Integrating Multimodal Fusion, Pretrained Language Models, and Cognitive Neuroscience for Ethical and Robust AI: Advances, Applications, and Future Directions

Abstract

This comprehensive survey delineates the state-of-the-art landscape and emerging frontiers in multimodal and pretrained language models (PLMs), emphasizing their technical architectures, cognitive alignments, application domains, and ethical frameworks. Motivated by the limitations of unimodal models and insights from human cognition, recent advances forgo traditional language-only paradigms to integrate heterogeneous modalities—including vision, audio, text, and biomedical data—within unified transformer-based frameworks. Key contributions include detailed examinations of multimodal fusion strategies (early, late, and cross-modal attention), retrieval-augmented transformer architectures, and geographic/sociocultural adaptation of language models to address linguistic diversity and bias.

Empirical studies reveal that multimodal large language models (MLLMs) learn embeddings partially mirroring neural representations in category-selective brain regions, aligning artificial semantic and perceptual features with human conceptual organization. This intersection of cognitive neuroscience and AI fosters enhanced interpretability and robustness, guiding the design of models that better emulate human knowledge structures. Furthermore, advances in retrieval-augmented generation and continual learning enhance factual consistency and long-range contextual understanding, while multi-objective pretraining integrating human preferences directly into model objectives improves alignment and reduces toxic or hallucinated outputs.

The survey comprehensively explores applications spanning healthcare—where multimodal AI supports diagnosis, personalized medicine, and surgical assistance—autonomous systems with real-time multimodal sensor fusion for safety, speech recognition, cross-lingual NLP, and emotion recognition, highlighting substantial gains and ongoing practical challenges. In parallel, it critically assesses explainable AI (XAI) frameworks centered on graph neural networks and causal inference to assure transparency and trustworthiness, alongside dynamic privacy-preserving and adaptive trust mechanisms essential for ethical deployment in sensitive contexts.

Notwithstanding these advances, the work identifies persistent challenges including data scarcity—particularly in low-resource languages and geographic regions—and computational scalability constrained by transformer self-attention complexity. Ethical imperatives demand frameworks curtailing bias, hallucinations, and privacy breaches, underscoring the need for multidisciplinary collaboration integrating technical innovation with domain expertise and regulatory considerations.

In conclusion, this synthesis articulates a cohesive narrative linking transformer-based architectural innovations, multimodal fusion paradigms, and interdisciplinary cognitive insights, situating them within the critical context of ethical AI development. The integration of scalable, interpretable, and culturally aware AI models portends transformative impacts across healthcare, education, transportation, and multilingual communication, charting a roadmap towards robust, transparent, and human-aligned artificial intelligence systems.

"'latex

ACM Reference Format:

. 2025. Integrating Multimodal Fusion, Pretrained Language Models, and Cognitive Neuroscience for Ethical and Robust AI: Advances, Applications, and Future Directions. In . ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The evolution of multimodal and pretrained language models (PLMs) represents a pivotal advancement in artificial intelligence. By integrating diverse data modalities with pretrained knowledge representations, these models have expanded AI capabilities beyond the constraints of traditional single-modality paradigms [4]. Early innovations in PLMs focused predominantly on language-only Transformer architectures, which established robust frameworks for capturing linguistic structure and semantics at scale [25]. Nonetheless, the limitations inherent in uni-modal training have spurred the development of multimodal large language models (MLLMs) that synthesize heterogeneous inputs—including text, images, audio, video, and structured biomedical data—to enable richer, contextually grounded understanding and generation [4**?**].

This transition is motivated in part by insights from cognitive neuroscience and human cognition, which emphasize the integrated, multisensory nature of conceptual representations in the brain [6**?**]. Empirical research reveals that human conceptual knowledge encompasses both semantic and perceptual features, systematically embedded across modality-specific and associative cortical areas [12**?**]. Recent studies utilizing representational similarity analysis (RSA) demonstrate that MLLMs develop object and concept embeddings that partially mirror human neural representational geometries in category-selective regions such as the extrastriate body area (EBA) and fusiform face area (FFA) [6]. This convergence between artificial and biological cognition underscores the utility of grounding computational architectures in neural and behavioral data to enhance model interpretability and functional alignment [**?**].

Beyond theoretical foundations, the integration of multimodal data within PLMs has driven substantial improvements in practical

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

https://doi.org/10.1145/nnnnnnn.nnnnnnn

domains. In healthcare, for example, MLLMs effectively leverage diverse medical modalities—including imaging, omics, electronic health records, and wearable sensor data—to facilitate nuanced clinical decision support, digital clinical trials, and pandemic surveillance [8]. These advances rely on sophisticated fusion mechanisms such as modality-specific encoders and cross-modal attention, combined with pretraining and fine-tuning strategies tailored to address data heterogeneity and scalability challenges [19]. Similarly, in social and collaborative learning contexts, incorporating nonverbal signals like posture and environmental cues through generative MLLMs enhances interaction analysis granularity, enabling more effective pedagogical interventions [**?**]. Collectively, these applications illustrate the broad impact of multimodal PLMs in complex, real-world tasks requiring integration and interpretation of diverse data streams [16].

From a technical perspective, the development of MLLM architectures has evolved from early fusion techniques to dynamic cross-modal attention and instruction tuning, reflecting increasingly sophisticated approaches to capture multimodal correlations and task adaptability [26]. Addressing efficiency concerns driven by the quadratic complexity of self-attention mechanisms, innovations such as sparse Mixture-of-Experts (MoE) models—exemplified by the Switch Transformer—enable scaling to trillion-parameter regimes through selective expert activation, thereby balancing performance with computational costs [17]. Concurrent foundational work confirms that Transformer architectures possess the expressivity and Turing completeness necessary for modeling long-range dependencies and universal computation essential for multimodal reasoning [**?**]. Nonetheless, challenges persist in data alignment, representation robustness, and interpretability, necessitating multi-level approaches that combine quantitative evaluations with cognitive assessments [5].

Equally important are ethical and trust considerations, especially in high-stakes domains such as healthcare and finance. Emerging frameworks embedding dynamic trust profiling, adaptive information sensitivity detection, and privacy-preserving output control mechanisms provide critical safeguards for responsible AI deployment [18]. Moreover, advances in explainable AI (XAI) tailored to multimodal data fusion—particularly through the use of graph neural networks that preserve causability—highlight the imperative for transparent, human-centered explanations aligned with domain expertise [**?**]. These dimensions emphasize the necessity of integrating ethical, privacy, and interpretability concerns early in model design and pretraining, alongside efforts to improve model alignment with human values and to mitigate undesirable behaviors [7].

