PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-zado-02wn **Comments Due:** May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0291
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Rice University, The Ken Kennedy Institute

General Comment

See attached file(s)

Attachments

Rice University Ken Kennedy Institute AI RnD Strategic Plan

A Response to the Request for Information on the Development of an Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan

The Ken Kennedy Institute
Rice University[1]

May 29, 2025

1. Overview

Recent, rapid progress has evolved Artificial Intelligence (AI) into a form that has begun to approach the lofty visions that the earliest AI proponents imagined more than half a century ago. After decades of development operating largely out of the public eye, there has been a sea change in both the capability and accessibility of AI that has thrust its capabilities (and its limitations) into the spotlight. There is now a public perception, correct or not, that all-purpose, autonomous systems might be just around the corner, ready to assist us in all aspects of our lives. A National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan must direct the current momentum and excitement surrounding AI and channel it along directions that maximize its potential benefits.

The demand for AI comes from many directions, as we all explore ways in which AI can improve our lives. Not surprisingly, we see a wide range of stakeholders rushing to fill the increased demand and promises for AI. Although the resulting large-scale, industry-wide investment has led to the most visible AI successes, notable gaps remain in the progress made so far. Filling those gaps requires a coordinated national investment that specifically prioritizes areas where progress is both impactful and feasible. We put forward the following topic areas as deserving of the highest priority in the AI R&D Strategic Plan being developed:

Energy Efficiency: Much of the excitement about AI has come from the success of large-scale AI models, but at the price of enormous increases in energy consumption needed to train and operate those models. Research in resource efficiency (both in software and hardware) is needed to keep pace with the ever-increasing scale and breadth of future AI applications (Section 2).

Assurance of AI Outputs: Excitement about AI is tempered by the mistakes that we see current systems make, which naturally introduce reluctance to deploy such systems in more critical applications. Providing systematic methods to measure

[1] This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI R&D Strategic Plan and associated documents without attribution.

and ensure the correctness and usefulness of AI models, while minimizing large and even existential risks, is needed to give us confidence in expanding the scope of what tasks we delegate to them (Section 3).

Human-AI Teaming: Another way to mitigate the risk of errors by AI systems is to team them with people whose expertise can cover for potential knowledge gaps in those systems. In such teams, AI systems can be a force multiplier for human skills, whether in the form of software systems for decision making or embodied systems for performing physical tasks (Section 4).

AI for Education, Science & Engineering, and Health: Research that improves core AI capabilities can theoretically benefit all aspects of our lives. However, research in AI for education, science & engineering, and health should be special priorities, because improvements in these sectors will maximize the impact of that research by benefiting our entire population (Section 5).

Although each of these areas are all important individually, we also prioritize them together as an interrelated system. For example, using AI to improve our education system will prepare our future workforce to understand and use AI systems better, making people more productive partners in human-AI teams. Having a human in the loop when training an AI system provides a mechanism for enforcing appropriate norms and preferences within the resulting model, leading to tighter assurances on its output. Such an assurance mechanism gives a more accurate picture of the cost-benefit tradeoff in energy consumption, allowing us to focus cost-reducing efforts where they have minimal impact on model accuracy. Improvements in energy efficiency can lead to smaller models that are both more efficient and more understandable, allowing AI systems to serve as more responsive, transparent teammates to people. This virtuous cycle illustrates the synergies among these areas that can amplify the investments that we make in each of them.

We prioritize these areas not just because of their importance, but also because making rapid progress in all of them is quite achievable. The following sections illustrate some of the progress we ourselves have made in each of these areas. At the same time, we are not the only ones pursuing these directions, as vividly demonstrated by the emergence of DeepSeek and the surprised reactions from the wider public that greeted it. We can expect similar surprises if we do not take leadership in the areas listed above. Fortunately, there is a golden opportunity to take leadership through an AI R&D Strategic Plan that prioritizes the funding and the talent pipeline needed for fundamental research in these four areas.

2. Energy Efficiency

The increasing energy consumption requirements of the latest generation of AI models are well publicized, as they have become a critical bottleneck in development and deployment. Recent estimates indicate that training a large language model (LLM) can consume energy equivalent to the lifetime emissions of five cars. While industry continues investing in data centers with enhanced power and cooling capabilities, physical and environmental constraints inevitably cap computational scalability. Therefore, reducing the computational cost of AI models is paramount for expanding

their training and operational scale.

