Recent Advances and Emerging Frontiers in Artificial Intelligence: 2024-2025

I. Foundational Model Architectures: Evolution and Innovations

The architecture of artificial intelligence models serves as the bedrock upon which their capabilities are built. Continuous refinement and radical rethinking of these foundational structures are paramount for advancing the field. This section examines recent progress in key architectural paradigms, including Vision Transformers (ViTs), the core Transformer model, and related concepts such as Residual and Reversible Networks. The focus is on enhancements in computational efficiency, adaptability to diverse data types, increased representational power, and the emergence of novel operational principles.

A. Vision Transformers (ViT): New Frontiers in Efficiency and Adaptability

Vision Transformers (ViTs) have rapidly ascended to prominence in computer vision, demonstrating remarkable performance on a wide array of tasks. However, early iterations often faced constraints related to fixed input resolutions and significant computational overhead. Research in 2024 and 2025 has concentrated on imbuing ViTs with greater flexibility, efficiency, and adaptability, extending their utility from standard image classification to more nuanced visual understanding and specialized domains such as medical imaging.
A notable development is **UniViTAR (Universal Vision Transformer for Arbitrary Resolution)**, a 2025 model that directly confronts the conventional ViT's limitation of standardizing input resolutions, a practice that can compromise spatial-contextual fidelity when dealing with the inherent variability of natural visual data.[1] UniViTAR incorporates several architectural enhancements inspired by advancements in large language models (LLMs), including 2D Rotary Position Embedding (RoPE), the SwiGLU activation function, and RMSNorm for normalization. Furthermore, it introduces a progressive training paradigm known as resolution curriculum learning, which transitions from fixed-resolution pretraining to native resolution tuning. This allows the model to effectively process images of variable resolutions and aspect ratios natively, leveraging ViT's inherent adaptability to variable-length sequences. This capability is particularly valuable in real-world applications where image dimensions are inconsistent, as it obviates the need for extensive preprocessing and preserves crucial image details.[1] The successful integration of components like 2D RoPE and SwiGLU underscores a growing synergy between vision and language model architectures, where innovations from one domain fertilize progress in the other. This trend suggests a

potential convergence towards more unified architectural primitives for foundation models across different modalities.

Addressing the computational demands of ViTs, the **SCHEMEformer (Scalable CHannEl MixEr for Vision Transformers)**, though introduced in late 2023, provides a foundational improvement highly relevant to 2024-2025 ViT variants.[2] SCHEME introduces an innovative channel mixer module that can be integrated into existing ViT architectures. It enables a more favorable trade-off between model complexity and performance by strategically controlling the MLP structure. A key feature is a channel attention branch that enhances feature mixing during training but can be discarded entirely during inference. This results in substantial accuracy and latency gains (up to 1.5% accuracy improvement or 20% latency reduction reported) without incurring additional computational costs at inference time. The focus on zero-cost inference improvements is indicative of a broader trend where practical deployability—encompassing speed and resource utilization—is gaining importance comparable to raw accuracy. This emphasis is likely to spur further research into "inference-aware" architectural designs.

The maturation of ViTs is further evidenced by their increasing application and success in specialized, high-stakes domains such as medical imaging. Multiple 2025 publications underscore this trend. One study demonstrates that ViT models can surpass traditional transfer learning approaches (using architectures like VGG16, ResNet50V2) for the classification of brain diseases from MRI data, achieving a notable accuracy of 94.39%.[3] Crucially, this work also employs Explainable AI (XAI) methods (GradCAM, LayerCAM, etc.) to interpret model predictions, enhancing transparency and providing insights for medical professionals. Another 2025 review focuses on the adaptation of Vision Foundation Models (VFMs), including ViT and the Segment Anything Model (SAM), for medical image segmentation tasks.[4] This review highlights persistent challenges such as domain adaptation—given the often-limited availability of large-scale, labeled medical datasets—and explores promising avenues like model compression and federated learning to address data scarcity and privacy concerns. The robust performance of ViTs in medical imaging, coupled with active research into overcoming domain-specific hurdles and integrating interpretability, signals their readiness for tackling complex real-world problems beyond general visual recognition. This may lead to the development of more domain-specific ViT variants tailored for scientific and healthcare applications.

B. Transformer Core: Enhancing Capacity and Rethinking Self-Attention

The Transformer architecture, first introduced by Vaswani et al. in 2017, has been a transformative force in sequence processing, particularly in natural language processing and increasingly in other domains. Despite its success, core components like the self-attention mechanism and the feed-forward network (FFN) present limitations, notably in computational complexity with increasing sequence length and in fully harnessing representational capacity. Recent research in 2024-2025 has focused on addressing these fundamental aspects, leading to innovative modifications and even alternatives to the standard Transformer design.

One significant challenge to the conventional Transformer information flow is presented by **LIMe (Layer-Integrated Memory)**, a 2025 model.[5] Standard Transformers primarily utilize representations from the immediately preceding layer, which, as the authors of LIMe argue, can lead to "representation collapse"—a phenomenon where distinct tokens or features become less distinguishable in deeper layers, especially with long sequences. LIMe proposes an extension to the masked multi-head self-attention mechanism that allows the model to retrieve and integrate representations from *all* earlier layers. This is achieved through a learned routing mechanism that blends multi-layer features for both keys and values, without substantial architectural overhaul or overhead. The capacity to access a richer history of representations from shallower layers—which might contain more direct lexical or syntactic cues that get diluted in deeper, more abstract layers—can effectively counter representation collapse and has been shown to improve performance on various language modeling benchmarks.[5] This suggests a potential shift away from strictly sequential information processing within Transformer blocks towards architectures that can more dynamically manage and utilize the full spectrum of their internal states.

The often attention-centric view of Transformer efficacy is being re-evaluated, with a 2025 ICLR submission (tentatively titled "Attention Is Not All You Need: The Importance of Feedforward Networks in Transformer Models" based on available information [6]) investigating the critical role of the Feedforward Network (FFN). This research demonstrates that Transformer models employing deeper FFNs (e.g., three-layer FFNs) but with fewer overall Transformer blocks can achieve superior performance, including lower training loss with fewer total parameters, compared to standard configurations with two-layer FFNs.[6] This finding underscores that the FFN is not merely a supplementary component but a crucial contributor to the model's capacity, and that strategic allocation of parameters and complexity within the FFN can lead to more efficient and powerful models.

The dominance of attention mechanisms, particularly self-attention, is also being challenged by alternative architectures designed for sequence modeling. **PROMPTCOT-MAMBA**, a 2025 paper, introduces a fully attention-free language model based on the Mamba-2 architecture's state space dual (SSD) layers.[8] This model entirely eschews self-attention and the associated key-value (KV) caching mechanism. The benefits are significant for inference efficiency: PROMPTCOT-MAMBA achieves fixed-memory consumption and constant-time inference per token, irrespective of sequence length. This directly addresses the quadratic complexity bottleneck of self-attention, which becomes particularly problematic for long chain-of-thought reasoning. Notably, PROMPTCOT-MAMBA has demonstrated performance superior to strong Transformer and hybrid Mamba-Transformer baselines of comparable scale on complex math and code reasoning benchmarks, even outperforming much larger models like Gemma3-27B on specific evaluations.[8] The maturation of State Space Models (SSMs) like Mamba as viable, high-performing alternatives to Transformers, especially for long-context tasks, may signal a diversification in sequence modeling architectures.

Further refinements to the self-attention mechanism itself are also being explored. A 2025 paper, "Does Self-Attention Need Separate Weights in Transformers?", proposes a **shared weight self-attention** mechanism.[9] Instead of using distinct weight matrices for generating

Keys (K), Queries (Q), and Values (V), this approach employs a single shared weight matrix (Ws). This seemingly simple modification leads to a substantial reduction in parameters within the attention block (a reported 66.53% reduction) and a significant decrease in overall training time. Despite this simplification, the model maintains competitive performance on GLUE benchmark tasks and, in some cases, even outperforms standard BERT baselines, particularly when dealing with noisy or out-of-domain data.[9] This research questions a long-standing design choice in self-attention, offering a pathway to more parameter-efficient and potentially more robust attention mechanisms. Such architectural simplifications that reduce intrinsic parameter counts while maintaining or improving performance contribute to the broader goal of developing more sustainable and accessible AI models.