The extensive body of research on multimodal and pretrained language models spans foundational cognitive parallels, technical innovations in model architectures and efficiency, diverse application domains, and ethical imperatives. This survey situates these multifaceted advances within a cohesive framework, elucidating both the potential and ongoing challenges of combining computational AI with human cognitive and neural insights. The remainder of this work is organized as follows: Section 2 elaborates on multimodal model architectures and training regimes; Section 3 discusses cognitive and neural alignment studies; Section 4 surveys healthcare and scientific applications; Section 5 addresses ethical and trust frameworks; and Section 6 outlines future research directions and open challenges.

2 Representation Learning and Multimodal Fusion

2.1 Multimodal Embeddings and Fusion Techniques

Multimodal representation learning aims to integrate heterogeneous data modalities—including vision, language, audio, and nonverbal behavioral signals—into cohesive embeddings that facilitate a wide range of downstream tasks in understanding and generation. Fusion strategies traditionally bifurcate into early fusion and late fusion techniques. Early fusion involves concatenating or merging modality-specific features at an initial processing stage to form joint embeddings, thereby enabling cross-modal feature interactions and learning. Conversely, late fusion combines modality-specific decision outputs at the inference stage, offering modularity and adaptability, particularly in leveraging pretrained unimodal models [6**?**].

Recent advances in cross-modal attention mechanisms have dramatically advanced fusion paradigms by enabling dynamic alignment and context-dependent weighting of heterogeneous features. Such mechanisms are especially effective in audiovisual and language integration tasks, underpinning improved semantic synchronization and interpretive fidelity in state-of-the-art multimodal architectures [12, 24**? ? ?**]. These attentions facilitate nuanced cross-modal interactions, enabling models to selectively emphasize pertinent information across modalities dynamically during inference.

Unified multimodal models increasingly leverage weak supervision and contrastive learning objectives over vast multilingual and multimodal corpora to scale their generalization capabilities [10, 16]. This approach permits the mapping of disparate modalities into a shared semantic embedding space that supports zero-shot and few-shot transfer learning. Notable examples such as Flamingo, PaLM-E, and GPT-4 embody this unified modeling philosophy: they employ single-stream or modular transformer backbones and train with multimodal language modeling and cross-modal contrastive losses, thereby demonstrating versatility across vision, language, audio, and other modalities [**?**]. These models manifest emergent generalist competencies, including image captioning, visual question answering, and audio-visual scene interpretation, highlighting the effectiveness of integrated large-scale multimodal training regimes.

Beyond canonical modalities, the incorporation of non-verbal multimodal signals—such as postural behavior—in contexts like collaborative work and education has become increasingly prominent. Generative AI-driven frameworks using multimodal large language models (LLMs) have proven capable of extracting meaningful features from complex, noisy non-verbal data collected in naturalistic environments [6**?**]. This integration not only broadens the applicability of multimodal fusion techniques but also underscores the necessity for adaptable architectures that process verbal and non-verbal information streams simultaneously.

Despite these promising developments, several challenges remain. Modality alignment and fusion complexity require sophisticated architectural designs and objective functions that balance the

representation of modality-specific nuances while fostering shared semantic abstractions. Scalability issues also arise as both model size and multimodal data volume increase training demands. Moreover, the issue of multimodal hallucination—where models generate semantically plausible but factually incorrect associations across modalities—raises concerns in sensitive applications, emphasizing the need for robust calibration methods and human-in-the-loop validation approaches [24].

2.2 Handling Noisy and Diverse Textual Inputs

The heterogeneity of textual inputs, particularly those originating from social media platforms like Twitter or from low-resource languages, imposes substantial challenges on representation learning frameworks. Textual noise manifests as deviations from standard language conventions, including irregular syntax, spelling variations, and code-switching, all of which undermine the assumptions underpinning pretrained language models (PLMs). To address these complications, specialized embedding methodologies have been developed that extract multi-layer latent features from models such as BERT, combining linguistic signals from diverse abstraction levels to create enriched sentence representations. This approach has been shown to improve downstream classification performance on noisy textual datasets [9]. Notably, intermediate transformer layers often encode richer linguistic information for noisy text than the final layers, revealing a form of layer-wise representational specialization exploitable for robust noisy text understanding.

Cross-lingual semantic similarity tasks present additional complications due to disparities in resource availability and structural differences across languages. Approaches leveraging multiple monolingual PLMs to independently embed sentences, followed by integration of these representations, have shown efficacy in capturing cross-lingual semantic alignments. This strategy minimizes dependence on extensive parallel corpora or heavy multilingual pretraining, making it particularly suitable for low-resource and multilingual contexts [**?**].

Fundamentally, large PLMs exhibit notable limitations including vulnerability to adversarial inputs, difficulties in compositional generalization, and challenges in interpretability [27]. These models' brittleness when presented with out-of-distribution or noisy linguistic phenomena highlights the insufficiency of purely statistical pattern recognition in complex languages. Such deficiencies motivate the exploration of hybrid architectures that combine symbolic reasoning and explicit knowledge with learned representations, bolstering robustness and generalization capabilities in intricate linguistic scenarios.

In summary, synergizing advanced multimodal fusion techniques with strategies specialized for handling noisy and diverse textual inputs is essential for constructing AI systems approximating human-like understanding. The integration of semantic and perceptual embeddings, supported by cross-modal attention and large-scale weakly supervised training, provides a foundational architectural blueprint. Nevertheless, achieving robust generalization across noisy, multilingual, and multimodal inputs demands ongoing innovations addressing representational alignment, interpretability, and scalability challenges [6, 9, 10, 12, 16, 24, 27**? ? ? ? ? ?**].

2.3 Applications of Multimodal AI and Large Language Models

2.3.1 Healthcare and Biomedical Domains. The integration of multimodal AI and large language models (LLMs) represents a fundamental shift from traditional unimodal methodologies toward comprehensive, personalized healthcare solutions. By harnessing diverse data modalities—such as genetic, proteomic, clinical, imaging, and environmental information—multimodal frameworks effectively capture complex pathophysiological interactions that single-source models fail to represent adequately [**?**]. This comprehensive approach advances personalized medicine by enabling enhanced patient stratification for clinical trials, dynamic pandemic surveillance, and the creation of virtual health assistants that provide nuanced clinical decision support.

A prominent illustration of this integration is the CONCH system, which synergizes patient data with contextual clinical information to improve diagnostic accuracy and therapeutic guidance [**?**]. Further exemplifying this progress are multimodal datasets in ophthalmology combining fundus autofluorescence (FAF), infrared (IR), and spectral-domain optical coherence tomography (SD-OCT) imaging. The Eye2Gene deep learning system, trained on such heterogeneous imaging data, significantly surpasses expert ophthalmologists by achieving an 83.9% top-five accuracy in predicting gene classes underlying rare inherited retinal diseases across diverse international cohorts [23]. This success largely stems from modality-specific Convolutional Neural Network (CoAtNet0) ensembles, which address data imbalance through weighted loss functions and ensemble averaging, while UMAP visualizations reveal meaningful genotype-phenotype correlations. Eye2Gene's interpretability, facilitated by attention maps, supports clinical trust, though it remains limited by gene coverage gaps and reliance on image-only data comparisons [23].