Furthermore, increasing the energy efficiency of AI algorithms brings additional benefits that could be even more impactful in the long run. While having large-scale AI models available through the cloud makes those models widely accessible, that same cloud-based access limits their usefulness for domains that require local computation. For example, when we deploy systems that must sense, learn, and act in remote or hazardous environments, high-throughput wireless communication and computation are luxuries we cannot always afford. Furthermore, even in domains with no resource barrier to communicating with the cloud, there can still be a desire to keep the computation and (especially) data local. For example, in medical domains, privacy safeguards will limit what data (if any) can be transmitted. Smaller and more efficient algorithms can operate in these currently underexplored domains, without the data ever leaving the room.

Even in those domains where the end users are willing and able to make use of cloud services, the computational cost of model training discourages them from specializing those models for their particular domain, not to mention updating those models over time. When using off-the-shelf AI models is not an option, resource efficiency is critical to put model training (and re-training) directly within the grasp of specialists with knowledge of the domain. We identify four distinct approaches to energy efficiency in AI systems, each operating at different levels of the AI development stack and offering unique trade-offs between implementation complexity and potential energy savings.

One approach is to use hardware-aware algorithmic adaptations that work within existing training protocols while leveraging hardware-specific optimizations [Liu et al., 2023]. For example, workload-aware power management adjusts hardware power states based on the computational needs of different model components. These types of methods are particularly valuable as they can be applied post-hoc to existing models with minimal retraining. Standardization of such hardware-aware optimization techniques can ensure widespread adoption across different platforms and architectures.

Parameter-efficient methods minimize the number of distinct parameters to be learned, reducing both the amount of training needed and the size of the learned models. For example, *parameter reduction* carefully chooses lower-dimensional approximations of parts of an AI model to produce smaller, faster models while still achieving comparable performance [Celaya et al., 2022]. *Parameter sharing*, in contrast, does not prune parameters away through approximation, but instead reuses the same parameters in multiple places across the AI model and has outperformed parameter reduction in certain cases [Desai and Shrivastava, 2024].

Decentralized training protocols fundamentally alter how models are trained. For example, *federated learning* approaches naturally support distributed training across resource-constrained devices, as each node needs to handle only a fraction of the full model capacity [Kim et al., 2023, Hu et al., 2023]. In addition to the greater efficiency, secure protocols for decentralized machine learning have the additional advantages of being able to operate across heterogeneous devices while also preserving privacy.

Compositional approaches challenge the monolithic model paradigm by combining specialized models through mixture-of-experts or agent-based architectures [Dun et al., 2023]. Rather than training a single large model, this approach routes inputs to appropriate expert models, allowing for efficient resource utilization as only relevant experts are activated for each input. For example, a document analysis system might

use different AI experts for layout analysis, text recognition, and semantic understanding. The AI R&D StrategicPlan can encourage modular development and reuse of specialized models by establishing benchmarks and interfaces for compositional AI systems.

These priorities should be supported by investment in both basic research to develop novel efficiency techniques and applied research to translate theoretical advances into practical implementations. Success in these areas will not only reduce the energy costs of AI systems but also enable their deployment in resource-constrained environments where cloud computing is not feasible or desirable.

3. Assurance of AI Outputs

Regardless of how efficient an algorithm may be, it will not be useful if it does not lead to accurate AI models. Unfortunately, predicting the accuracy of an AI model must account for the complexity of modern machine-learning algorithms, the vast amounts of data fed into them to train AI models, and the breadth of contexts to which the resulting models will be applied. Current AI pipelines are very good at providing output that seems plausible, but whose flaws are easily visible to domain experts. While such “good enough” output may be sufficient in non-critical applications, we need stronger assurances before deploying AI models to assist in decision-making scenarios with weightier consequences. Fortunately, there are many alternative paths to such assurances that model builders can choose from, allowing them to instill confidence in their end users without being overly restrictive during model development.

One such path is the use of multimodal data, where datasets of different forms (e.g., language and images) are used simultaneously for training and execution of AI models. While the amount of data being ingested and the size of the resulting model grow with each additional modality, the interdependencies between the modalities induce constraints on what the model can learn and infer. For example, when using AI to generate a text description of a visual data stream, we can use the image content to ground the language generation in a way that reduces hallucinations [Xiao et al., 2024]. This synergy also works in the opposite direction, as generating images from text can similarly improve by leveraging other input streams (e.g., additional text, edge maps) [He et al., 2024].