These developments collectively indicate a dynamic period of innovation within and beyond the Transformer paradigm. Researchers are not only fine-tuning existing components but are also fundamentally questioning established design choices and exploring entirely new mechanisms for sequence processing and representation learning. The emphasis on representational capacity, computational efficiency, and alternative architectural philosophies suggests a future where a more diverse toolkit of foundational models will be available.

C. Beyond Transformers: Advances in Residual and Reversible Networks

While Transformer models have become a dominant force in many AI domains, research into other architectural paradigms, such as Residual Networks (ResNets) and Reversible Networks, continues to yield significant advancements. ResNets remain relevant, particularly for tasks where their convolutional inductive biases are advantageous or where the global attention of Transformers is computationally prohibitive. Reversible architectures, on the other hand, directly address the critical issue of memory overhead in training extremely deep neural networks.

The enduring utility of ResNet architectures is highlighted in a 2024/2025 study on **ResNet in Speech Recognition**.[10] This work demonstrates the application of transfer learning using a pre-trained ResNet model for robust speech recognition, especially in noisy environments. The authors propose an extended architecture designed to effectively learn from both clean and noisy speech data, evaluating it across various noise scenarios (e.g., suburban train, babble) and signal-to-noise ratios (SNRs). This showcases how established architectures like ResNet can be adapted and remain competitive for specific, challenging problem domains through techniques like transfer learning and targeted architectural modifications.

A significant challenge in training very deep networks is the memory required to store activations for backpropagation. **PETRA (Parallel End-to-End Training with Reversible Architectures)**, an ICLR 2025 paper, introduces a novel method to alleviate this bottleneck by parallelizing gradient computations within reversible architectures.[13] Reversible architectures, by their nature, can reconstruct input activations from output activations, thus avoiding the need to store all intermediate activations during the forward pass. PETRA leverages this by allowing parameters at different stages of the network to evolve in parallel across multiple devices. It employs an approximate inversion of activations during the backward pass, which,

combined with the inherent properties of reversible layers, facilitates efficient model parallelism with minimal computational overhead and constant communication costs. This approach significantly reduces memory requirements and overall training time, making it feasible to train even larger and deeper models. This work exemplifies how the demand for training efficiency at extreme scales is driving co-design of architectures and training algorithms.

The powerful concept of residual connections, which enabled the training of much deeper networks by improving gradient flow, is also being extended to new mathematical domains. **LResNet (Lorentzian Residual Neural Networks)**, a 2025 paper, introduces a method for integrating residual connections into hyperbolic neural networks, specifically those operating in the Lorentz model of hyperbolic geometry.[14] Hyperbolic geometry is particularly well-suited for modeling hierarchical data structures, which are common in fields like biology (e.g., phylogenetic trees), social networks, and natural language (e.g., parse trees). LResNet enables the efficient use of residual connections in these non-Euclidean spaces, aiming to combine the benefits of deep residual learning with the representational strengths of hyperbolic embeddings. This could unlock improved performance for tasks involving inherently hierarchical data.

Furthermore, the standard formulation of the residual connection itself (output = x + F(x)) is being revisited and enhanced. The 2025 paper **LAuReL (Learned Augmented Residual Layer)** introduces a novel generalization of the canonical residual connection.[16] LAuReL is designed as an in-situ replacement for standard residual connections and aims to outperform them in both model quality (e.g., accuracy) and footprint metrics (e.g., parameters, latency, memory). The authors demonstrate its efficacy in enhancing model quality for both vision (ResNet-50 on ImageNet-1K) and language models, often with fewer added parameters and less overhead than naively increasing model size.[16] This suggests that the mechanism of information flow between layers can be made more expressive and efficient.

These advancements indicate that foundational architectural components like residual connections are evolving. Rather than simply stacking more layers, research is focusing on making the connections themselves "smarter" or adapting them to new contexts, such as non-Euclidean geometries or parallel training paradigms. This pursuit of more efficient and powerful building blocks is crucial for the continued scaling and applicability of deep learning models across diverse tasks and data types.

II. Perception and Generation: Pushing the Boundaries

This section delves into recent advancements in how AI models perceive specific and complex data types, such as human faces, and how they generate novel content, with a particular focus on Generative Adversarial Networks (GANs). It also touches upon the critical application of these perceptual and generative capabilities in specialized fields like medical imaging, where accuracy and reliability are of utmost importance.

A. Face Recognition: Novel Approaches to Identity and Robustness

Face recognition (FR) technology has achieved remarkable accuracy in controlled

environments. However, significant challenges persist, particularly in scenarios characterized by data scarcity, variations in pose and illumination, the effects of aging, and the detection of out-of-distribution (OOD) or unseen faces. While margin-based loss functions like ArcFace have become well-established benchmarks, recent research in 2024-2025 often seeks to build upon or go beyond these methods, tackling specific robustness issues and exploring novel sensing modalities.

One practical challenge in many specialized FR applications is the lack of large-scale, diverse training datasets. A 2025 paper addresses this by proposing a **GAN-based data augmentation method** specifically designed for face recognition in data-scarce situations.[18] This approach features a residual-embedded generator, aimed at alleviating issues like gradient vanishing or exploding during training, and an Inception ResNet-V1 based FaceNet discriminator for improved adversarial training. The core idea is to generate synthetic facial images that are not only realistic but also possess discriminative features crucial for recognition, thereby enhancing the robustness and generalization capabilities of FR models trained on limited data.[18] The co-evolution of generative models for data synthesis and recognition models for feature extraction continues to be a fruitful avenue for improving FR performance.

Exploring alternative sensing modalities, the **FARE (Face Recognition and Out-of-Distribution Detection using FMCW Radar)** system, introduced in a 2025 paper, presents a novel pipeline that utilizes short-range Frequency Modulated Continuous Wave (FMCW) radar for both face recognition and OOD detection.[21] Instead of relying on visual data, FARE processes Range-Doppler Images (RDIs) and micro-RDIs derived from radar signals. The system employs a unique architecture with a primary path for classifying in-distribution faces (trained using triplet loss) and intermediate paths for OOD detection (trained using reconstruction loss). This radar-based approach offers potential robustness against challenges that typically affect visual FR systems, such as variations in illumination, occlusions, and presentation attacks (spoofing). The integration of OOD detection is also a critical feature for ensuring the reliability of FR systems in real-world deployments where unknown faces or non-face inputs might be encountered.[21] This exploration of non-visual modalities signifies a move towards more robust and versatile FR systems.

The domain of generative face manipulation is also advancing beyond simple face swapping. A 2025 paper introduces a method for **complex head swapping**, which involves transferring the entire head—including face identity, facial shape, and hairstyle—from a source image to a target body image.[22] This is a more challenging task than traditional face swapping, requiring seamless blending and realism. The proposed approach builds upon the PhotoMaker V2 model and utilizes a dataset featuring upper body images with diverse facial orientations, rather than just face-centered crops. A key contribution is the IOMask, designed to automatically generate context-aware masks for more natural head-body integration. Such advancements are relevant for applications in virtual avatar generation, movie and advertisement synthesis, and social media content creation, while also highlighting the need for more comprehensive datasets to train these sophisticated generative models.[22]

Addressing the persistent challenge of aging in face recognition, a 2025 paper proposes the

use of an **additive transformer loss** to supplement existing metric learning loss functions like ArcFace.[23] The core hypothesis is that while facial appearance changes with age, the underlying spatial relationships between facial features remain relatively consistent. The transformer component of the loss is designed to capture these long-range dependencies and global facial structure, thereby learning age-invariant features. This approach aims to make the learned embeddings more robust to age-related variations without discarding the discriminative power of established margin-based losses. This exemplifies a trend of augmenting, rather than entirely replacing, successful existing paradigms to tackle specific weaknesses.

