers of change: big picture challenges and thinkers. As AI advancement and adoption into organizations — and moreover everyday life — moves at an unprecedented rate, business professionals and leaders increasingly require a solid understanding of AI to fill the growing skills gap and to provide influence in the workplace. Further, as more institutions and corporations reorganize to leverage the research, they will look to those who are comfortable with the tools and confident in the field to guide and align strategies, while helping others to upskill and embrace the transformation.

The Lab cultivates these individuals and teams through initiatives, such as internships and collaborative research projects.

Technology that began as a tremor has grown to shake up and disrupt industries and economies, making way for Lab students and young researchers to innovate and shape the future of AI and its implementation.

Some tools are better suited for a specific job — this is certainly true for AI systems. Before a large language model (LLM) can be selected from the suites of available models, stewards need to strike a balance between deployment price point and accuracy for a type of task, like understanding chemical equations versus helping with computer code debugging. Working with Lab researchers Mikhail Yurochkin and Veronika Thost, master's student Jessica Wu SB'23 is automating this process, making model selection more efficient and targeted. By designing a meta-router and companion classifiers generated with the RouterBench dataset, Wu can provide a prompt to the meta-router and produce a sequence of LLMs that will most likely offer the highest accuracy at the lowest price. Once ordered, Wu's pipeline reuses that prompt, feeding it to one LLM at a time, and based on its response, predicts the accuracy and the likelihood that that system will perform well for the prescribed task family. The process continues until a balance is reached. "Just seeing how the whole pipeline comes together — it's like a giant maze, and figuring out the little components that make a difference is [satisfying,]" says Wu.

William Brandon, a PhD student alongside Lab researchers Jonathan Ragan-Kelley and Rameswar Panda, is looking to lower the cost of frontier LLMs by developing computing architecture interventions and optimizations that improve LLM efficiency and scalability, while preserving accuracy. His team rethinks hardware utilization at training and inference to reduce overall memory usage and communication across GPUs to engage parallel computing as well as remove memory bottlenecks. Taking a low-level systems engineering perspective targeted at transformers, Brandon's team has developed a technique that overlaps computation and data communication, unlocking and squeezing hardware capacity where possible. Other work attacks the KV cache, minimizing duplicated information and lengthening chat sequences.

"In addition to staying grounded in the system, the algorithmic, and quality considerations, you also, to some extent, need to keep an eye on the business and the economic considerations of what are the different regimes in which you could possibly deploy these models and how the tradeoffs that you'd want to make in those settings differ," says Brandon. "Can you carve out a niche for a particular technique, where you can articulate in this setting that this [prescribed method] will deliver big wins?"

In a domain where data rapidly turns over, like finance, techniques that can accurately and quickly parse through heterogenous, temporal information in an explainable and privacy-preserving manner benefit institutions and governing bodies. Collaborating with Lab researchers Yada Zhu and Julian Shun, MEng student Samuel Mitchell is accelerating how fraud is identified within financial transaction graph data. Using synthetic data, Mitchell explores nearest neighbor approaches to speed up and scale TG Editor, a graph editing algorithm to improve embedding, correct missing data, and remove noisy data, in such a way that it can scale to tens of millions of edges without latency. By evaluating inductive biases in previous literature, Mitchell says, he discovered that, a dot product for the distance metric would better capture the multitude of node types, helping to illuminate fraudulent relationships across the graph faster. "Fraudsters are quick to adapt; if they have a method that no longer works, they quickly change," says Mitchell. This innovation helps to reliably discern patterns and players in evolving fraud schemes.

Instead of connections between people, PhD student Jessy Han pursues causal relationships in the criminal justice system, as well as in areas of social and life sciences. With Lab researchers Devavrat Shah and Kristjan Greenewald as mentors, Han reappropriated a survival analysis framework from healthcare to quantitatively analyze how racial disparities and other interventions, like educational programs, may influence the likelihood and timing of recidivism of individuals after release from incarceration. She's paying particular attention to the effect of COMPAS scores assigned to individuals, generated by a risk assessment algorithm for re-offending — one that is suggested to be biased but widely used. Han points out that there are frequently many judgements and actions taken by individuals, bystanders, and law enforcement before a person is arrested, but current frameworks portray what is a multi-step process as a single timestamp and recidivism as a binary quantification.