Multimodal AI also benefits the surgical domain, particularly in intraoperative environments where real-time recognition of surgical instruments enhances workflow efficiency and patient safety. Comparative evaluations of LLMs—including ChatGPT-4 and Google's Gemini variants—show that while category-level instrument recognition attains promising accuracy rates (e.g., 89.1% accuracy for ChatGPT-4o), fine-grained subtype identification presents substantial challenges, with accuracies dropping to approximately 33–39% [**?**]. These results highlight the inherent complexity of nuanced visual pattern recognition in surgical settings and suggest that hybrid retrieval-augmented generation strategies, combining LLMs with domain-specific knowledge bases and data augmentation, are necessary to improve performance.

Beneath these technical achievements lie critical ethical, legal, and deployment challenges. Privacy concerns are paramount in biomedical AI due to the sensitivity of health data, which must be safeguarded without compromising model utility. Strategies such as differential privacy, federated learning, and transparency frameworks are actively explored to mitigate bias and maintain confidentiality. However, practical deployment remains complex owing to computational overheads and real-time operational demands [2**?**]. Trustworthy AI frameworks that dynamically regulate data access based on user roles and data sensitivity—by integrating

attribute-based and role-based access control with semantic sensitivity detection—represent promising directions to balance privacy with information utility in healthcare LLM applications [2].

Looking ahead, progress depends on assembling curated multimodal biomedical datasets and fostering collaborative data-sharing frameworks that strictly adhere to privacy standards while enabling clinically validated AI systems. The development of pretrained multimodal biomedical models that incorporate domain-specific reasoning capabilities is crucial for creating scalable and generalizable AI solutions in medicine [**?**].

2.3.2 Real-Time Safety and Autonomous Systems. Multimodal AI's role in real-time safety management within autonomous and semi-autonomous systems leverages the fusion of heterogeneous data types to enhance situational awareness and enable timely interventions. The integration of drone-acquired imagery, vehicular telemetry, and environmental sensor data—processed via convolutional neural networks and advanced sensor fusion methodologies—facilitates robust detection of traffic hazards, including congestion and accidents. These systems achieve mean average precisions exceeding 90%, with marked performance improvements under adverse or complex environmental conditions [7].

Real-time decision frameworks that combine rule-based reasoning with learning algorithms provide safety alerts with latencies below 200 milliseconds, an essential threshold for effective highway safety interventions. Nevertheless, technical challenges persist in managing limited data bandwidth, ensuring communication reliability (particularly in unmanned aerial vehicle (UAV) networks), and preserving privacy amid extensive data collection [7]. The increasing complexity of traffic scenarios calls for sophisticated predictive analytics and multi-agent coordination mechanisms capable of proactive risk anticipation and mitigation, constituting a vibrant frontier for ongoing research.

2.3.3 Speech Recognition and Cross-Lingual Natural Language Processing. Pretrained language models (PLMs) have substantially advanced speech recognition and cross-lingual natural language processing (NLP), notably in low-resource and linguistically diverse contexts. The incorporation of PLMs such as Chinese BERT into non-autoregressive (NAR) automatic speech recognition (ASR) models alleviates the classical trade-off between decoding speed and transcription accuracy. By enriching acoustic representations with linguistic context provided by PLMs, these systems attain character error rates competitive with traditional autoregressive baselines (e.g., 6.9% vs. 6.5%) while achieving lower real-time factors conducive to rapid inference [24]. Such approaches address inherent challenges in tonal and homophonic features characteristic of Chinese speech without sacrificing computational efficiency.

Cross-lingual adaptation studies reveal that models pretrained on English and fine-tuned systematically outperform native-language models trained from scratch on low-resource languages, demonstrating the potency of transfer learning in exploiting resource-rich linguistic representations [10]. Additionally, geographic adaptation of PLMs through fine-tuning on carefully curated regional corpora tackles entrenched biases inherent in predominantly North American and European English pretrained models. This geographically informed fine-tuning improves performance on underrepresented English variants by over 4 F1-score points and reduces perplexity and error rates related to regional lexical and syntactic variations [**?**]. Collectively, these advances underscore the importance of accounting for linguistic diversity to develop robust, equitable NLP systems.

2.3.4 Text Generation and Emotion Recognition. Text generation frameworks powered by PLMs encompass varied paradigms including open-ended, conditional, and controllable generation tasks. Each paradigm presents distinct challenges regarding output coherence, diversity, and ethical constraints [**?**]. The integration of reinforcement learning (RL) techniques—particularly Reinforcement Learning from Human Feedback (RLHF) and Proximal Policy Optimization (PPO)—has facilitated the alignment of model outputs with human preferences, enhancing multi-step reasoning capabilities and mitigating undesirable biases [**?**]. Notwithstanding, challenges such as sample inefficiency, reward design complexity, and safety concerns persist, necessitating hybrid strategies that integrate symbolic reasoning and hardware acceleration.

In the domain of emotion recognition, combining PLMs with deep neural architectures has extended capabilities beyond single-label classification to nuanced multi-label emotion detection. Utilizing contextualized embeddings alongside sigmoid-activated output layers, these models outperform traditional baselines by macro F1-score improvements of 5–7%, effectively managing overlapping emotional states such as joy, sadness, and anger while offering enhanced interpretability [22]. Nonetheless, challenges remain in handling data imbalance and differentiating semantically close emotional states, motivating future research into incorporating multimodal inputs and explainability methods.

Together, these advances delineate a comprehensive landscape in which multimodal AI and LLMs significantly enhance predictive and generative tasks while foregrounding ethical, privacy, and fairness considerations essential for responsible deployment. Their broad applicability across healthcare, autonomous systems, and natural language domains attests to their transformative potential in modern AI research and applications.

2.4 Explainable AI (XAI), Trustworthiness, and Ethical Considerations

2.4.1 Multimodal Explainable AI (MXAI). The field of explainable AI (XAI) has evolved significantly, transitioning from classical feature attribution techniques based on handcrafted features to more advanced neural visualization and attention-based interpretability methods. More recently, generative post-hoc reasoning techniques have emerged, enabling the synthesis of explanations that are coherent and aligned with human-understandable rationales across multiple data modalities [1, 21, 25]. This progression reflects a necessary adaptation to the inherent complexity of heterogeneous biomedical data — integrating genomic, clinical, imaging, and environmental information — which is essential for addressing multifaceted clinical questions.

Graph Neural Networks (GNNs) have played a central role in advancing biomedical explainability by fusing multi-omics, clinical, and environmental data into heterogeneous knowledge graphs. Leveraging their ability to encode cross-modal relationships and propagate messages across graph structures guided by domain expertise, GNNs enhance the *causability* of models—that is, their

capacity to provide causal, rather than purely correlational, explanations understandable to human experts [1]. This human-centered approach to explainability represents a substantial shift away from focusing solely on technical interpretability metrics, fostering greater trust in clinical decision-support systems by linking predictions directly to established biomedical knowledge.