Finally, improving robustness to pose variations without costly retraining is addressed by **Pose-TTA (Test-Time Augmentation for Pose-invariant Face Recognition)**, detailed in a 2025 paper.[24] Pose-TTA enhances the performance of pre-trained face recognition models *during inference* by aligning faces at test time. It utilizes a portrait animator to generate images with canonical poses from input images with arbitrary poses, and then employs a weighted feature aggregation strategy to combine features from the original and augmented images. This method avoids the need for retraining models for different pose conditions and reduces dependency on explicit pose estimators, offering a practical solution for improving the performance of existing FR systems in unconstrained environments where pose variability is common.[24]

These diverse research directions indicate that the field of face recognition is actively working to overcome long-standing challenges related to data, robustness, and the very definition of identity representation. The integration of novel sensing modalities, advanced generative models, and refined loss functions points towards more reliable and versatile FR systems in the near future.

B. Generative Adversarial Networks (GANs): Towards Stable and Modern Baselines

Generative Adversarial Networks (GANs) have been instrumental in advancing the field of image generation, producing highly realistic and diverse imagery. However, their training process is notoriously unstable and often requires a host of empirical tricks and careful hyperparameter tuning. Recent research in 2024-2025 has focused on demystifying GAN training, placing it on a more principled theoretical footing, and modernizing GAN architectures to keep pace with advancements seen in other generative models, such as diffusion models.

A significant contribution in this direction is the 2025 paper introducing **R3GAN ("Re-GAN" - A Modern Baseline GAN)**.[26] This work directly challenges the widely held belief that GANs are inherently difficult to train. The authors first derive a well-behaved regularized relativistic GAN loss function. This loss addresses common GAN training pathologies like mode dropping and non-convergence, which were previously tackled with an assortment of ad-hoc techniques. Crucially, their proposed loss comes with mathematical analysis demonstrating local convergence guarantees, a property lacking in many existing relativistic losses. Armed with this more stable loss function, the authors argue that many of the empirical tricks

commonly employed in GAN training become unnecessary. This newfound stability allows for a more principled approach to architectural design. They then present a roadmap for simplifying and modernizing existing GAN architectures, using StyleGAN2 as a case study. By stripping away non-essential features and incorporating design elements from modern convolutional networks (ConvNets) and Transformers (e.g., increased width with depthwise convolution, inverted bottlenecks, fewer activation functions, separate resampling layers), they develop R3GAN, a minimalist yet powerful baseline. Despite its simplicity and lack of "tricks," R3GAN is shown to surpass the performance of StyleGAN2 on several benchmark datasets (FFHQ, ImageNet, CIFAR, Stacked MNIST) and compares favorably against other state-of-the-art GANs and even diffusion models.[26] This work represents a pivotal step towards making GANs more reliable, easier to train, and potentially more competitive by demonstrating that principled design can lead to superior results.

Another area of focus for improving GAN training, particularly in scenarios with limited data, is the refinement of core architectural components. The CVPR 2024 paper (with potential updates into 2025) on **CHAIN (Improving Batch Normalization in GANs via Disentangled Whitening Transform)** re-examines the role of Batch Normalization (BN) within the discriminator of GANs.[30] While BN is known to improve generalization in standard supervised learning, its application in GANs, especially with scarce data, can lead to issues like discriminator overfitting. CHAIN introduces a novel approach that modifies BN by incorporating a disentangled whitening transform. This technique aims to stabilize GAN training by moderating the gradient of latent features and improve generalization by lowering the gradient of the weights, thereby addressing some of the problematic interactions between BN and the adversarial training dynamic. By improving this fundamental normalization component, CHAIN enhances GAN performance and stability, particularly in practical, data-constrained settings.[30]

The drive towards more principled loss functions and modernized architectures, as seen in R3GAN, suggests a maturation in GAN research. By systematically integrating proven components from the broader computer vision and deep learning landscape, GANs can benefit from years of architectural advancements. This could lead to a new wave of GAN development characterized by greater stability, efficiency, and performance, ensuring their continued relevance in the generative modeling toolkit.

C. Advanced AI in Medical Imaging: Diagnostics and Segmentation

The application of advanced AI techniques, particularly Vision Foundation Models (VFMs) like Vision Transformers (ViTs) and Segment Anything Models (SAM), is rapidly transforming the field of medical image analysis. These models are being deployed for a range of critical tasks, including automated disease detection, precise tumor classification, and accurate segmentation of anatomical structures or pathologies. Key challenges in this domain revolve around the scarcity of large, annotated medical datasets, the need for models to adapt to different imaging modalities and patient populations (domain adaptation), the imperative for high reliability, and the crucial requirement for interpretability to gain clinical trust.

As discussed in Section I.A, ViTs are demonstrating strong performance in medical imaging. A

2025 study highlighted a ViT outperforming established transfer learning models for brain MRI classification, achieving 94.39% accuracy.[3] A critical aspect of this research was the integration of Explainable AI (XAI) methods, such as GradCAM, to provide visual explanations for the ViT's predictions. This transparency is essential for medical professionals to understand and trust the AI's diagnostic suggestions.

A broader perspective is offered by a 2025 review on the use of VFMs in medical image segmentation.[4] This survey details how large models like ViT and SAM, often pre-trained on massive general-domain image datasets, are being adapted for the nuanced task of segmenting medical images. The review underscores significant challenges, including the domain shift between general images and specific medical scans, which can hinder out-of-the-box performance. To address this, research is focusing on efficient adaptation techniques, such as the use of adapters (lightweight modules added to a pre-trained VFM and fine-tuned on the target medical data), knowledge distillation (transferring knowledge from a large VFM to a smaller, more efficient model), and multi-scale contextual feature modeling to capture details at various resolutions. Furthermore, the paper discusses the limitations imposed by small medical image datasets and explores solutions like federated learning, which allows models to be trained on decentralized data without sharing sensitive patient information, thereby addressing both data scarcity and privacy concerns.[4]

The increasing adoption of VFMs in medical imaging suggests they are becoming a new baseline for tackling complex analytical tasks in this field. However, their successful deployment hinges on specialized adaptation strategies tailored to the unique characteristics of medical data. This includes not only technical solutions for domain shift and data limitations but also a strong emphasis on building trust through interpretability. The integration of XAI is not merely an add-on but a core requirement for clinical adoption, as it allows medical practitioners to scrutinize and validate the reasoning behind AI-driven outputs. This trend is fostering the emergence of "Medical VFMs," a subfield dedicated to making these powerful models effective, reliable, and trustworthy within the healthcare ecosystem.

III. The Convergence of Modalities: Towards Unified AI

A significant trajectory in contemporary AI research is the development of models capable of understanding, processing, and generating information from multiple types of data simultaneously. This pursuit of multimodal AI—encompassing text, images, audio, video, and even actions—is a crucial step towards creating more general, adaptable, and human-like artificial intelligence. This section explores advancements in multimodal and cross-modal foundation models, including their architectures and applications, and delves into the synergistic domain of embodied AI and multimodal reinforcement learning.

A. Multimodal and Cross-Modal Foundation Models: Architectures and Applications

The primary objective in this area is to build models that can learn joint representations across different modalities and execute tasks that necessitate an understanding of the complex interplay between them. This includes well-established pairings like vision-language models,

as well as emerging audio-language and truly omni-modal systems designed to handle a multitude of data types.

Research into the internal workings of these models is providing valuable insights. A 2024 study investigated the internal representations of three recent multimodal models, analyzing activations from semantically equivalent inputs across text and speech modalities.[38] The findings revealed that cross-modal representations tend to converge in the deeper layers of these models, suggesting the formation of a shared semantic space. However, the initial layers often remain specialized for processing their respective modalities. The study also highlighted that effective length adaptation mechanisms are crucial for reducing the cross-modal gap, particularly between text and speech, though current approaches show limitations, especially for low-resource languages.[38] Understanding how these models achieve multimodality internally is key to designing more effective architectures.