Han says that we can make qualitative and circumstantial arguments that racial bias is occurring, but "we need a quantitative framework to understand what exactly is 'racial bias'. How do you define it? Can you find evidence from real world data saying that there is, indeed, this amount of racial bias?" With this information, policymakers can implement actionable change.

In order to make a measurable impact, it's "really important to understand the domain ... when we want to apply a methodology to a real-world scenario because we really want to capture the mechanism and/or how the system actually works," says Han. "You have to understand how things work in the real world, and then you have to abstract it."

In some cases, the MIT-IBM Watson AI Lab has the opportunity to make a more extensive mark on the student experience when “alumni” of the Lab return for another internship or even join the research cohort at the Lab or an IBM group after graduation. In such a position, these individuals bring invaluable knowhow to the table, not only to build on their previous work and/or pursue exciting new research directions but also to, potentially, remain on Lab projects as collaborators and help guide students who now stand where they once stood. Seungwook Han, Ji Lin PhD ’23, Ching-Yun (Irene) Ko PhD ’24, and Maohao Shen SM ’23 are some of those that can share in these insights.

Seungwook Han entered the Lab in a less conventional way — from the industry side with an interest in using elements from neuroscience to enhance AI systems. As an AI Research Engineer at IBM working with the Lab, Han began by focusing on energy-efficient generative adversarial neural networks. Here, he started by developing a two-step generative model for high-dimensional image creation, matching the performance of state-of-the-art models while reducing training time and compute requirements. “Even though it’s an industry lab, it felt like an academic lab in the sense that it’s very small, very organic in terms of research interest and pursuits,” says Han, who, at the time, was considering doctoral studies but grappling with choosing a research direction. Ultimately, he landed on building foundational algorithms that can scale. While still at IBM, he and Lab researcher Akash Srivastava engaged in other visual projects, one of which was self-supervised contrastive learning. Together, they created a framework for equivariances, so that image representations change with image transformations, perturbations, and orientation — work that would benefit future vision foundation models on downstream tasks like classification and segmentation.

On the academic MIT side, interning with Lab researchers Pulkit Agrawal and Srivastava, Han now investigates large language models (LLMs) with projects like HIP (compositional foundation models for hierarchical planning), where he’s imbuing a vision-language foundation model with multi-level task planning abstraction to assist in long-horizon jobs carried out by robots. Other pursuits included reward methods to align and personalize foundation language models to the values and preferences of individuals, and providing new domain knowledge to LLMs at fine-tuning time while mitigating hallucinations.

“I wanted to go through the whole process cycle of refining what the important [research] questions are for the future and try to tackle them one by one — maybe not the whole grand question — with my own set of solutions,” says Han.

Maohao Shen has been seeing through one research direction with the Lab over the course of his master’s and now PhD work: improving the reliability and trustworthiness of machine learning models, including classical and foundation models, through uncertainty quantification. During Shen’s master’s projects, he worked with the teams of Lab researchers Subhro Das, Prasanna Sattigeri, and Gregory Wornell to improve the uncertainty quantification capabilities of both classical models and LLMs. In one case, they developed a meta-model that, when injected into a classical image classification model, can help predict and detect misclassification.

As a PhD interning with the Lab again, he developed a universal calibration method that adjusts an LLM’s confidence level for response alignment. This method has also been used in IBM’s foundation model development. Additionally, beyond uncertainty quantification and calibration, Shen and his team are exploring novel approaches to improve LLMs’s reasoning capabilities through self-correct mechanisms.

Shen expressed how, through the Lab, he was exposed to the “frontline” of AI research directions. “The meaningful research is always looking ahead. I believe it’s important to first consider how your work will make a large impact and then dive into addressing the relevant research problems and technical details.”