Providing assurance on the performance of an AI model usually relies on some assumption that future conditions will resemble the past data used to train that model. That assumption breaks down in highly dynamic environments. In such environments, we instead want our AI models to be able to assimilate new data, thus being able to continually learn as conditions change, rather than being stuck with out-of-date reasoning [Wolfe and Kyrillidis, 2022].

It is not just the combination of data with more data that leads to more robust AI models: combining data with scientific models can provide similar benefits. The natural sciences have distilled centuries' worth of data into mathematical models, and AI models should exploit that knowledge whenever possible. For example, we can use validated mathematical models of physical systems to generate “data” for training AI models and to infuse physical laws into loss functions [Celaya et al., 2024].

Progress in these directions would significantly advance our ability to deploy reliable AI systems in practice. The AI R&D Strategic Plan should prioritize research that

addresses these critical challenges in scaling AI capabilities while maintaining reliability. This research should emphasize both theoretical foundations for understanding modeling errors and practical techniques for building robust systems. This combination of theory and systems would enable new applications of AI in domains where reliability is paramount, such as healthcare, financial systems, and critical infrastructure management.

4. Human-AI Teaming

One way to mitigate the risk of deploying an AI model that we know will make mistakes is to also give it a human teammate. By having people and AI systems working together on a task, we can exploit both human expertise and AI's algorithmic (and, in the case of embodied systems, physical) strengths, with each hopefully covering for the weaknesses of the other. Just like a good team of people, a good team of AIs and people can perform tasks faster, more reliably, and at greater scale than any individual team member could on their own.

For any team, it is not enough for individual team members to be good at their tasks; they must also be good at working together. In a human-AI team, working well together requires that the human members understand enough of their AI teammates' operation to make good decisions themselves. As a result, transparency of AI models is a priority. Progress has been made on making AI models *explainable* by automatically generating descriptions of their content and operation in terms that are more understandable by a human teammate [Linder et al., 2021, Yang et al., 2021]. Furthermore, more recent work has begun to account for the individual differences and dynamic states of human users to deliver explanations of an AI model's behavior based on their expertise [Rong et al., 2024] or their current understanding of the model [Qian et al., 2024]. The AI R&D Strategic Plan should prioritize explainability that is similarly personalized and dynamic, as these methods greatly enhance the transparency of AI models, which in turn will improve the decisions made by the humans working with those models.

Today there is also great excitement for embodied AI systems (robots) that can carry out tasks specified by humans, collaborate with humans, and physically assist in manufacturing, fulfillment, transportation, and, in the long run, everyday tasks. Robotics provides a tangible platform where abstract algorithms and data-driven insights are tested in real-world scenarios. Through the direct interaction of AI with physical objects and complex environments, AI models will be refined to handle unpredictability, adapt to changing conditions, and reason about complex real-world tasks. Advances in AI allow the use of natural language to instruct and collaborate with such systems, overcoming one of the current barriers in task specification [Guo et al., 2025]. Work in this area combines perception [Bhowmick et al., 2020] with advanced motion planning and reasoning capabilities for long-horizon missions [Pan et al., 2024]. By pushing the limits of what intelligent machines can sense and do, robotics not only strengthens existing AI methods but also inspires new approaches to adaptive collaborative systems, driving us toward a future where humans and machines work together to achieve unprecedented breakthroughs.

5. AI for Education, Science & Engineering, and Health

AI has already made substantive contributions to education, science and engineering, and health, and our entire population has benefited from those contributions. Continuing to direct AI research toward these sectors will maximize the downstream positive impact of progress made at the algorithmic foundations of AI, in addition to fueling progress in the respective sectors.

5.1 AI for Education

Education is a key vehicle for improving human flourishing (i.e., enhancing individuals' potential and well-being) in a supportive environment that enables students to cultivate and maintain meaningful relationships and activities [Singh et al., 2025]. The use of AI in education has been prevalent for over two decades for use cases such as automated writing evaluations [Shermis & Burstein, 2013]. Presently, with the mainstream availability of Generative AI (GenAI), with its improved capabilities for personalized and dynamic interactions, these technologies stand to have a transformative impact on education when used responsibly.