Efficiency in adapting and deploying multiple large foundation models is a major practical concern. The 2025 paper on **CROSSAN (Cross-modal Side Adapter Network)** proposes a solution for this challenge in the context of sequential recommendation systems that leverage diverse modalities like text, images, video, and audio.[39] CROSSAN is a plug-and-play side adapter network designed for efficiently adapting multiple pre-trained unimodal foundation models (such as ViT for images, BERT for text, VideoMAE for videos, and AST for audio) without requiring full fine-tuning of these large backbones. It utilizes a Mixture of Modality Expert Fusion (MOMEF) mechanism to optimize the integration of information from different modalities. This adapter-based approach offers a parameter-efficient pathway to building powerful multimodal systems by reusing existing expert models, which is critical for scalability and resource management.

Another promising strategy for efficiently combining the capabilities of different pre-trained models is **model merging**. Several 2025 papers, including "Unifying Multimodal Large Language Model Capabilities and Modalities via Model Merging" [3], explore this technique as a data-free or data-light method to create more comprehensive Multimodal Large Language Models (MLLMs). The core idea is to combine the parameters or task vectors of multiple specialized MLLMs (e.g., a vision-language model and an audio-language model) to produce a single, more versatile omni-language model. This approach can significantly reduce the computational cost and data requirements associated with training a large omni-modal system from scratch, while allowing for the integration of diverse capabilities and modalities.[42] The emphasis on modularity and efficiency in both adapter-based methods and model merging points towards a future where large-scale multimodal AI systems are composed from pre-trained expert modules, focusing on effective inter-modal communication and fusion.

The push towards truly omni-modal systems is further exemplified by **Nexus**, an industry-level omni-modal LLM pipeline introduced in 2025.[46] Nexus is designed to integrate auditory, visual, and linguistic modalities within a modular, end-to-end framework. This framework allows for flexible configuration of various encoder-LLM-decoder architectures. A key aspect of Nexus is its lightweight training strategy, which involves pre-training audio-language alignment on top of an existing state-of-the-art vision-language model (Qwen2.5-VL), thereby avoiding the costly pre-training stages typically required for vision-specific modalities. The pipeline also

includes an audio synthesis component to generate high-quality audio-text data, supporting applications like Automatic Speech Recognition (ASR) and Speech-to-Speech chat.[46] Such efforts signify a move beyond dual-modality systems towards models that can seamlessly handle a richer set of sensory inputs.

The integration of action as a modality is critical for enabling AI to interact with the physical world. A 2025 survey on **Vision-Language-Action (VLA) Models** comprehensively details the evolution of this rapidly advancing field.[48] VLA models aim to unify perception (vision), natural language understanding, and embodied action within a single computational framework. The survey traces the progress of VLAs through distinct phases: foundational integration (2022-2023) focusing on basic visuomotor coordination; specialization and embodied reasoning (2024) incorporating domain-specific inductive biases and more complex reasoning; and the current phase of generalization and safety-critical deployment (2025) prioritizing robustness and human alignment.[48] The development of VLAs is crucial for robotics and embodied AI, allowing agents to understand complex instructions and execute appropriate actions in dynamic environments.

These advancements collectively illustrate a clear trend towards more integrated and comprehensive AI systems. The focus on efficiency, modularity, the incorporation of more modalities, and the grounding of perception and language in action are all driving the development of AI that can interact with and understand the world in a more holistic and human-like manner.

B. Embodied AI and Multimodal Reinforcement Learning: Bridging Perception and Action

Embodied AI focuses on agents that can perceive their environment through multiple senses (such as vision, language, and audio), reason about this perception, and take actions within that environment to achieve specific goals. Reinforcement Learning (RL) provides a natural and powerful framework for training such agents, allowing them to learn optimal behaviors through interaction and feedback. However, the integration of multimodal inputs and the generation of complex, often continuous, actions add significant complexity to the RL problem. Recent research in 2024-2025 has made notable strides in addressing these challenges, particularly by leveraging the capabilities of large language models and developing scalable offline RL techniques.

The reasoning capabilities of LLMs are being increasingly framed within an RL context. A 2025 paper discusses how chain-of-thought (CoT) reasoning, commonly employed in modern LLMs to solve complex problems, can be naturally viewed as an RL problem.[52] In this perspective, each intermediate reasoning step is considered an "action" contributing to a final answer or decision. The RL objective function then captures how well the model performs over the sequence of these reasoning steps. While not exclusively focused on multimodal RL, this perspective is highly relevant for designing agents that need to perform complex, multi-step reasoning based on multimodal inputs before taking an action in the environment.

A critical enabler for training embodied AI agents is the ability to learn from large, pre-existing datasets of interactions, which are often multimodal in nature (e.g., containing video, text

commands, and recorded actions). Offline RL is the paradigm that addresses learning from such fixed datasets. **SORL (Scalable Offline Reinforcement Learning)**, introduced in a 2025 paper, is a novel offline RL algorithm that leverages "shortcut models"—a new class of generative models—to effectively scale both the training and inference processes.[53] SORL is designed to handle diverse, multimodal datasets and exhibits positive scaling behavior with increased test-time compute. Its one-stage training procedure and ability to perform inference under varying compute budgets make it a promising approach for learning policies from the rich multimodal data available from robot interaction logs or internet-scale video datasets.[53]

Natural language is emerging as a powerful modality for supervising and guiding RL agents. The **T2DA (Text-to-Decision Agent)**, proposed in a 2025 paper, is a framework designed for supervising offline meta-RL with natural language instructions.[55] T2DA first employs a generalized world model to encode multi-task decision data into a dynamics-aware embedding space. Then, inspired by contrastive learning methods like CLIP, it aligns textual descriptions of tasks with these decision embeddings through contrastive language-decision pre-training. This process effectively bridges the semantic gap between language and environment dynamics. Once the text-conditioned generalist policy is trained, the T2DA agent can perform zero-shot text-to-decision generation, meaning it can execute tasks based on novel language instructions it has not seen during training. This leverages the vast knowledge encoded in LLMs to provide task understanding and facilitate generalization in RL agents, a particularly powerful paradigm for multimodal RL where tasks can be intuitively specified through language.[55] This trend points towards "Language-Conditioned RL" becoming a significant research direction.

The Vision-Language-Action (VLA) models, as discussed previously [48], inherently combine multimodal perception (vision and language) with action generation, and are often trained using imitation learning or RL. The evolutionary trajectory of VLAs towards more sophisticated embodied reasoning (a focus in 2024) and enhanced safety and generalization (a focus in 2025) is central to progress in multimodal RL for robotics.

Furthermore, for MRL agents that may need to retrieve and utilize external knowledge (which could be multimodal) to inform their decisions, understanding the utility of this retrieved information is crucial. The ICLR 2025 paper on **SePer (Semantic Perplexity Reduction for Retrieval Utility)** introduces a method to measure the contribution of retrieved information in Retrieval-Augmented Generation (RAG) systems.[58] It does so by estimating the shift in an LLM's internal knowledge distribution (specifically, the reduction in semantic perplexity) when provided with the retrieved context. While not directly an MRL algorithm, SePer's approach to quantifying information utility is relevant for MRL agents that might employ RAG-like mechanisms for planning and decision-making based on multimodal queries to knowledge bases.[58]

Collectively, these advancements underscore a strong trend towards integrating sophisticated perception, language understanding, and reasoning capabilities within RL agents. The emphasis on leveraging LLMs for task specification and knowledge transfer, the development of scalable offline RL methods for learning from diverse multimodal data, and the increasing

role of generative models in representing policies and world models are all paving the way for more intelligent, adaptable, and versatile embodied AI systems.

IV. AI for Science and Understanding: New Paradigms

Artificial intelligence, particularly the recent advancements in Large Language Models (LLMs), is poised to revolutionize the scientific discovery process. These models, trained on vast corpora of text that include extensive scientific literature, offer the potential to assist researchers by synthesizing complex information, identifying hidden patterns, generating novel hypotheses, and even aiding in the interpretation of complex experimental data. This section explores how LLMs are being applied to accelerate scientific endeavors and enhance the interpretability of sophisticated AI models themselves.

A. Large Language Models in Scientific Discovery: Hypothesis Generation and Reasoning

The integration of LLMs into the scientific workflow is rapidly moving beyond simple literature search and summarization towards more active roles in knowledge creation.