Early on in his graduate studies, Ji Lin was introduced to the Lab through researchers Song Han and Chuang Gan, who were shrinking vision-based machine learning models to run on microcontrollers, smartphones, and other edge devices. They addressed the tinyML issue through a lens of system-algorithm co-design with autoML-based methods to create neural

architectures and optimized scheduling for different hardware constraints. Lin’s projects, like MCUNet, included focusing on inference, as well as on-device training for model customization that are compiler-centric to reduce memory usage and improve efficiency.

Lin’s group found that this work naturally lent itself to the world of LLMs and quantization. “In the large language model area, we’re actually facing exactly the same bottleneck compared to the tinyML area. Basically, they’re both bounded by memory, but not by the compute,” says Lin. “We find actually a lot of the work we did in the tinyML area can directly transfer, and the impact on the large language model side could be quite big.” In the pivot, one of their works, SmoothQuant, helped to smooth out outlier values, nearly doubling the model’s speed. In addition to other weight and activation quantization methods, their work enabled LLMs in the 10-billion parameter range to be deployed locally.

Naturally, Lin found himself in an industry research setting after graduation, applying insights gained from LLM work. “I think my biggest improvement is finding the right problem to solve, in terms of research,” says Lin. “What is the critical question that better fits your skills at this moment, that is worth your time to solve. Secondary, it might be my overall management of the research projects; it includes both research skills and also managing collaboration with teammates.”

Before Ching-Yun (Irene) Ko graduated and joined the Thomas J. Watson Research Center in Yorktown Heights, NY, she interned with the Lab and then with IBM two more times, investigating the interplay between robustness, fairness, and accuracy of neural networks. In order for foundation models to be trustworthy, researchers like Ko need to understand the boundaries where classification breaks down, potentially causing harm, and where they are certain of the response accuracy. Ko began her work with deterministic approaches for establishing the certified radius in recurrent neural networks in collaboration with Lab researchers Luca Daniel, Sijia Liu, and Pin-Yu Chen. She further extended her studies to probabilistic methods, such as randomized smoothing techniques and uncovered a significant lost in fairness amongst classes.

In another project spanning the Lab and the Thomas J. Watson Research Center in Yorktown Heights, NY, Ko’s team developed an evaluation framework for foundation models: SynBench for vision models and SynTextBench, for language models to measure the quality of pretrained representations using synthetic images and sentences, in terms of accuracy and robustness tradeoffs without relying on specific downstream tasks. On the whole, Ko’s work has helped to identify and address potential model vulnerabilities, which touch applications ranging from computer vision to natural language.

As someone who sees herself as highly adaptable, Ko felt that at the Lab and during her PhD, she was in a place to explore other novel ideas as necessary. “I don’t restrict myself to one specific topic throughout my research ... I let my thoughts pave the way for my research.”

In the Media

The MIT-IBM Watson AI Lab and its researchers have been cited and featured across media outlets, including *Bloomberg, CNN, Marketplace, New Scientist, Popular Mechanics*, and *The Economist.*

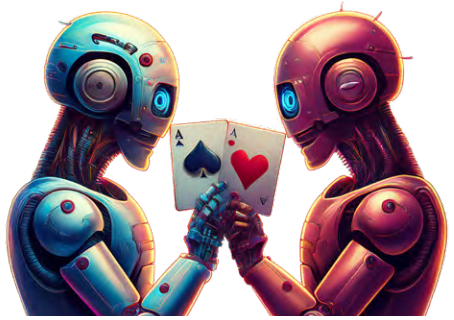
Interesting Engineering

MIT's new powerful chip thwarts millions of data theft attacks in tests

Health-monitoring apps, like health and fitness trackers, collect sensitive information to help individuals work toward fitness goals or manage chronic conditions, but they also impede the phone's general performance. This is because data is moved frequently between a central memory server and the edge device; machine-learning accelerators are employed to assist here but are susceptible to data attacks. MIT-IBM Watson AI Lab researchers have developed a machine-learning accelerator with digital in-memory compute that can help protect against side-channel attacks and bus-probing attacks.

Quanta Magazine

Game theory can make AI more correct and efficient



Some large language models might provide different answers based on how a query is posed. In order to help the model be more consistent and reliable, MIT-IBM Watson AI Lab researchers have developed a consensus game based on Diplomacy. The method uses a reward system, with an answer generator and discriminator, to align the responses and make them more accurate, regardless of how a question is asked of the model.