The rapid pace at which AI is evolving is a stark contrast from the pace at which changes typically affect educational systems [Basu Mallick et al., 2025]. Yet, AI more broadly, and GenAI specifically, have the potential to support educators in identifying and providing individual and timely attention to students who need help. There is rich literature in the field of AI for education, including the impact of intelligent tutoring systems [Aleven et al., 2023] that combine models from learning sciences, cognitive sciences, and education research to make personalized, on-demand tutoring available to students while using a variety of active learning and assessment strategies. GenAI is particularly useful for generating instructional content, including assessments [Wang et al., 2022], as well as multimodal materials, as illustrated by Duolingo's learning application, Google's Gemini, and OpenAI's GPT 4o.

With high-stakes applications like education, a human-centered approach and appropriate safeguards that help mitigate some of the known limitations of GenAI models are essential. Some of these limitations include: LLM hallucinations that produce incorrect information [Ji et al., 2023]; privacy vulnerabilities due to adversarial attacks that could reveal private information about learners and educators [Bender et al., 2021; Carlini et al., 2021; Zou et al., 2023]; mathematical, logical, and reasoning errors limiting their educational utility in certain domains [Hendrycks et al., 2021; Lu et al., 2022]; and challenges with algorithmic fairness that affect learning experiences [Belzak et al., 2023; Johnson & Brun, 2022].

Crucial to supporting the effective use of AI systems in education is a deep understanding of the individual needs of the students and educators, the content students are learning, and the instructional design strategies used by educators [McNamara et al., 2022]. Attaining this deep understanding requires significant data about the learning process, context, trajectory, and individual differences, with strict data protection protocols in place. R&D infrastructures like SafeInsights, with their nontraditional privacy-first approach, create opportunities not only for evaluating which

AI tools and models are working well for educational settings, but also for developing models for the educational context.

Additionally, the privacy-first approach facilitates further algorithmic development in areas such as synthetic data generation (i.e., data that are algorithmically generated to represent the properties of the real underlying dataset). Synthetic data are of value in education and other domains where there is an abundance of protected data (e.g., healthcare) [Liu et al., 2024]. Currently, synthetic data are being used for generating synthetic student records [Khalil et al., 2025] , to improve the reproducibility of research [Grund et al., 2022]. As data become more complex and multimodal, synthetic data have immense value when data are limited and/or private [Qian et al., 2023]. With the growth of multimodal systems (e.g., student-teacher dialog systems, student handwriting) [Ochoa, 2022], there is increasing interest in developing algorithms to generate multimodal synthetic data, which is a challenging problem given their high-dimensional nature.

5.2 AI for Science & Engineering

Research investments into AI and ML will also trigger fundamental advances outside of computer science. For example, research into the mathematical underpinnings of ML must address several open mathematics problems in ML — the interpretability of neural networks, optimization problems in parameter estimation, and the generalization ability of learning models — and progress in solving these problems continually fuels new mathematical theories. In the natural sciences, AI has begun showing its great potential for generating novel hypotheses in physics and chemistry. Combining this potential with pre-existing scientific theories and models (as described in Section 3) will lead to the discovery and design of new materials with targeted properties for broad applications in energy, biomedicine, construction, transportation, national security, spaceflight, and other domains. To offer a concrete example from engineering, combining physics knowledge and data into an AI model for flood prediction can lead to orders-of-magnitude speedups in predictions when compared against comparable traditional simulations [Kazadi et al., 2024b, Kazadi et al., 2024a], paving the way for not only faster but also more accurate simulations to inform critical decisions for resilient urban infrastructure.

5.3 AI for Health

AI has shown similar promise for generating novel analyses, discoveries, and designs when addressing medical challenges. AI's ability to unify theoretical models with large amounts of data makes it a powerful tool for amplifying the research and clinical expertise that medical practitioners have. As a result, we have already begun to see payoffs from AI applications across a wide range of healthcare challenges: omics and big data analysis [Yao et al., 2018]; understanding the origins of disease, particularly in mental health, autoimmune conditions, rare diseases, chronic or degenera tive disease [Roussarie et al., 2020], and cancer [Mallory et al., 2020, Edrisi et al., 2023]; accelerated drug discovery by analyzing molecular interactions [Fasoulis et al., 2024, Conev et al., 2023]; early detection of pathogens [Wu et al., 2025, Balaji et al., 2022]. The AI R&D Strategic Plan should maintain the momentum of these preliminary successes to achieve

AI's full potential for population-wide improvement in health outcomes.