A 2025 study quantifies the impact of LLMs on scientific knowledge production by analyzing the publication patterns of researchers involved in LLM development ("insiders") versus those applying LLMs in other domains ("outsiders").[59] The findings indicate that researchers from non-AI domains are increasingly leveraging LLMs to pursue application-focused research, engage in transdisciplinary work, address social accountability related to their findings, and experiment with new evaluation practices. This demonstrates the broad and transformative potential of LLMs across a multitude of scientific disciplines, acting as a catalyst for new research directions and methodologies.[59]

To better understand and guide the role of LLMs in the scientific process, a 2025 survey examines LLM-based hypothesis discovery through the philosophical lens of Charles S. Peirce's framework of abduction (generating explanatory hypotheses), deduction (deriving predictions from hypotheses), and induction (verifying and refining hypotheses against evidence).[60] This structured approach aims to conceptualize how LLMs might evolve from being mere "information executors" to becoming genuine "engines of innovation," capable of contributing to the formulation and validation of new scientific knowledge. Such frameworks are essential for systematically advancing LLMs as tools for discovery.

One of the challenges in using LLMs for hypothesis generation is scalability and computational cost, especially when relying on prompting with numerous examples. **HypotheSAEs (Sparse Autoencoders for Hypothesis Generation)**, a 2025 work, introduces a more efficient method.[61] This approach first trains a sparse autoencoder (SAE) on text embeddings (e.g., from scientific articles or datasets) to identify a set of interpretable features or concepts present in the data. Then, it selects those SAE features that are statistically predictive of a target variable of interest (e.g., a disease outcome, a material property). Finally, an LLM is prompted with texts that strongly activate these predictive features to generate natural language interpretations of what these features represent. These interpretations then serve as human-understandable hypotheses. HypotheSAEs has been

shown to identify reference hypotheses more effectively and produce more predictive hypotheses on real datasets compared to baseline methods that directly prompt LLMs, while requiring significantly less compute (1-2 orders of magnitude reduction reported).[62] This hybrid approach, combining statistical feature discovery with LLM-based interpretation, offers a more scalable pathway for LLM-assisted hypothesis generation.

A concrete application of LLMs in a hybrid system for complex scientific discovery is presented in a 2025 bioRxiv paper focusing on **automating AI discovery for biomedicine**.[66] This framework combines the exploration of large biomedical knowledge graphs (KGs) with a multi-agent system of specialized LLMs. The system aims to systematically identify hidden relationships between biomedical entities (e.g., genes, diseases, drugs) within the KG. Once potential relationships are found, the LLM agents are tasked with extracting supporting facts from scientific literature and even designing AI predictors (e.g., machine learning models) to further investigate and validate these discovered pathways. This demonstrates how LLMs can be integrated with structured knowledge sources and automated workflows to accelerate research in complex domains like biomedicine.

These developments suggest that the role of LLMs in science is evolving. Rather than LLMs operating in isolation, hybrid approaches that combine their broad knowledge and generative capabilities with structured data (like KGs) or statistical methods (like SAEs for feature discovery) are emerging as a robust strategy. This can help ground LLM outputs, mitigate hallucinations, and lead to more scientifically rigorous and testable hypotheses. Furthermore, the emphasis on computational efficiency in methods like HypotheSAEs is crucial for the practical and widespread adoption of these tools. As LLMs become more adept at not just retrieving information but actively participating in the synthesis of new knowledge and the generation of novel hypotheses, they hold the promise of significantly accelerating the scientific discovery cycle. This, however, also brings to the forefront important considerations regarding the validation, authorship, and potential biases of AI-generated scientific insights.

B. Explainable AI (XAI): Deepening Interpretability with LLMs

As artificial intelligence models, particularly deep learning systems, become increasingly complex and are deployed in high-stakes decision-making scenarios (e.g., medical diagnosis, financial forecasting, autonomous systems), the need for understanding their internal workings and the rationale behind their predictions becomes paramount. Explainable AI (XAI) is the field dedicated to developing methods that make AI model outputs more transparent and understandable to humans. Large Language Models (LLMs) are emerging as powerful tools within XAI, primarily due to their ability to generate human-understandable narratives from complex, often numerical or abstract, model outputs.

A comprehensive 2025 survey on **LLMs for XAI** highlights this potential, detailing how LLMs can transform intricate machine learning outputs into accessible and easy-to-understand narratives.[68] The survey categorizes approaches into: (1) post-hoc explanations, where LLMs are used to explain the outputs of an already trained model by analyzing why a specific input led to a particular output; (2) intrinsic explainability, which involves designing ML model architectures (potentially incorporating LLMs) to be inherently more interpretable; and (3)

human-centered narratives, where LLMs enhance the explanations for ML model outputs by framing them in natural language, thereby fostering user trust and making the outputs more comprehensible.[69] The core value proposition is that LLMs can act as a bridge between the sophisticated behavior of complex models and human interpretability.

The practical need for XAI is evident in specialized applications. For instance, the 2025 study on ViT for brain disease detection (discussed in Sections I.A and II.C) explicitly employed XAI techniques like GradCAM, GradCAM++, LayerCAM, and ScoreCAM to interpret the ViT model's predictions from MRI data.[3] While this particular study did not use LLMs for generating the explanations, it underscores the critical demand for model transparency, especially in fields like medicine where understanding the basis of an AI's decision is crucial for clinical acceptance and responsible use. LLMs could potentially take the visual outputs of such gradient-based methods and provide textual summaries of the highlighted regions and their significance.

The primary role LLMs are beginning to play in XAI can be seen as that of a "universal translator." Many traditional XAI techniques produce outputs such as feature importance scores (e.g., from SHAP), local rule-based explanations (e.g., from LIME), or saliency maps. These outputs, while informative to experts, can still be cryptic or highly technical for non-expert users or even for domain experts who are not AI specialists. LLMs, with their strong capabilities in natural language generation, summarization, and contextual understanding, can ingest these primary XAI outputs and translate them into tailored explanations suitable for different audiences—be it a developer debugging the model, an end-user affected by its decision, or a regulator assessing its compliance. This has the potential to significantly broaden the adoption and practical impact of XAI by making explanations more accessible, actionable, and ultimately more useful.

However, the use of LLMs in XAI also introduces new challenges. If an LLM is used to generate an explanation, the faithfulness of that explanation to the underlying model's true reasoning process becomes a critical concern. There is a risk that the LLM might generate a plausible-sounding but inaccurate or misleading explanation (a form of "explanation hallucination"). Therefore, future research will need to focus not only on leveraging LLMs to explain other models but also on developing methods to ensure the fidelity and reliability of LLM-generated explanations. This includes making the reasoning processes of the "explaining LLM" itself more transparent, a challenge known as "explainable LLMs." This may lead to the development of meta-XAI techniques, where AI systems (possibly other LLMs or formal methods) are used to validate, critique, and improve the explanations generated for AI systems, ensuring they are both understandable and truthful.

V. Ensuring Trustworthy AI: Safety, Alignment, and Ethics

The rapid advancement and increasing deployment of powerful AI systems necessitate a strong focus on ensuring these technologies are safe, aligned with human values and intentions, and used ethically. This section addresses the critical research area of AI safety,

including the mitigation of systemic risks and emerging threats from systems like embodied AI, and the ongoing challenges in AI alignment, particularly the growing recognition of data-centric approaches.

A. AI Safety: Addressing Systemic Risks and Emerging Threats (including Embodied AI)

AI safety research aims to understand, mitigate, and prevent negative outcomes associated with artificial intelligence. These risks span a wide spectrum, from the misuse of AI for generating harmful content or disinformation, to systemic societal impacts such as job displacement and economic disruption, and even to potential existential risks from future, highly capable AI systems.

A notable shift in the discourse on AI safety is the call to broaden its scope beyond purely technical or catastrophic risks. A 2025 position paper argues compellingly that **AI safety should prioritize the future of work**.[70] The authors contend that the rapid replacement of human labor by AI could systematically disrupt human agency and economic dignity. This could lead to what Kasirzadeh (2025, cited in [70]) terms "accumulative x-risks"—gradual, systemic erosions of societal structures that, while perhaps not immediately catastrophic in the way a "rogue AI" scenario might be, can have profound and destabilizing long-term consequences. This perspective challenges an exclusive focus on future superintelligence, highlighting more immediate and tangible socio-economic risks that require urgent attention from the AI safety community, policymakers, and society at large. This suggests a need for AI safety to become more interdisciplinary, integrating insights from economics, sociology, and ethics to develop proactive transition support and governance frameworks.