New Scientist

AI costs too much to automate vision-related jobs – for now

People have long speculated about AI taking over jobs, but it wasn't until recently that the concern became a reality. Work from MIT-IBM Watson AI Lab researchers examined the economic costs of such an overhaul across many US industries and sectors, when it came to computer vision tasks. They found that, while over 400 tasks could be at risk for automation by AI, most are not cost-effective yet to do so, but as the technology improves and costs come down, it is likely that we'll see more implementation.

IEEE Spectrum

MIT and IBM find clever AI ways around brute-force math

Partial differential equations (PDEs) help to elegantly explain the world around us, like natural phenomena and physical systems — from the waves in the ocean to quantum mechanics and more. Running models and solving them with numerical methods requires time and computational power. MIT-IBM Watson AI Lab researchers have developed a "physics-enhanced deep surrogate" (PEDS) model that uses a physics simulator with a neural network generator, which is then trained to match the performance of numerical solvers. The technique affords significant savings in computation and data, with the added benefit of explainability.

Hackster.io

See what you mean

Videos provide a wealth of information; however, searching for a particular detail or action within them can be tedious. To address this need, MIT-IBM Watson AI Lab researchers developed a self-supervised AI method that harnesses raw video and auto-generated transcripts, allowing users to pinpoint moments of interest through simple text prompts. Their spatio-temporal grounding approach deciphers when and where key actions occur, creating a bridge between human language and video comprehension that could assist in fields ranging from medicine to education.

Decrypt

Scientists uncover biological echoes in powerful AI transformer models

Work from the MIT-IBM Watson AI Lab explored the parallels between a popular transformer architecture and biological neurons. A team of researchers investigated and appreciated the role of the brain's astrocytes and neuron–astrocyte networks — how they communicate and process inputs. Further, the researchers examined signal processing over time and how this could allow for self-attention, a key behavior of transformers.

MarkTechPost

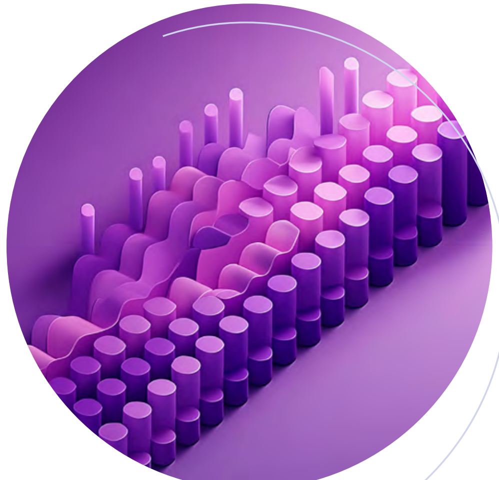
This AI paper introduces the scientific generative agent: A unified machine learning framework for cross-disciplinary scientific discovery

Scientific exploration and discovery require domain-specific and general area knowledge of tools, data, and procedures, as well as creativity and the ability to generate, test, and analyze a hypothesis. To assist scientists in this work, MIT-IBM Watson AI Lab researchers have developed a framework that combines the knowledge and reasoning abilities of a large language model with observations from model simulation experiments. The method creates a hypothesis and optimizes variables to help discover constitutive relationships and design new molecules, accelerating research capabilities.

MarkTechPost

QoQ and QServe: A new frontier in model quantization transforming large language model deployment

On account of their size, deploying large language models requires efficient handling of data and computation; quantization helps to facilitate this, particularly at inference. Previous methods often lead to further computation and accuracy losses. To accelerate computation while maintaining accuracy, MIT-IBM Watson AI Lab researchers developed a framework — the key innovations include progressive group quantization, adjustment of activation keys, as well as compute-aware weight reordering and fused attention mechanisms. Together, the algorithm and system significantly increase throughput and reduce costs, enabling faster and more efficient AI processing for real-world applications.