In parallel, sensor-based multimodal classification research demonstrates that integrating XAI techniques such as SHAP and LIME with rigorous data governance frameworks substantially improves explanation fidelity and accountability. For example, environmental monitoring applications that combine multimodal sensor inputs with explainable gradient boosting and attention mechanisms achieve not only significant accuracy gains but also enhanced interpretability metrics [21]. Nonetheless, these successes highlight ongoing challenges: maintaining a balance between increasing model complexity and preserving user comprehensibility remains difficult, especially when models scale to incorporate high-dimensional heterogeneous data.

Moreover, addressing noise management and mitigating hallucinations in large multimodal models requires hybrid strategies that integrate symbolic reasoning alongside continual learning paradigms [1, 21, 25]. Promising directions include richer context encoding and feature compression techniques, as exemplified by context-aware transformer frameworks, which help constrain model complexity and improve explanation precision [13]. However, disentangling the contributions of intertwined features while preserving interpretability remains a critical and unsolved challenge, emphasizing the need for standardized evaluation benchmarks that focus on explanation faithfulness and human-grounded assessments.

2.4.2 Trust, Privacy, and Security Frameworks. Deploying trustworthy AI necessitates dynamic, context-sensitive frameworks that govern data access and output disclosure based on detailed assessments of user trust and data sensitivity. A notable approach combines Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) with real-time user trust profiling, producing adaptive trust scores informed by credentials, user behavior, and contextual factors [2]. This hybrid trust mechanism enables fine-grained control over sensitive information flow, which is particularly critical in privacy-sensitive domains such as healthcare and finance.

Complementing trust profiling, advanced sensitivity detection modules employing Named Entity Recognition (NER) enhanced with domain-specific lexicons and semantic analysis have achieved high accuracy (exceeding 92%) in detecting sensitive content [2]. When integrated with adaptive output controls — such as differential privacy, information redaction, and summarization — this framework effectively balances the trade-off between data utility and privacy protection. Experimental evaluations confirm that the system maintains responsiveness with minimal latency overhead, typically under 12%, thereby meeting the demands of real-time operations.

Despite these advances, several challenges remain unresolved. Accurately disambiguating sensitive information in evolving or ambiguous contexts is still problematic. There is also a pressing need for continual adaptation of privacy parameters to comply with evolving data governance policies. Machine learning–based trust modeling offers potential for increased adaptability; however, its deployment must remain transparent and auditable to avoid opaque, inscrutable decision-making processes [2]. Additionally, systemic concerns persist regarding transparency, reproducibility, and intellectual property rights in black-box model access, particularly within Language-Models-as-a-Service (LMaaS) paradigms where proprietary models severely limit user visibility into internal operations [11]. Overcoming these barriers calls for standardized benchmarking of black-box models, stronger regulatory oversight, and advanced privacy-preserving techniques to ensure accountable and trustworthy AI ecosystems.

2.4.3 Ethical and Robustness Challenges. Large-scale English language models (LLMs) reveal a diverse spectrum of ethical and robustness vulnerabilities originating from their training data and architectural paradigms. Documented risks include inadvertent memorization of sensitive or private content, systemic biases reflecting the underlying data distributions, the propagation of toxic or false information, and model failure modes when exposed to adversarial inputs [3]. These vulnerabilities significantly complicate efforts to deploy LLMs responsibly in sensitive or high-stakes contexts.

To address these complexities, a mechanistic interpretability framework beyond superficial behavioral benchmarks is essential. This involves in-depth analysis of internal representations and decision pathways to ensure fairness, transparency, and safety in model design and deployment [2, 3, 25, 28**?**]. Achieving this level of rigor faces practical hurdles, including the high computational costs of such analyses and the difficulty of balancing multitask learning objectives without compromising robustness [20].

Emerging research emphasizes improving computational efficiency, developing multimodal grounding for language understanding, and enhancing robustness against adversarial and distributional shifts. Crucially, these technical advances must be embedded within ethical frameworks prioritizing inclusivity and transparency. Cross-disciplinary insights drawn from linguistics, cognitive science, and social sciences play a vital role in addressing bias and toxicity at their origin, ensuring these challenges are tackled both technically and socially [3, 28]. Only through such integrated approaches—melding technical innovation with principled ethical guidelines—can AI systems become truly trustworthy and resilient.

2.4.4 Early Integration of Human Preferences in Large Language Models. A critical advancement in improving LLM alignment and trust involves embedding human preferences early during model pretraining, rather than relying solely on post hoc fine-tuning approaches such as Reinforcement Learning from Human Feedback (RLHF). Recent studies demonstrate that incorporating pairwise human preference judgments directly into the pretraining objective via a multi-objective framework — optimizing both next-token prediction and preference ranking simultaneously — produces models with measurably reduced toxicity and improved alignment to human values [4].

Formally, this approach optimizes the likelihood under a Bradley-Terry model applied to human preference pairs, promoting the internalization of aligned behavior directly within model parameters instead of as an external corrective process. Empirical evidence

shows that such jointly pretrained models outperform those fine-tuned on preference data post-pretraining, achieving lower rates of undesirable outputs without sacrificing language modeling quality. This paradigm shift suggests that early preference integration reinforces intrinsic model trustworthiness and reduces dependency on costly, potentially unstable fine-tuning methods.

Nevertheless, challenges persist including the limited availability and scalability of high-quality preference datasets, and the complex task of balancing competing optimization objectives during large-scale pretraining. Promising future directions involve developing scalable protocols for preference data collection, hybrid frameworks that combine early integration with selective RLHF refinement, and extending these techniques to multimodal architectures. Such advances will enable more nuanced alignment of complex AI systems with diverse human values [4].

3 Behavioral, Cognitive, and Neuroscientific Insights

Representational Similarity Analysis (RSA) has become a pivotal technique for bridging human neural representations and those learned by artificial systems, providing profound insights into conceptual knowledge across different modalities. Research employing RSA demonstrates that embeddings generated by large language models (LLMs) and multimodal large language models (MLLMs) substantially align with neural activation patterns in category-selective brain regions such as the extrastriate body area (EBA), parahippocampal place area (PPA), retrosplenial cortex (RSC), and the fusiform face area (FFA) [6**?**]. This alignment indicates that, while computational models do not replicate human representational space perfectly, they capture essential semantic and perceptual dimensions—including semantic categories like animals and food, as well as perceptual features such as hardness and texture—that underpin organizing principles of human conceptual knowledge [6]. Notably, multimodal models enhance this alignment by incorporating spatial and color information, underscoring the importance of rich sensory data in forming internal representations analogous to those found in the brain [**?**].