To facilitate more systematic research and practical adoption of safety measures, the development of standardized tools and frameworks is crucial. The 2025 paper introducing **AISafetyLab** presents such a resource.[73] AISafetyLab is a unified framework and toolkit designed to integrate representative attack methodologies (e.g., jailbreak attacks against LLMs), defense strategies (both training-based and inference-time), and evaluation techniques (e.g., safety scoring methods like ShieldGemma and WildGuard, specialized benchmarks like Agent-SafetyBench). By providing an intuitive interface and an extensible codebase (publicly available at https://github.com/thu-coai/AISafetyLab), AISafetyLab aims to enable developers and researchers to seamlessly apply and compare various safety techniques, thereby fostering more rigorous and reproducible AI safety research.[73] The availability of such practical tools signals a move towards more empirical and evidence-based approaches in the field.

As foundation models become increasingly integrated into robotic systems that interact with the physical world (Foundation Model-enabled Robotics, or FMRs), a new set of safety concerns emerges. A comprehensive 2025 survey on **physical risk control for FMRs** systematically summarizes current approaches to mitigate these risks across the entire lifecycle of such systems: the pre-deployment phase (e.g., safe data curation, simulation, red-teaming), the pre-incident phase (runtime monitoring before an accident occurs), and the post-incident phase (recovery and learning from failures).[74] The survey highlights several

recent (2024-2025) techniques, particularly in runtime monitoring. These include leveraging Vision-Language Models (VLMs) to detect task success or failure (e.g., Kanazawa et al., 2023, cited in [76]), using VLMs for constraint monitoring (e.g., Code-as-monitor, Zhou et al., 2024a, cited in [76]), utilizing pre-trained video prediction models to assess the plausibility of observed state transitions (e.g., Huang et al., 2024, cited in [76]), and training specialized failure classifiers from human intervention data (e.g., Liu et al., 2024c, cited in [76]). The unique and urgent physical safety challenges posed by embodied AI—where an unsafe action can lead to direct physical harm or property damage—necessitate a specialized focus within AI safety, emphasizing robust real-time monitoring, predictive failure detection, and safe intervention strategies.

These developments indicate that AI safety is evolving into a multifaceted discipline. It is expanding its purview from technical alignment of hypothetical future systems to addressing the concrete socio-economic impacts of current AI, developing practical tools for assessing and improving the safety of deployed systems, and tackling the distinct challenges posed by AI agents operating in the physical world.

B. AI Alignment: Data-Centric Strategies and Overcoming Challenges

AI alignment is the endeavor to ensure that artificial intelligence systems act in accordance with human values, preferences, and intended goals. While much research has focused on algorithmic approaches to alignment, such as Reinforcement Learning from Human Feedback (RLHF), there is a growing recognition that the quality, representativeness, and reliability of the data used in these alignment processes are critically important, and often a primary bottleneck.

A 2025 paper titled "Challenges and Future Directions of Data-Centric AI Alignment" strongly advocates for a **shift towards data-centric AI alignment**.[77] The authors emphasize that enhancing the quality and representativeness of the human feedback data used to align AI systems can have a more significant positive impact than merely tweaking alignment algorithms. The paper provides a qualitative analysis identifying multiple sources of unreliability in human feedback, which can undermine the alignment process. These sources include:

1. **Mis-labeling by humans:** Clear errors where annotators select a suboptimal or incorrect response as preferred.
2. **High subjectivity and lack of context:** For subjective queries (e.g., travel recommendations), preferences can vary widely, and without sufficient context about the user or situation, assessing which response is "better" becomes unreliable.
3. **Different preference criteria:** Annotators may have different personal preferences regarding desirable response attributes (e.g., directness vs. inquisitiveness, emphasis on helpfulness vs. harmlessness).
4. **Different thresholds of criteria:** Annotators might agree on the content but disagree on the severity of certain flaws or the importance of specific qualities, leading to inconsistent preferences, especially when differences between responses are subtle.
5. **Harmful suggestions in both responses:** If both AI-generated response options offer

harmful advice, annotators may be forced to make an arbitrary choice if a "both are bad" option is unavailable, leading to unreliable preference signals.[77]

The core message is that the "Garbage In, Garbage Out" principle applies acutely to AI alignment. If the human preference data fed into alignment algorithms like RLHF is noisy, inconsistent, biased, or poorly contextualized, the learned reward model will inherit these flaws. Consequently, an LLM fine-tuned against such a flawed reward model may exhibit misaligned behaviors, fail to generalize appropriately, or engage in "alignment washing" (appearing aligned during training but behaving undesirably in deployment).

This data-centric perspective suggests that future research in AI alignment will need to significantly invest in several key areas. First, developing more sophisticated and robust methods for collecting high-quality, diverse, and reliable human feedback is essential. This may involve better annotator training, more nuanced data collection protocols that capture context and the reasoning behind preferences, and mechanisms for identifying and resolving inter-annotator disagreements. Second, techniques for cleaning, de-biasing, and augmenting existing preference datasets will be crucial. Third, there is a need for alignment algorithms that are inherently more robust to noise and inconsistencies in the preference data. Finally, while AI-assisted feedback generation and verification can help scale the data collection process, it must be approached with caution, being mindful of the AI's own limitations and potential biases, as highlighted in the survey.[78] The challenges of subjectivity and context dependency in human preferences also point towards a need for AI systems that can perhaps handle a diverse range of human values or allow for personalization of alignment, rather than being optimized for a single, monolithic notion of "aligned" behavior.

VI. Synthesis and Future Trajectories

The advancements across various AI domains in 2024-2025 reveal a field characterized by rapid evolution, increasing cross-pollination of ideas, and a growing focus on efficiency, robustness, and trustworthiness. This concluding section synthesizes the overarching themes and interconnected breakthroughs identified throughout this report, highlights promising underrepresented research frontiers, and discusses unresolved challenges that will likely shape the future trajectory of artificial intelligence.

A. Key Interconnected Breakthroughs and Paradigm Shifts Across Domains

Several key themes and paradigm shifts have emerged from the recent literature, indicating significant interconnections between different areas of AI research:

1. **The Symbiotic Evolution of Architectures:** There is a clear trend of architectural innovations in one domain rapidly influencing others. For instance, components proven effective in Large Language Models (LLMs), such as Rotary Position Embeddings (RoPE) and SwiGLU activations, are being successfully integrated into Vision Transformers (ViTs), enhancing their capabilities.[1] This cross-pollination extends to fundamental mechanisms. Researchers are not only adopting components but also re-evaluating and improving core building blocks like residual connections (e.g., LAuReL [16], LResNet [14]) and

the attention mechanism itself (e.g., LIMe allowing access to all prior layer representations [5], Shared Weight Self-Attention for parameter reduction [9], and attention-free alternatives like Mamba in PROMPTCOT-MAMBA [8]). This suggests a move towards a more unified understanding of effective architectural primitives for foundation models.