5.4 Educating the Workforce about AI

Although it is critical to train AI systems to benefit people in domains such as employment, education, and health,, there also needs to be improvement in the other direction: training people on how to work effectively with AI. The benefit of human-AI teaming hinges on the ability to find people who are knowledgeable about both the application domain and the AI system they are working with. This AI knowledge can take on a variety of forms, such as knowledge of algorithms and the data to which they are applied, interpreting output from explainable AI systems, and learning interactions with AI that are effective for meeting one's goals (e.g., appropriate prompts provided to a large language model interface). A priority of the AI R&D Strategic Plan should be the education of domain experts to be able to apply their knowledge effectively when sharing decision-making with an AI model.

The need for such education extends well beyond the subset of people who develop or train AI models. The likely ubiquity of AI in the future means that everyone will be interacting with AI systems at work, and the increasing degree of automation will transform the job landscape [Hanna et al., 2024]. However, there is still much that we do not know about what skills are required to team with AI systems in a way that boosts productivity. We need a greater understanding of how to develop AI literacy in both education and the workforce. To maximize the effectiveness of human-AI teams, the AI R&D Strategic Plan should invest in attaining and leveraging that understanding to ensure that we have a sufficiently trained workforce.

References

[Aleven et al., 2023] Aleven, V., Baraniuk, R., Brunskill, E., Crossley, S., Demszky, D., Fancsali, S., Gupta, S., Koedinger, K., Piech, C., Ritter, S., Thomas, D. R., Woodhead, S., & Xing, W. (2023). Towards the Future of AI-Augmented Human Tutoring in Math Learning. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (pp. 26–31). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_3

[Aleven et al., 2023] Aleven, V., Baraniuk, R., Brunskill, E., Crossley, S., Demszky, D., Fancsali, S., Gupta, S., Koedinger, K., Piech, C., Ritter, S., et al. (2023). Towards the future of AI-augmented human tutoring in math learning. In *International Conference on Artificial Intelligence in Education*, pages 26–31.

[Balaji et al., 2022] Balaji, A., Kille, B., Kappell, A. D., Godbold, G. D., Diep, M., Elworth, R. L., Qian, Z., Albin, D., Nasko, D. J., Shah, N., Pop, M., Segarra, S., Ternus, K. L., and Treangen, T. (2022). SeqScreen: Accurate and sensitive functional screening of pathogenic sequences via ensemble learning. *Genome Biology*, 23(1):133.

[Basu Mallick et al., 2025] Basu Mallick, D., Burstein, J., Woodhead, S., Sharpnack, J., & Wang, Z. (2025). Preface. *Proceedings of the Innovation and Responsibility in AI-Supported*

Education Workshop, i–vi. https://proceedings.mlr.press/v273/basumallick25a.html

[Belzak et al., 2023]Belzak, W. C., Naismith, B., & Burstein, J. (2023). *Ensuring Fairness of Human-and AI-Generated Test Items*. 701–707.

[Bender et al., 2021] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[Bhowmick et al., 2020] Bhowmick, S., Nagarajaiah, S., and Veeraraghavan, A. (2020). Vision and deep learning-based algorithms to detect and quantify cracks on concrete surfaces from UAV videos. *Sensors*, 20(21):6299.

[Carlini et al., 2021] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2021). *Extracting training data from large language models*. 2633–2650.

[Celaya et al., 2022] Celaya, A., Actor, J. A., Muthusivarajan, R., Gates, E., Chung, C., Schellingerhout, D., Riviere, B., and Fuentes, D. (2022). PocketNet: A smaller neural network for medical image analysis. *IEEE Transactions on Medical Imaging*, 42(4):1172–1184.

[Celaya et al., 2024] Celaya, A., Kirk, K., Fuentes, D., and Riviere, B. (2024). Solutions to elliptic and parabolic problems via finite difference based unsupervised small linear convolutional neural networks. *Computers & Mathematics with Applications*, 174:31–42.

[Conev et al., 2023] Conev, A., Rigo, M. M., Devaurs, D., Fonseca, A. F., Kalavadwala, H., de Freitas, M. V., Clementi, C., Zanatta, G., Antunes, D. A., and Kavraki, L. E. (2023). EnGens: A computational framework for generation and analysis of representative protein conformational ensembles. *Briefings in Bioinformatics*, 24(4):1–11.