However, the degree of alignment revealed by RSA varies across representation levels and model architectures. More granular analyses show that early layers of deep neural networks predominantly encode low-level visual attributes such as shape and texture, while higher layers increasingly represent semantic categories, a hierarchy that echoes processing stages in human visual cognition [**?**]. Despite this hierarchical consistency, important discrepancies remain: certain representational dimensions exclusive to human cognition are underrepresented or absent in current computational models. These gaps emphasize challenges stemming from dataset biases and the inherent complexity of integrating visual and semantic cues, highlighting the need for models embedding richer multimodal context combined with augmented behavioral data [**? ? ?**].

The necessity for enriched multimodal context extends beyond architectural considerations to include the nature and quality of training datasets and behavioral signals guiding representation learning. Empirical evidence suggests that training on multimodal datasets aligned closely with human perceptual and conceptual experiences enhances AI models' capacity to capture human-like representations [**? ?**]. Moreover, this interdisciplinary interface between cognitive neuroscience and artificial intelligence not only promotes interpretability and robustness in artificial agents but also informs ethical frameworks crucial for the development of artificial general intelligence (AGI), where alignment with human cognitive principles is paramount [**? ? ?**].

Complementing these neuroscientific insights, linguistic and cognitive evaluations of pretrained language models (PLMs) shed light on their capabilities and limitations in handling complex syntactic structures, semantic content, and multi-step reasoning tasks prior to any fine-tuning. The bulk of evidence points to robust mastery of fundamental syntactic rules (e.g., subject-verb agreement) and semantic compositionality, as well as competence in analogical reasoning and basic logical inference [3]. Conversely, PLMs struggle significantly with complex syntactic phenomena, negations, implicit pragmatics, and multi-step inferencing, exposing inherent limits in their innate linguistic and conceptual comprehension [3]. These findings dovetail with neuroscientific evidence regarding incomplete semantic alignment, suggesting that prevailing model architectures and pretraining strategies insufficiently capture the full extent of human linguistic and cognitive processing.

Systematic assessments of model consistency—measuring stability of outputs over factual restatements, paraphrases, and negations—further highlight challenges in reliability and interpretability. Empirical benchmarks reveal that prominent pretrained models, including GPT-2, BERT, RoBERTa, and GPT-3, exhibit significant inconsistencies, with factual consistency rates frequently falling below 80% [28]. Such inconsistencies jeopardize trustworthiness, particularly in applications demanding precise and stable knowledge representation. Quantitative analyses identify miscalibrated confidence distributions as a primary source of inconsistency rather than fundamental knowledge deficits. In response, probabilistic calibration methods like temperature scaling have been applied post hoc to adjust confidence levels, yielding improvements of up to 20% in consistency metrics without modifying underlying model weights [28]. This advance reinforces the importance of integrating uncertainty quantification and calibration mechanisms to bolster model robustness and practical deployment viability.

Together, these results portray a nuanced landscape: AI models approximate human-like conceptual embeddings and linguistic competence but remain constrained by limitations in semantic depth, multimodal integration, and output consistency. Addressing these challenges requires interdisciplinary frameworks synthesizing insights from behavioral science, neuroscience, computational modeling, and linguistic theory. Such integrative approaches are essential for developing AI systems whose internal representations and outputs more faithfully mirror the complexity and richness of human cognition and communication [3, 28**? ? ?**].

4 Advances in Retrieval-Augmented and Geographic Adaptation of Language Models

4.1 Retrieval-Pretrained Transformer Architectures

Recent developments in transformer architectures address the inherent limitations of fixed-length context windows in conventional

models by incorporating retrieval mechanisms that dynamically query relevant contextual information. The Retrieval-Pretrained Transformer (RPT) exemplifies such innovation. At each decoding step, RPT derives a query vector from the decoder's hidden state, which attends over a memory bank comprised of past hidden states to retrieve salient information. This self-retrieval framework effectively extends the model's context beyond the conventional fixed window, leading to substantial improvements in long-range language modeling. Empirically, RPT demonstrates significant perplexity reductions on large-scale scientific text corpora, including arXiv and PubMed [17]. Alongside these quantitative gains, RPT exhibits enhanced zero-shot retrieval-augmented generation capabilities, integrating distant context to bolster factual coherence and generation consistency. Crucially, this approach offers scalable and efficient alternatives to full attention mechanisms, whose quadratic complexity limits application on extensive documents, thereby positioning RPT as a promising solution for document-level understanding and generation.

Despite these strengths, the RPT architecture faces notable challenges. Reliance on memory indexing necessitates sophisticated management of retrieval noise and scalable indexing strategies for multi-document retrieval. Retrieval noise can propagate errors into generated outputs, undermining quality, while naïve memory storage proves prohibitively expensive at large scales. Addressing these challenges requires advanced memory selection techniques and integration with external knowledge bases, which remain key directions for future research [17]. Collectively, this body of work lays the foundation for retrieval-augmented language models that maintain coherence across extensive textual spans and enable grounded, knowledge-intensive generation.

In parallel, the Regression Transformer (RT) introduces a versatile foundation model unifying regression of continuous numerical properties with conditional sequence generation within a single transformer framework [15]. RT reframes regression tasks as conditional sequence modeling by employing numerical encodings preserving decimal ordering, thereby inducing an inductive bias favoring numerical proximity. This facilitates multitask learning across diverse scientific domains—spanning small molecule property prediction, protein characterization, and reaction yield estimation—without necessitating task-specific heads. A self-consistency loss applied during training further aligns generated sequences with target property values, enhancing robustness.

The RT model's capacity to generate novel molecules and proteins conditioned on target properties marks a significant advance in molecular engineering and materials science. Its superior performance on benchmark datasets such as MoleculeNet—characterized by high novelty and structural fidelity in generated samples—demonstrates the efficacy of integrating continuous property regression with generative modeling [15]. RT's adaptability and multitasking flexibility complement retrieval-augmented architectures by extending foundation models to diverse data modalities beyond natural language.

4.2 Geographic and Sociocultural Language Adaptation

Mitigating pervasive geographic and sociocultural biases in pretrained language models (PLMs) constitutes a critical challenge toward developing equitable and robust natural language processing systems. Standard PLMs often underperform on region-specific variants due to training data skewed toward dominant linguistic areas, exacerbating disparities in language technology accessibility and accuracy. Recent research in geographic adaptation leverages finely curated, regionally annotated corpora combined with targeted finetuning techniques. These strategies yield measurable improvements in model performance on underrepresented English variants, including African, Indian, and Caribbean English dialects [**?**].

Empirical evaluations demonstrate that region-aware finetuning enhances task-specific F1 scores by approximately 4–5 points across diverse benchmarks such as sentiment analysis and named entity recognition. Concurrently, these adaptations reduce perplexity and error rates linked to region-specific lexical and syntactic phenomena [**?**]. These outcomes underscore that incorporating linguistic context sensitivity through geographic adaptation can mitigate biases without compromising general language understanding capabilities.