2. **Efficiency as a Primary Design Constraint:** The sheer scale of modern AI models has made computational and parameter efficiency a primary design consideration, not merely an optimization afterthought. This is evident across domains:
    - In ViTs, with models like UniViTAR designed for arbitrary resolutions without retraining [1] and SCHEMEformer offering inference-time performance gains at no extra cost.[2]
    - In Transformer core architectures, with attention-free models like PROMPTCOT-MAMBA promising constant-time inference [8] and shared weight self-attention significantly reducing parameters.[9]
    - In ResNet-related research, with PETRA enabling parallel training of reversible architectures to reduce memory overhead.[13]
    - In GANs, with R3GAN demonstrating that principled design with modern backbones can lead to simpler, more efficient, yet powerful models.[26]
    - In LLM pre-training, where new techniques focus on parameter and memory-efficient methods.[79] This pervasive focus is driven by the dual needs of scaling models to unprecedented sizes and deploying them effectively in resource-constrained real-world environments.
3. **Multimodality and Embodiment as Drivers of Generalization:** The rapid progress in Multimodal Foundation Models (e.g., CROSSAN for efficient adaptation [39], Nexus for omni-modality [46], and model merging techniques [42]) and Vision-Language-Action (VLA) models [48] signals a clear trajectory towards AI systems that can perceive, reason, and act in more human-like ways. By integrating diverse data streams (text, image, audio, video) and grounding language and perception in physical action, these models are pushing the boundaries of AI generalization. This is tightly coupled with advancements in Multimodal Reinforcement Learning, where agents learn to make decisions based on rich, multimodal inputs (e.g., T2DA leveraging language for task specification [55], SORL scaling offline RL with multimodal data [53]).
4. **Data-Centricity in Specialized and Safety-Critical Domains:** In fields requiring high precision and reliability, such as medical imaging [3], AI for scientific discovery (e.g., HypotheSAEs using LLMs to interpret data-driven features [61]), and AI Alignment [77], the quality, nature, and intelligent utilization of data are becoming paramount. This includes not only the volume of data but also its relevance, diversity, and the methods used to extract information or learn from it. Examples include GANs for generating better training data in face recognition under data scarcity [18], and the critical focus on the reliability of human feedback data for aligning LLMs.[78]
5. **AI Safety Broadening its Horizons:** The field of AI safety is expanding its focus. While

technical alignment of potential future superintelligence remains a concern, there is increasing attention on the immediate societal impacts of current AI, such as the future of work.[70] Concurrently, practical toolkits and benchmarks (e.g., AISafetyLab [73]) are being developed to assess and improve the safety of contemporary systems. Furthermore, the unique and urgent physical safety challenges posed by embodied AI and foundation model-enabled robotics are giving rise to specialized research in controlling physical risks.[74]

These interconnected breakthroughs are summarized in Table 1.

Table 1: Key Paradigm Shifts and Cross-Domain Innovations (2024-2025)

Paradigm Shift/Innovation	Description	Key Exemplar Papers (Source ID)	Primary Domain(s) Impacted	Broader Implication
Architectural Convergence & Component Sharing	Adoption of successful architectural components (e.g., RoPE, SwiGLU) from one domain (LLMs) into others (ViTs), and fundamental improvements to core mechanisms.	UniViTAR [1], LIMe [5], LAuReL [16]	CV, NLP, Foundational Models	More unified and powerful architectural primitives, faster diffusion of best practices.
Shift from Attention-Only to Hybrid/Alternative Seq. Processors	Exploration of State Space Models (e.g., Mamba) and architectural rebalancing (e.g., FFN importance) to overcome attention's limitations or improve efficiency.	PROMPTCOT-MAMBA [8], FFN in Transformers [6]	NLP, Sequence Modeling	Potential for more efficient and scalable long-context models, diversification of sequence processing architectures.
Principled Design Replacing Heuristics in Generative Models	Focus on mathematically grounded, stable loss functions and modern backbones for GANs, reducing	R3GAN [26]	CV, Generative Modeling	Improved trainability, stability, and performance of GANs, making them more competitive and

	reliance on empirical tricks and improving trainability.			accessible.
Data-Centric AI Alignment	Prioritizing the quality, diversity, and reliability of human feedback data as fundamental to achieving robust and genuinely aligned AI systems.	Data-Centric AI Alignment Survey [78]	AI Alignment, NLP, Ethics	More robust and genuinely aligned AI systems; a necessary shift in research focus towards data quality and collection methodologies.
Modular and Efficient Composition of Foundation Models	Utilizing adapters, model merging techniques, or lightweight training strategies to combine multiple specialized foundation models for complex multimodal tasks.	CROSSAN [39], MLLM Merging [42], Nexus [46]	Multimodal AI, NLP, CV, Audio	Scalable development of powerful, comprehensive AI systems without the need for full retraining from scratch, fostering modularity and reusability.
Language as a Key Interface for Embodied Agent Control & Learning	Leveraging natural language for task specification, knowledge transfer, and supervision in reinforcement learning and robotics, enabling more intuitive interaction.	T2DA (Text-to-Decision Agent) [55], VLA evolution [48]	Robotics, RL, Multimodal AI	More intuitive, flexible, and generalizable robots and autonomous agents capable of understanding and acting upon linguistic instructions.
Expansion of AI Safety Concerns	Broadening the scope of AI safety from purely technical alignment of	AI Safety & Future of Work [70], AISafetyLab [73], FMR Safety Survey [74]	AI Safety, Ethics, Robotics, Policy	A more holistic and practical approach to ensuring AI benefits humanity,

	future AGI to include immediate socio-economic impacts and the safety of embodied AI.			addressing both long-term and near-term risks.

B. Promising Underrepresented Research Frontiers and Unresolved Challenges

Despite the rapid progress, several research frontiers remain relatively underrepresented or pose significant unresolved challenges:

- **Multimodal Reinforcement Learning with Complex Reasoning:** While advancements like T2DA [55] and SORL [53] are pushing the boundaries of offline MRL, the integration of deep, long-horizon, and compositional reasoning (akin to the chain-of-thought capabilities seen in LLMs [52]) directly into MRL agents that operate in rich, continuous, and partially observable multimodal environments remains a formidable challenge. How can agents learn to "think" multimodally, plan over extended horizons, and adapt their reasoning strategies in dynamic settings?
- **True Omni-Modal Generalization and Compositionality:** Current MLLMs, such as Nexus [46] and those created via model merging [42], are making strides towards handling an increasing number of modalities. However, achieving truly seamless generalization to new, unseen combinations of modalities and tasks, particularly those requiring strong compositional understanding (i.e., understanding how concepts from different modalities combine to form new meanings), is still an open frontier.
- **Causal Reasoning in Foundation Models:** While LLMs can generate plausible hypotheses based on correlations found in vast datasets [60], embedding robust causal reasoning capabilities is largely an unsolved problem. For AI to be a true partner in scientific discovery or to make reliable decisions in complex systems, it needs to distinguish correlation from causation, reason about the effects of interventions, and understand underlying causal mechanisms.
- **Scalable and Verifiable AI Safety for Embodied Systems:** The survey on FMR safety [74] highlights many emerging techniques for mitigating physical risks. However, ensuring the safety of highly autonomous robots interacting with unpredictable human environments at scale, especially with formal or verifiable guarantees, remains a monumental task. Runtime monitoring needs to be proactive and predictive, not just reactive, and capable of handling novel, out-of-distribution situations gracefully.
- **Neuro-Symbolic Hybrids at Scale for Enhanced Robustness and Interpretability:** While some works are exploring the integration of LLMs with structured knowledge like KGs (e.g., in biomedicine [66]), the development of large-scale foundation models that natively and seamlessly integrate symbolic reasoning structures with neural learning paradigms is a long-standing but increasingly relevant goal. Such hybrid models could offer improved robustness, better generalization from fewer examples, enhanced

interpretability, and the ability to incorporate explicit domain knowledge and constraints more effectively.

- **Lifelong and Continual Multimodal Learning:** How can these large foundation models be designed to continually learn from new multimodal data streams over extended periods without catastrophically forgetting previously learned knowledge? How can they adapt efficiently to evolving environments, changing task requirements, and new modalities without requiring complete retraining? This is particularly crucial for embodied agents that must operate and learn persistently in the real world.

Addressing these challenges will require concerted effort and innovation across multiple disciplines within AI. The pursuit of more capable, efficient, robust, and trustworthy AI systems continues to drive the field towards exciting and uncharted territories. The developments of 2024-2025 indicate a vibrant research landscape actively tackling these frontiers.