Case Studies

Data is no longer just a byproduct of operations; it's a core asset. By analyzing macro- and microscale patterns in industry trends, supply chain efficiency, customer behavior, and more, the MIT-IBM Watson AI Lab and its corporate members gain insights that allow our industry colleagues to anticipate changes and make assessments proactively. AI amplifies this power by processing vast amounts of information at speeds that wouldn't otherwise be possible — identifying inefficiencies, anomalies, disruptions, and actionable items in near real-time. The resulting AI tools and strategies created equip our members to make more informed and effective decisions, reduce risk, and seize opportunities. Embracing this capability means not just keeping up — we're leading.

Precision blood vessel measurement for stent sizing and placement

During interventional medicine, a fraction of a millimeter can sometimes make or break a procedure, like treatments of cardiovascular conditions i.e. coronary atherosclerosis or deep vein thrombosis that affect millions of Americans today. A popular one is stenting in which specialists reopen a narrowed or blocked blood vessel. The sizing and placement of these are crucial to the intervention's success as under-sizing can lead to thrombosis while oversizing can cause vessel tears and other injuries. Currently, intravascular ultrasound (IVUS) imaging is used to visualize and evaluate the inner wall of the vein or artery to locate affected regions in real time, and to measure lumen diameter to assess stent sizing. Until now, most of this characterization and evaluation has been done qualitatively through visual inspection during IVUS, which can be inconsistent and imprecise, affecting outcomes. To address the need, the Lab developed AI methods using hybrid convolutional neural network architectures that combine geometric reasoning, spatial priors, and temporal modeling with the domain knowledge of stent diameters in order to more accurately and precisely segment the lumen boundaries for executing image-guided interventions by these specialists, thus showing AI can be helpful in such high precision tasks as well.

Key Metrics:

- Achieved a major and minor axis diameter error within 0.25/0.5/0.75mm for 66/84/90% of all normal frames
- 200% improvement in performance for normal lumen segmentation over state-of-the-art methods

Better geospatial forecasting for renewable energy generation

From sea surface temperatures and topography to vegetation cover and air pressure — combined, diverse spatiotemporal data paint a picture of our complex Earth system. Geospatial foundation models are trained on troves of Earth system data and provide a highly versatile base to address intricate scientific problems and expedite broader deployment of AI models across diverse Earth system applications. Leveraging this paradigm, the Lab sought to enhance IBM and NASA's geospatial foundation model for local wind and solar energy generation forecasts by grounding the model in more data modalities. The Lab developed a message-passing graph neural network that uses multimodal and historical weather data to forecast near-surface wind dynamics at off-grid locations. The technique addresses the limitations of traditional gridded weather models, which often miss local variations due to topography, buildings, vegetation and other factors. It also captures spatial and temporal nuances in weather patterns and excels in coastal and inland regions with complex topography.

Key Metrics:

- 92% error reduction compared to ERA5, a traditional gridded weather model
- Greater than 50% error reduction versus a multi-layer perceptron (MLP) combined with ERA5