Nonetheless, significant challenges persist. The scarcity of high-quality, geographically representative corpora in low-resource regions remains a major bottleneck. Moreover, capturing intersectional sociocultural factors presents complex modeling difficulties. Future directions encourage integrating societal and cultural metadata alongside geographic signals to further improve model fairness and robustness. This multidimensional adaptation paradigm advances technical performance and promotes inclusive language technologies that recognize and respect diverse linguistic identities [**?**].

4.3 Synthesis and Outlook

Together, advancements in retrieval-augmented transformer architectures, multitask scientific foundation models, and geographic adaptation strategies illustrate a progressive shift toward more context-aware, precise, and equitable language modeling paradigms. These developments transcend static, monolithic architectures by enabling dynamic, multifaceted systems capable of nuanced understanding across diverse domains and demographic contexts, marking a critical evolution in the design and utility of large-scale language models.

5 Challenges and Future Directions

5.1 Data and Computational Limitations

The advancement of sophisticated multimodal and multilingual language models faces significant impediments due to the paucity of large-scale annotated datasets that encompass diverse modalities and a wide range of languages, particularly in low-resource and cross-lingual contexts. This scarcity restricts model generalizability and robustness when processing heterogeneous inputs and complicates the alignment of modalities and cultural representations within AI systems [8, 16**? ? ? ?**]. Moreover, insufficient dataset diversity exacerbates biases and underrepresentation, especially concerning geographic and societal factors, thus highlighting the critical need for dataset expansion and the development of equitable benchmarking protocols [7**? ?**].

In parallel, the computational scalability associated with transformer-based architectures remains a vital bottleneck. The quadratic time and memory complexities of standard self-attention mechanisms hinder efficient training and inference on long sequences and multimodal inputs [5**? ? ?**]. Recent innovations, including sparse transformers and kernel-based attention approximations, have emerged to mitigate these challenges. For instance, Sparse Mixture-of-Experts models such as the Switch Transformer employ expert routing, activating only subsets of parameters per input token, significantly reducing computational overhead while facilitating scaling to trillion-parameter models [5, 26]. Although these advances are promising, they require meticulous tuning to maintain expressive power and ensure balanced load distribution across experts [26]. Therefore, addressing data scarcity and computational constraints demands integrated strategies involving curated datasets, architectural innovations, and cross-modal optimization.

5.2 Interpretability, Ethics, and Safety

Ensuring interpretability and ethical compliance in multimodal AI systems is an urgent and complex challenge, especially as these models increasingly influence sensitive sectors including healthcare, finance, and safety-critical domains [2, 25**?**]. The opaque, black-box nature of large models complicates understanding their internal decision-making processes, necessitating the development of domain-specific interpretability frameworks that align technical explanations with user-centered transparency [3, 16]. Multimodal explainable AI (MXAI) techniques have evolved to integrate explanations across modalities, employing causal inference and counterfactual reasoning to harmonize model rationales with human cognitive expectations [2, 25]. Nonetheless, the heterogeneity of multimodal data and the complexities introduced by fusion layers present substantial obstacles in generating explanations that are both faithful and unbiased [16].

Ethical considerations simultaneously impose stringent requirements to reduce bias, preserve privacy, secure informed consent, and comply with evolving regulatory landscapes [2, 25**?**]. Emerging frameworks embedding dynamic trust mechanisms have demonstrated potential by balancing information disclosure with privacy preservation, utilizing adaptive controls guided by user trust profiles and data sensitivity assessments [3]. Despite these advances, real-world deployment mandates robust safeguards against misuse, fairness violations, and harmful outcomes. These necessities call for continuous monitoring coupled with transparent accountability mechanisms. Consequently, progress in ethical considerations and interpretability research must advance in concert with technical model development to achieve responsible AI aligned with societal values.

5.3 Model Advancements and Emerging Research Frontiers

Recent research efforts have centered on developing unified architectures capable of seamless cross-modal and cross-lingual integration within a single framework. These models facilitate zero-shot and few-shot learning, as well as refined multi-document retrieval, thereby enhancing transferability and generalization across tasks [16, 17, 26**? ? ?**]. A core technical innovation underpinning such models involves embedding space alignment and contrastive learning paradigms that improve representation quality in multilingual and multimodal contexts [10**? ? ?**]. These approaches transcend isolated modality modeling to capture complementary semantic and perceptual cues crucial for achieving human-like understanding.

Furthermore, integrating neuroscientific and cognitive insights into model architectures presents promising avenues toward interpretable and robust generalization that aligns AI representations with human conceptual knowledge [2, 26**? ? ?**]. Empirical studies demonstrating the alignment of multimodal embeddings with neural representations localized in category-selective brain regions underscore the potential for cognitive-inspired architectures to enhance semantic and perceptual grounding. Additionally, research into temporal dynamics and the development of lightweight models tailored for real-time applications address critical operational constraints and the requirements of dynamic environments [14].

5.4 Integration and Multidisciplinary Collaboration

The progression of scalable sparse transformer architectures, typified by innovations such as the Switch Transformer, demands integration with advanced explainability, trust, and privacy frameworks to cultivate transparent and secure AI systems [26]. Such synergy enables models that operate efficiently at scale while complying with ethical norms and regulatory expectations. Given the multifaceted technical and ethical challenges, collaboration across sectors—including AI researchers, clinicians, ethicists, linguists, and security experts—is indispensable for ensuring domain-appropriate and safe AI deployment [2, 3, 26**?**]. This multidisciplinary engagement facilitates the establishment of standards and best practices tailored to specific fields such as healthcare, finance, and education, thereby promoting responsible AI adoption.

5.5 Domain-Specific Prospects

Multimodal fusion and retrieval-augmented approaches have been increasingly adopted across diverse domains, including biomedical research, digital health, collaborative learning, autonomous driving, safety management, speech recognition, and multilingual natural language processing [6, 7, 10, 17, 24**? ? ?**]. These applications leverage integrated data streams—spanning medical imaging, environmental sensors, and other sources—to generate enriched insights and predictive power unattainable by unimodal systems. Nevertheless, the success of such methods depends critically on the expansion of datasets with equitable geographic and cultural representation to mitigate biases and improve model generalizability [7**? ? ?**].

In the realm of biomolecular AI, prospects include the integration of structural bioinformatics with class II Human Leukocyte Antigen (HLA) prediction models, such as transformer-based peptide-HLA binding predictors, which offer promise for advancing vaccine design and immunotherapy. Additionally, automated mutation optimization pipelines that exploit transformer attention to enhance binding affinity predictions reveal valuable avenues for experimental validation and iterative model refinement [**?**]. Thus, domain-specific research continues to advance technical frontiers while

emphasizing equitable data representation and adherence to ethical imperatives.