[Desai and Shrivastava, 2024] Desai, A. and Shrivastava, A. (2024). In defense of parameter sharing for model-compression. In *International Conference on Learning Representations*.

[Dun et al., 2023] Dun, C., Garcia, M. H., Zheng, G., Awadallah, A. H., Sim, R., Kyrillidis, A., and Dimitriadis, D. (2023). FedJETs: Efficient just-in-time personalization with federated mixture of experts. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

[Edrisi et al., 2023] Edrisi, M., Ogilvie, H. A., Li, M., and Nakhleh, L. (2023). MoTERNN: Classifying the mode of cancer evolution using recursive neural networks. In *RECOMB Inter national Workshop on Comparative Genomics*, pages 232–247. Springer.

[Fasoulis et al., 2024] Fasoulis, R., Rigo, M. M., Lizee, G., Antunes, D. A., and Kavraki, L. E. ´ (2024). APE-Gen2.0: Expanding rapid class I peptide–major histocompatibility complex modeling to post-translational modifications and noncanonical peptide geometries. *Journal of Chemical Information and Modeling*, 64(5):1730–1750.

[Grund et al., 2022] Grund, S., Lüdtke, O., & Robitzsch, A. (2022). Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychological Methods*. https://doi.org/10.1037/met0000526

[Guo et al., 2025] Guo, W., Kingston, Z., and Kavraki, L. E. (2025). CaStL: Constraints as specifications through LLM translation for long-horizon task and motion planning. *International Conference on Robotics and Automation*.

[Hanna et al., 2024] Hanna, A., Nye, C. D., Samo, A., Chu, C., Hoff, K. A., Rounds, J., and Oswald, F. L. (2024). Interests of the future: An integrative review and research agenda for an automated world of work. *Journal of Vocational Behavior*.

[He et al., 2024] He, X., Zheng, J., Fang, J. Z., Piramuthu, R., Bansal, M., Ordonez, V., Sigurdsson, G. A., Peng, N., and Wang, X. E. (2024). FlexEControl: Flexible and efficient multimodal control for text-to-image generation. *Transactions on Machine Learning Research*.

[Hendrycks et al., 2021] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (No. arXiv:2009.03300). arXiv. http://arxiv.org/abs/2009.03300

[Hu et al., 2023] Hu, E., Tang, Y., Kyrillidis, A., and Jermaine, C. (2023). Federated learning over images: Vertical decompositions and pre-trained backbones are difficult to beat. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19385–19396.

[Ji et al., 2023] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

[Johnson et al., 2022] Johnson, B., & Brun, Y. (2022). Fairkit-learn: A fairness evaluation and comparison toolkit. *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, 70–74. https://doi.org/10.1145/3510454.3516830

[Kazadi et al., 2024a] Kazadi, A., Doss-Gollin, J., and Da Silva, A. L. (2024a). Pluvial flood emulation with hydraulics-informed message passing. In *International Conference on Machine Learning*.

[Kazadi et al., 2024b] Kazadi, A., Doss-Gollin, J., Sebastian, A., and Silva, A. (2024b). FloodGNN-GRU: A spatio-temporal graph neural network for flood prediction. *Environmental Data Science*, 3.

[Khalil et al., 2025] Khalil, M., Vadiee, F., Shakya, R., & Liu, Q. (2025). *Creating Artificial Students that Never Existed: Leveraging Large Language Models and CTGANs for Synthetic Data Generation* (No. arXiv:2501.01793). arXiv. https://doi.org/10.48550/arXiv.2501.01793

[Kim et al., 2023] Kim, J. L., Toghani, T., Uribe, C. A., and Kyrillidis, A. (2023). Adaptive federated learning with auto-tuned clients via local smoothness. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.

[Linder et al., 2021] Linder, R., Mohseni, S., Yang, F., Pentyala, S. K., Ragan, E. D., and Hu, X. B. (2021). How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, 2(4).

[Liu et al., 2023] Liu, Z., Desai, A., Liao, F., Wang, W., Xie, V., Xu, Z., Kyrillidis, A., and

Shrivastava, A. (2023). Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342– 52364.

[Liu et al., 2024] Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., Zheng, S., Peng, D., Yang, D., Zhou, D., & Dai, A. M. (2024). *Best Practices and Lessons Learned on Synthetic Data* (No. arXiv:2404.07503). arXiv. http://arxiv.org/abs/2404.07503

[Lu et al., 2022] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., & Kalyan, A. (2022). *Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering*.