Works cited

1. UniViTAR: Unified Vision Transformer with Native Resolution - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.01792v1
2. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2312.00412
3. An Exploratory Approach Towards Investigating and Explaining Vision Transformer and Transfer Learning for Brain Disease Detection - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.16039v1
4. Vision Foundation Models in Medical Image Analysis: Advances and Challenges - arXiv, accessed June 2, 2025, https://arxiv.org/html/2502.14584v2
5. arxiv.org, accessed June 2, 2025, https://arxiv.org/html/2502.09245v1
6. www.arxiv.org, accessed June 2, 2025, https://www.arxiv.org/pdf/2505.06633
7. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.06633
8. www.arxiv.org, accessed June 2, 2025, http://www.arxiv.org/pdf/2505.22425
9. arxiv.org, accessed June 2, 2025, http://arxiv.org/pdf/2412.00359
10. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2505.01632
11. accessed January 1, 1970, https://arxiv.org/pdf/2505.01632.pdf
12. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.01632
13. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2406.02052
14. Lorentzian Residual Neural Networks - arXiv, accessed June 2, 2025, https://arxiv.org/html/2412.14695v2
15. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2412.14695
16. LAuReL: Learned Augmented Residual Layer - arXiv, accessed June 2, 2025, https://arxiv.org/html/2411.07501v3
17. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2411.07501
18. Facial Recognition Leveraging Generative Adversarial Networks - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.11884v1
19. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2505.11884.pdf
20. [2505.11884] Facial Recognition Leveraging Generative Adversarial Networks - arXiv, accessed June 2, 2025, https://arxiv.org/abs/2505.11884
21. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2501.08440
22. arXiv:2503.00861v1 [cs.CV] 2 Mar 2025, accessed June 2, 2025,

https://arxiv.org/pdf/2503.00861
23. Transformer-Based Auxiliary Loss for Face Recognition Across Age Variations - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2412.02198?
24. Test-Time Augmentation for Pose-invariant Face Recognition - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2505.09256
25. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.09256
26. The GAN is dead; long live the GAN! A Modern Baseline GAN - arXiv, accessed June 2, 2025, https://arxiv.org/html/2501.05441v1
27. [Literature Review] The GAN is dead; long live the GAN! A Modern GAN Baseline, accessed June 2, 2025, https://www.themoonlight.io/en/review/the-gan-is-dead-long-live-the-gan-a-modern-gan-baseline
28. accessed January 1, 1970, https://arxiv.org/abs/2501.05441
29. accessed January 1, 1970, https://arxiv.org/pdf/2501.05441.pdf
30. arXiv:2404.00521v6 [cs.LG] 15 Mar 2025, accessed June 2, 2025, https://arxiv.org/pdf/2404.00521?
31. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2404.00521
32. accessed January 1, 1970, https://arxiv.html/2404.00521v6
33. accessed January 1, 1970, https://arxiv.org/pdf/2404.00521.pdf
34. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2505.16039.pdf
35. [Literature Review] An Exploratory Approach Towards Investigating and Explaining Vision Transformer and Transfer Learning for Brain Disease Detection - Moonlight | AI Colleague for Research Papers, accessed June 2, 2025, https://www.themoonlight.io/review/an-exploratory-approach-towards-investigating-and-explaining-vision-transformer-and-transfer-learning-for-brain-disease-detection
36. accessed January 1, 1970, https://arxiv.org/pdf/2502.14584.pdf
37. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2502.14584
38. How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations - arXiv, accessed June 2, 2025, https://arxiv.org/html/2411.17666v2
39. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2504.10307
40. accessed January 1, 1970, https://arxiv.org/pdf/2504.10307.pdf
41. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2504.10307
42. Unifying Multimodal Large Language Model Capabilities and Modalities via Model Merging - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.19892v1
43. accessed January 1, 1970, https://arxiv.org/pdf/2505.19892.pdf
44. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.19892
45. Unifying Multimodal Large Language Model Capabilities and ..., accessed June 2, 2025, https://paperswithcode.com/paper/unifying-multimodal-large-language-model
46. Nexus: An Omni-Perceptive And -Interactive Model for Language, Audio, And Vision - arXiv, accessed June 2, 2025, https://arxiv.org/html/2503.01879v3
47. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2503.01879
48. Vision-Language-Action Models: Concepts, Progress, Applications and

Challenges - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.04769v1
49. Vision-Language-Action Models: Concepts, Progress, Applications and Challenges - arXiv, accessed June 2, 2025, https://arxiv.org/abs/2505.04769
50. (PDF) Vision-Language-Action Models: Concepts, Progress, Applications and Challenges, accessed June 2, 2025, https://www.researchgate.net/publication/391575814_Vision-Language-Action_Models_Concepts_Progress_Applications_and_Challenges
51. accessed January 1, 1970, https://arxiv.org/pdf/2505.04769.pdf
52. LLM Post-Training: A Deep Dive into Reasoning Large Language Models - arXiv, accessed June 2, 2025, https://arxiv.org/html/2502.21321v2
53. Scaling Offline RL via Efficient and Expressive Shortcut Models - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.22866v1
54. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.22866
55. Text-to-Decision Agent: Offline Meta-Reinforcement Learning from Natural Language Supervision - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.15046v3
56. [2504.15046] Text-to-Decision Agent: Offline Meta-Reinforcement Learning from Natural Language Supervision - arXiv, accessed June 2, 2025, https://arxiv.org/abs/2504.15046
57. Artificial Intelligence - arXiv, accessed June 2, 2025, https://arxiv.org/list/cs.AI/new
58. SePer: MEASURE RETRIEVAL UTILITY THROUGH THE LENS OF SEMANTIC PERPLEXITY REDUCTION - OpenReview, accessed June 2, 2025, https://openreview.net/pdf?id=ixMBnOhFGd
59. Adapting to LLMs: How Insiders and Outsiders Reshape Scientific Knowledge Production, accessed June 2, 2025, https://arxiv.org/html/2505.12666v1
60. www.arxiv.org, accessed June 2, 2025, https://www.arxiv.org/pdf/2505.21935
61. Sparse Autoencoders for Hypothesis Generation - arXiv, accessed June 2, 2025, https://arxiv.org/html/2502.04382v1
62. Sparse Autoencoders for Hypothesis Generation - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2502.04382
63. Sparse Autoencoders for Hypothesis Generation | Papers With Code, accessed June 2, 2025, https://paperswithcode.com/paper/sparse-autoencoders-for-hypothesis-generation
64. accessed January 1, 1970, https://arxiv.org/pdf/2502.04382.pdf
65. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2502.04382
66. Automating AI Discovery for Biomedicine Through Knowledge Graphs And LLM Agents - bioRxiv, accessed June 2, 2025, https://www.biorxiv.org/content/10.1101/2025.05.08.652829.full.pdf
67. Automating AI Discovery for Biomedicine Through Knowledge ..., accessed June 2, 2025, https://www.biorxiv.org/content/10.1101/2025.05.08.652829v1
68. LLMs for Explainable AI: A Comprehensive Survey - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.00125v1
69. LLMs for Explainable AI: A Comprehensive Survey - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2504.00125

70. AI Safety Should Prioritize the Future of Work - arXiv, accessed June 2, 2025, https://arxiv.org/html/2504.13959v1
71. accessed January 1, 1970, https://arxiv.org/pdf/2504.13959.pdf
72. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2504.13959
73. arxiv.org, accessed June 2, 2025, https://arxiv.org/pdf/2502.16776
74. A Comprehensive Survey on Physical Risk Control in the Era of Foundation Model-enabled Robotics - arXiv, accessed June 2, 2025, http://www.arxiv.org/pdf/2505.12583
75. arxiv.org, accessed June 2, 2025, https://arxiv.org/abs/2505.12583
76. A Comprehensive Survey on Physical Risk Control in the Era of Foundation Model-enabled Robotics - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.12583v1
77. Challenges and Future Directions of Data-Centric AI Alignment - arXiv, accessed June 2, 2025, https://arxiv.org/html/2410.01957v2
78. Challenges and Future Directions of Data-Centric AI Alignment - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2410.01957
79. Scalable Parameter and Memory Efficient Pretraining for LLM: Recent Algorithmic Advances and Benchmarking - arXiv, accessed June 2, 2025, https://arxiv.org/html/2505.22922v1
80. Scalable Parameter and Memory Efficient Pretraining for LLM: Recent Algorithmic Advances and Benchmarking - arXiv, accessed June 2, 2025, https://arxiv.org/pdf/2505.22922