5.6 PLM-Specific Innovations and Challenges

Pretrained language models (PLMs) encounter notable challenges when processing noisy and informal textual inputs, such as social media content. Layered BERT-based representations, which capture diverse linguistic features across model layers, have demonstrated effectiveness in improving understanding of non-standard language use [9]. Complementarily, emotion recognition models that integrate multimodal signals with PLMs have achieved improved detection performance, though difficulties related to data imbalance and limited interpretability persist [22]. Hybrid symbolic-connectionist approaches have also emerged to enhance robustness, controllability, and interpretability in natural language generation and reasoning tasks, aiming to mitigate deep models' intrinsic black-box characteristics [**?**].

To reduce reliance on large supervised datasets, automated prompt construction methods—including neural prompt synthesis and zero-shot prompting—constitute critical innovations. For example, NPPrompt automatically mines and synthesizes external task-related knowledge into coherent prompts, outperforming naive zero-shot baselines and nearing few-shot learning performance without manual prompt engineering [**?**]. These advancements delineate a path toward more efficient and adaptable PLMs that generalize effectively across diverse tasks and domains with minimal labeled data, thereby enhancing scalability and practical applicability.

In summary, the evolving landscape of multimodal and multilingual AI is shaped by intertwined challenges of data scarcity, computational demands, ethical considerations, architectural innovation, and domain-specific complexities. The continuous integration of technical, cognitive, and ethical insights, supported by multidisciplinary collaboration and innovative methodologies, will be paramount in overcoming these obstacles and unlocking the transformative potential of large-scale AI systems across diverse applications.

6 Conclusions

This review has outlined significant advances across three interconnected domains: transformer architectures, multimodal large language models (MLLMs), and foundational pretrained language models (PLMs). Transformer-based models have progressively revolutionized large-scale learning, driven by innovations such as sparse Mixture-of-Experts exemplified by the Switch Transformer. This architecture achieves extreme parameter scaling through efficient routing mechanisms that reduce computational loads without compromising model expressivity [2]. Moreover, the theoretical proof of transformer Turing completeness solidifies their potential for universal computation, reinforcing their status as powerful foundational architectures for a broad spectrum of AI tasks [**?**]. Despite their transformative impact, challenges arising from quadratic self-attention complexity have propelled the development of efficient "X-formers," which employ techniques like sparse attention, kernel approximations, and memory mechanisms to enable more scalable and practical transformer deployments [25]. In computer vision, transformers have supplanted traditional convolutional paradigms by effectively capturing long-range dependencies and hierarchical features; however, their substantial data requirements and training costs underscore the urgency for developing efficient and lightweight variants suitable for real-world applications [26].

The emergence of multimodal large language models marks a pivotal transition from unimodal language understanding to the synergistic integration of vision, audio, text, and other modalities within unified architectures. Surveys emphasize foundational principles underpinning state-of-the-art MLLMs, including modality-specific encoders, cross-modal interaction modules, and fused pretraining schemas [**?**]. Empirical progress across tasks such as image captioning, visual question answering, and audio-visual speech recognition underscores the capacity of these integrative architectures to transcend text-only model limitations [**?**]. Nevertheless, critical technical challenges remain, particularly in addressing multimodal data scarcity, alignment complexity, and elevated computational demands. Addressing these hurdles necessitates innovations in self-supervised learning, parameter efficiency, and the establishment of robust cross-modal evaluation benchmarks [**?**]. In healthcare and biomedical domains, MLLMs leverage heterogeneous data streams—including clinical imaging, omics data, and wearable sensors—to facilitate personalized medicine and real-time patient monitoring, demonstrating transformative applications that harness the richness of multimodal data while also highlighting challenges related to privacy, interpretability, and regulatory compliance [8, 18].

A crucial insight arising from recent intersections between cognitive neuroscience and artificial intelligence is the synergy between artificial and human conceptual representations. Multimodal embeddings generated by MLLMs capture semantic categories comparable to those humans employ while simultaneously encoding perceptual features such as texture, color, and spatial attributes. This results in representational similarities with neural activations observed in category-selective brain areas [19]. However, despite this potent semantic alignment, current models only partially replicate the full complexity of human conceptual understanding—particularly regarding nuanced semantic dimensions and context-dependent reasoning. This gap underscores the importance of context-aware, multimodal architectures supplemented with rich behavioral datasets [1]. Such convergence of computational AI with cognitive and neural insights provides a promising roadmap to enhance AI's human-likeness and interpretability. Notably, methods such as multimodal explainable AI (MXAI) and graph neural network-driven causal explanations promote transparency and causability across data modalities [21**?**].

Ethical and robust AI development emerges as a unifying theme linking architectural innovations with societal impact. Integrating computational advances with cognitive principles facilitates the design of trust-aware frameworks that dynamically tailor data disclosure according to user trust profiles and data sensitivity—a feature especially critical in regulated sectors such as healthcare and finance [7]. Concurrently, efforts in explainability addressing multimodal data fusion foster accountability and user trust by balancing model complexity and interpretability through hybrid XAI and governance frameworks [16]. Persistent challenges—including hallucinations, bias, privacy concerns, and ethical alignment—demand

sustained interdisciplinary research, informed policy development, and transparent model design and deployment [17**?**].

The downstream impact of these integrated efforts manifests across multiple pivotal sectors:

- **Healthcare**: MLLMs enhance clinical decision-making, early diagnostics, and personalized treatments by integrating medical imaging and genomic data analyses [8].
- **Education**: Analysis of non-verbal cues via MLLMs improves collaborative learning analytics, enabling more effective educational technologies [18].
- **Transportation**: Real-time multimodal sensor fusion supports safety interventions, reducing accident risks [10, 18].
- **Multilingual NLP**: PLMs adapted to geographic and cultural specificities mitigate language biases and improve performance in underrepresented linguistic regions [5, 24].
- **Resource-efficient adaptation**: Zero-shot and few-shot learning methods, augmented by retrieval-augmented generation and automated prompt synthesis, exploit pretrained knowledge to facilitate efficient adaptation across diverse domains [3, 28].

Realizing the full societal benefits of these advancements requires sustained interdisciplinary collaboration spanning computer science, cognitive neuroscience, ethics, and domain expertise. It is imperative to develop comprehensive evaluation frameworks that assess not only performance but also fairness, interpretability, and trustworthiness [22]. Enforcing transparency regarding data provenance, model limitations, and decision rationales must accompany ethical stewardship to mitigate potential harms and ensure equitable AI deployment [9**?**]. Furthermore, the continual refinement of foundational models by embedding human preferences enhances alignment and reduces undesirable outputs such as toxic or hallucinated content, thereby increasing model reliability [**?**]. Finally, addressing outstanding challenges in scalability, interpretability, multimodal data fusion, and domain transfer will propel new frontiers in AI, fostering systems that are powerful, efficient, and aligned with human values and cognitive frameworks [4**?**].