[Mallory et al., 2020] Mallory, X. F., Edrisi, M., Navin, N., and Nakhleh, L. (2020). Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biology*, 21:1–22.

[McNamara et al., 2022] McNamara, D., Arner, T., Butterfuss, Reese, Basu Mallick, D., Lan, A., Roscoe, R., Roediger, H., & Baraniuk, R. (2022). Situating AI (and Big Data) in the Learning Sciences: Moving toward Large-Scale Learning Sciences. In *Artificial Intelligence in STEM Education* (pp. 289–308). CRC Press.

[Ochoa et al., 2022] Ochoa, X. (2022). Multimodal Learning Analytics: Rationale, Process, Examples, and Direction. In C. Lang, G. Siemens, & A. F. Wise (Eds.), *The Handbook of Learning Analytics* (2nd ed., pp. 54–65). SOLAR. https://doi.org/10.18608/hla22.006

[Pan et al., 2024] Pan, T., Shome, R., and Kavraki, L. E. (2024). Task and motion planning for execution in the real. *IEEE Transactions on Robotics*, 40:3356–3371.

[Qian et al., 2024] Qian, P., Huang, H., and Unhelkar, V. (2024). PPS: Personalized policy summarization for explaining sequential behavior of autonomous agents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1167–1179.

[Qian et al., 2023] Qian, Z., Cebere, B.-C., & van der Schaar, M. (2023). *Synthcity: Facilitating innovative use cases of synthetic data in different data modalities* (No. arXiv:2301.07573). arXiv. http://arxiv.org/abs/2301.07573

[Rong et al., 2024] Rong, Y., Qian, P., Unhelkar, V., and Kasneci, E. (2024). I-CEE: Tailoring explanations of image classification models to user expertise. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21545–21553.

[Roussarie et al., 2020] Roussarie, J.-P., Yao, V., Rodriguez-Rodriguez, P., Oughtred, R., Rust, J., Plautz, Z., Kasturia, S., Albornoz, C., Wang, W., Schmidt, E. F., et al. (2020). Selective neuronal vulnerability in Alzheimer's disease: a network-based analysis. *Neuron*, 107(5):821–835.

[Shermis et al., 2013] Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*.

[Singh et al., 2025] Singh, N. C., Gilead, T., Chakraborty, A., Van Herwegen, J., van Atteveldt, N., Borst, G., Bugden, S., Jasinska, K., Kay, J., Pugh, K., & Duraiappah, A. (2025). A new education agenda based on The International Science and Evidence Based Education Assessment. *Npj Science of Learning*, *10*(1), 1–9. https://doi.org/10.1038/s41539-024-00288-w

[Wang et al., 2022] Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards Human-Like Educational Question Generation with Large Language Models. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 153–166). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_13

[Wolfe and Kyrillidis, 2022] Wolfe, C. R. and Kyrillidis, A. (2022). Cold start streaming learning for deep networks. *arXiv preprint arXiv:2211.04624*.

[Wu et al., 2025] Wu, C.-T., Shropshire, W. C., Bhatti, M. M., Cantu, S., Glover, I. K., Anand, S. S., Liu, X., Kalia, A., Treangen, T. J., Chemaly, R. F., et al. (2025). Rapid whole genome characterization of antimicrobial-resistant pathogens using long-read sequencing to identify potential healthcare transmission. *Infection Control & Hospital Epidemiology*, 46(2):129–135.

[Xiao et al., 2024] Xiao, Z., Gong, M., Cascante-Bonilla, P., Zhang, X., Wu, J., and Ordonez, V. (2024). Grounding language models for visual entity recognition. In *European Conference on Computer Vision*, pages 393–411. Springer.

[Yang et al., 2021] Yang, F., Alva, S. S., Chen, J., and Hu, X. (2021). Model-based counterfactual synthesizer for interpretation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1964–1974.

[Yao et al., 2018] Yao, V., Kaletsky, R., Keyes, W., Mor, D. E., Wong, A. K., Sohrabi, S., Murphy, C. T., and Troyanskaya, O. G. (2018). An integrative tissue-network approach to identify and test human disease genes. *Nature Biotechnology*, 36(11):1091–1099.

[Zou et al., 2023] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models* (No. arXiv:2307.15043). arXiv. https://doi.org/10.48550/arXiv.2307.15043