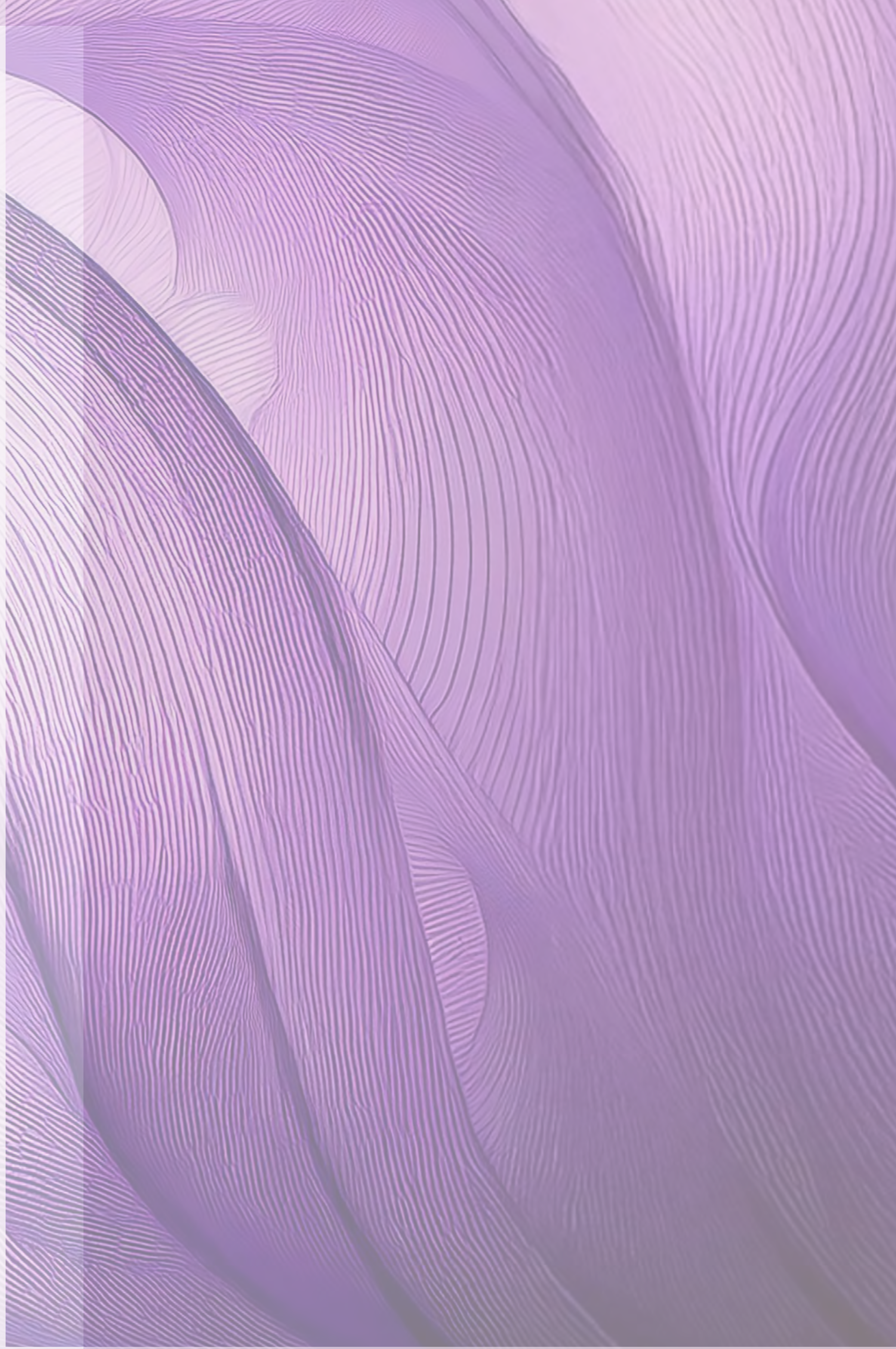
Aligning learning with downstream tasks for revenue and resource forecasting and optimization

Machine learning can enhance data-driven solutions by considering its impact on optimization through end-to-end learning. This approach, integrating prediction and optimization, often surpasses traditional methods focused on prediction accuracy. One application is fine-grain revenue forecasting, which supports trend identification, efficient resource allocation, and strategic planning for business operations. Using this, the Lab developed a hierarchal time-series forecasting method, reducing manual effort and enhancing the neural network's output coherency. The Lab also looked at uncertainties, independent and dependent on downstream decisions, as seen in inventory and revenue optimization. The team designed a meta-optimization method for exogenous uncertainty that learns a surrogate function to approximate the optimization objective, drastically reducing computational time during learning, as well as a learning method for endogenous uncertainty that incorporated domain knowledge and structural constraints for simpler, interpretable models.

Key Metrics:

- 2-4% lower test RMSE on large datasets and 10-40% improvement on smaller datasets compared to hierarchical methods; 3-percentage point improvement in revenue forecasts with member data
- 2-10x faster solutions for stochastic optimization problems with no accuracy loss; over 7% performance improvement with simpler models compared to traditional two-step methods

Boston Scientific Advancing science for life™ EVONIK Leading Beyond Chemistry NEXPLORE EXPANDING HORIZONS WELLS FARGO woodside



MIT-IBM Watson AI Lab

MIT School of Engineering

MIT Schwarzman College of Computing

IBM Research