In summation, the convergence of advancing transformer-based architectures, emergent multimodal large language models, and integrative cognitive and neural insights marks a transformative era in AI research. When combined with principled ethical frameworks and continuous technical innovations, this synergy promises unprecedented potential to reshape healthcare, education, transportation, and language technologies—enabling AI systems that are robust, transparent, and beneficial across diverse societal domains.

References

[1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol. 2022. Multimodal biomedical AI. *Nature Medicine* 28 (2022), 1773–1784. doi:10.1038/s41591-022-01981-2

[2] R. AlSaad, A. Abd-alrazaq, S. Boughorbel, A. Ahmed, M.-A. Renault, R. Damseh, and J. Sheikh. 2024. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. *Journal of Medical Internet Research* 26 (2024), e59505. https://www.jmir.org/2024/1/e59505/

[3] J. Born and M. Manica. 2023. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence* 5, 4 (April 2023). https://www.nature.com/natmachintell/volumes/5/issues/4

[4] X. Chen, H. Xie, and B. Lei. 2024. Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. *Artificial Intelligence Review* 57, 4 (2024), 91. doi:10.1007/s10462-024-10591-5

[5] Y. Chu, Y. Zhang, Q. Wang, L. Zhang, X. Wang, Y. Wang, D. R. Salahub, Q. Xu, J. Wang, X. Jiang, Y. Xiong, and D.-Q. Wei. 2022. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence* 4 (2022), 300–311. doi:10.1038/s42256-022-00459-7

[6] C. Du, K. Fu, and H. He. 2025. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence* 7, 6 (2025), 548–559. https://www.nature.com/articles/s42256-025-00435-2

[7] W. Fedus, B. Zoph, and D. P. Kingma. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* 23, 1 (2022), 1–39. https://jmlr.org/papers/volume23/21-0998/21-0998.pdf

[8] G. Feretzakis, A. Rivas, S. D. Georgakopoulos, and S. Mitrokotsa. 2024. Trustworthy AI: Securing Sensitive Data in Large Language Models. *AI* 5, 4 (2024), 134. https://www.mdpi.com/2673-2688/5/4/134

[9] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Siddiqui, and M. Shah. 2022. Transformers in Vision: A Survey. *Comput. Surveys* 54, 10s (2022), 1–41. doi:10.1145/3505244

[10] Z. Li, Y. Han, T. Liu, C. Ding, Q. Li, and J. Yin. 2022. Transformer-based Context-Aware Feature Interactions for Click-Through Rate Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2022), 5433–5446. doi:10.1109/TPAMI.2021.3123474

[11] M. Liu, X. Chen, and Q. Huang. 2022. Towards Multimodal Large Language Models: Advances, Challenges, and Opportunities. *AI* 3, 4 (2022), 407–420. https://www.mdpi.com/2673-9541/3/4/407

[12] F. P. Mahner, L. Muttenthaler, and M. N. Hebart. 2025. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence* 7, 6 (2025), 575–588. https://www.nature.com/articles/s42256-025-00437-4

[13] E. La Malfa, A. Petrov, S. Frieder, C. Weinhuber, R. Burnell, R. Nazar, A. Cohn, N. Shadbolt, and M. Wooldridge. 2024. Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges. *Journal of Artificial Intelligence Research* 80 (2024). doi:10.1613/jair.1.15865

[14] P. Moschoula, P. Singh, A. Chakraborty, Y. Oruganti, M. Milletari, S. Bapat, and K. Jiang. 2024. The RL/LLM Taxonomy Tree: Reviewing Synergies Between Reinforcement Learning and Large Language Models. *Journal of Artificial Intelligence Research* 80 (2024). doi:10.1613/jair.1.15960

[15] N. Pontikos, W. A. Woof, and M. Michaelides. 2025. Next-generation phenotyping of inherited retinal diseases from multimodal imaging with Eye2Gene. *Nature Machine Intelligence* 7, 6 (2025), 594–608. https://www.nature.com/articles/s42256-025-01040-8

[16] J. Pérez, R. L. Uria, P. Pollakis, J. Marecek, K. Muroya, and N. Durrani. 2021. Attention is Turing Complete. *Journal of Machine Learning Research* 22, 1 (2021), 1–24. https://jmlr.org/papers/volume22/20-302/20-302.pdf

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 1 (2020), 1–67. https://jmlr.org/papers/volume21/20-074/20-074.pdf

[18] G. Salierno. 2025. Generative AI and Large Language Models in Industry 5.0. *AI* 5, 1 (2025), 30. https://www.mdpi.com/2673-8392/5/1/30

[19] K. Shah, S. Russell, and M. Lakshmanan. 2024. Large Language Model Prompting Techniques for Clinical Decision Support. *J. Clin. Med.* 13, 17 (2024), 5101. https://www.mdpi.com/2077-0383/13/17/5101

[20] J. Sublime. 2024. The AI Race: Why Current Neural Network-based Architectures are a Poor Basis for Artificial General Intelligence. *Journal of Artificial Intelligence Research* 80 (2024). https://jair.org/index.php/jair/article/view/15315

[21] S. Sun, W. An, F. Tian, F. Nan, Q. Liu, J. Liu, N. Shah, and P. Chen. 2024. A Review of Multimodal Explainable Artificial Intelligence: Past, Present and Future. arXiv preprint arXiv:2412.14056. https://arxiv.org/abs/2412.14056 Accessed: 2024-06-20.

[22] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. 2022. Efficient Transformers: A Survey. *Comput. Surveys* 55, 6 (2022), 1–28. doi:10.1145/3530811

[23] R. Whitehead, A. Nguyen, and S. Järvelä. 2025. Utilizing Multimodal Large Language Models for Video Analysis of Posture in Studying Collaborative Learning: A Case Study. *Journal of Learning Analytics* 12, 1 (2025), 186–200. doi:10.18608/jla.2025.8595

[24] H. Wu, W. Wang, F. Wang, X. Chen, and W. Chen. 2022. End-to-End Transformer-Based Framework for Facial Action Unit Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022), 1196–1209. doi:10.1109/TPAMI.2020.3033120

[25] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu. 2023. Multimodal Large Language Models: A Survey. In *IEEE BigData 2023*. 1–10. https://arxiv.org/abs/2311.13165

[26] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. 2023. A Survey on Multimodal Large Language Models. *Nat. Sci. Rev.* 11, 6 (2023). doi:10.1093/nsr/nwae403

[27] Y. Zhang, L. Wang, and J. Hu. 2023. Multimodal Large Language Models for Medical Visual Question Answering: A Survey. *AI* 4, 2 (2023), 287–311. https://www.mdpi.com/2673-9541/4/2/287

[28] Z. Zhang, W. Xiang, and M. Zitnik. 2024. Efficient generation of protein pockets with PocketGen. *Nature Machine Intelligence* 6, 4 (Nov 2024). https://www.nature.com/natmachintell/articles