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, 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

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

This progression draws inspiration from cognitive neuroscience and human cognition, emphasizing the integrated, multisensory nature of conceptual representations in the brain [6**?**]. Behavioral and neural evidence shows that human conceptual knowledge systematically embeds both semantic and perceptual features across modality-specific and associative cortical areas [12**?**]. For instance, recent representational similarity analysis (RSA) studies demonstrate that MLLMs naturally develop object and concept embeddings closely mirroring human neural representational geometries in category-selective brain regions such as the extrastriate body area (EBA) and fusiform face area (FFA) [6]. This alignment between artificial and biological cognition highlights the benefits of grounding computational architectures in neural and behavioral data, enhancing model interpretability and functional convergence [**?**].

Beyond theoretical foundations, the integration of multimodal data within PLMs has driven substantial improvements across practical domains. In healthcare, MLLMs leverage diverse medical

Conference’17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

modalities—including imaging, omics, electronic health records, and wearable sensor data—to support nuanced clinical decision-making, digital clinical trials, and pandemic surveillance [8]. Achieving these capabilities requires sophisticated fusion mechanisms such as modality-specific encoders and cross-modal attention, combined with training strategies that address data heterogeneity and scalability [19]. Similarly, in social and collaborative learning, incorporating nonverbal cues like posture and environmental signals through generative MLLMs enhances interaction analysis granularity and facilitates more effective pedagogical interventions [?]. Collectively, these applications illustrate the broad impact of multimodal PLMs in tackling complex real-world tasks demanding the integration and interpretation of diverse data streams [16].

From a technical perspective, MLLM architectures have evolved from early fusion methods to dynamic cross-modal attention and instruction tuning, showcasing increasingly advanced techniques to capture multimodal correlations and adapt to tasks [26]. To mitigate efficiency challenges posed by the quadratic complexity of self-attention, sparse Mixture-of-Experts (MoE) models such as the Switch Transformer enable scaling to trillion-parameter regimes through selective expert activation, balancing computational cost and model performance [17]. Foundational research also confirms that Transformer architectures possess the expressivity and Turing completeness necessary to model long-range dependencies critical for multimodal reasoning [?]. Nevertheless, ongoing challenges include aligning multimodal data, ensuring robust representations, and enhancing interpretability, motivating multi-level evaluation frameworks that combine quantitative metrics with cognitive benchmarks [5].

Ethical and trust considerations play a vital role, particularly in sensitive areas like healthcare and finance. Emerging frameworks focus on dynamic trust profiling, adaptive sensitivity detection, and privacy-preserving output controls to safeguard responsible AI deployment [18]. Advances in explainable AI (XAI) tailored for multimodal data fusion—especially through graph neural networks preserving causability—underscore the importance of transparent, human-centered explanations aligned with domain expertise [?]. These dimensions highlight the necessity of incorporating ethical, privacy, and interpretability aspects early in model design and pretraining, alongside efforts to enhance alignment with human values and mitigate undesired behaviors [7].

The extensive literature on multimodal and pretrained language models encompasses foundational cognitive parallels, technical innovations in architectures and efficiency, diverse application domains, and ethical imperatives. This survey organizes these multifaceted advances within a coherent framework, clarifying both the opportunities and continuing challenges in integrating computational AI with human cognitive and neural insights. The remainder of this paper is structured as follows: Section 2 details 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

Representation learning plays a crucial role in multimodal fusion by extracting meaningful features from diverse data sources, which can be effectively combined to enhance model performance across various applications. Multimodal fusion methods integrate information from multiple modalities—such as text, images, audio, and sensor data—leveraging complementary strengths to achieve more robust representations and improved inference capabilities.

Fusion techniques broadly fall into three categories: early fusion, late fusion, and hybrid fusion. Early fusion methods combine raw data or low-level features before applying a learning algorithm. This approach enables the model to learn joint representations but may suffer from challenges due to heterogeneous data incompatibility, increased dimensionality, and sensitivity to noisy or missing data. Late fusion methods aggregate output decisions or high-level features from individual modality-specific models, allowing for flexible and independent processing pipelines; however, this may lead to suboptimal integration as joint feature learning is limited. Hybrid fusion attempts to alleviate these limitations by combining features at multiple stages, balancing joint representation learning with modularity, but it may increase model complexity and computational overhead.

Applications of multimodal fusion span diverse domains including multimedia retrieval, emotion recognition, medical diagnosis, and autonomous systems. By effectively combining complementary information from different modalities, fusion methods improve robustness to noise and missing data, enhance interpretability, and facilitate richer context understanding. The selection of an appropriate fusion strategy depends on the nature of the data, computational resources, and specific application needs. For instance, early fusion may be preferred when strong feature-level correlations exist and data is clean, whereas late fusion offers greater modularity and flexibility in heterogeneous systems. Hybrid approaches seek to integrate these benefits but require careful architectural design.

In summary, understanding the strengths and weaknesses of fusion methods, alongside their suitability to application domains, provides valuable guidance for designing effective multimodal systems. Future work should emphasize rigorous comparative evaluations and tailored fusion strategies to address the evolving complexity and heterogeneity of multimodal data.

2.1 Multimodal Embeddings and Fusion Techniques

Multimodal representation learning seeks to integrate heterogeneous data modalities—including vision, language, audio, and nonverbal behavioral signals—into unified embeddings that enhance performance across diverse downstream tasks such as understanding, reasoning, and generation. Fusion strategies have traditionally been categorized into early fusion and late fusion techniques. Early fusion combines modality-specific features at initial processing stages, enabling joint embeddings that facilitate cross-modal feature interactions and deeper integrated learning. In contrast, late fusion merges modality-specific decision outputs during inference, promoting modularity and flexibility, especially when leveraging pretrained unimodal models [6?].

Table 1: Summary of Multimodal Fusion Methods with Limitations and Applications

Fusion Method	Description	Limitations	Applications
Early Fusion	Combines raw or low-level features before learning; captures joint feature space	Sensitive to heterogeneous data and noise; high dimensionality	Multimedia retrieval, medical image analysis
Late Fusion	Integrates decisions or high-level features from separate modality models	Limited joint learning; possible information loss	Emotion recognition, autonomous driving
Hybrid Fusion	Combines early and late fusion advantages via multi-stage integration	Increased complexity and computational cost	Multimodal sentiment analysis, healthcare monitoring

Recent progress in cross-modal attention mechanisms has significantly advanced fusion paradigms by enabling dynamic alignment and context-sensitive weighting of heterogeneous features. These mechanisms excel particularly in audiovisual and language integration applications, supporting improved semantic synchronization and interpretive fidelity in state-of-the-art multimodal architectures [12, 24**? ? ?**]. By allowing models to selectively emphasize salient information across modalities during inference, these attention-based techniques foster nuanced cross-modal interactions and richer representations.

Unified multimodal models increasingly employ weak supervision combined with contrastive learning objectives over extensive multilingual and multimodal corpora to achieve scalable generalization [10, 16]. This strategy facilitates mapping diverse modalities into a shared semantic embedding space, empowering powerful zero-shot and few-shot transfer learning capabilities. Prominent examples such as Flamingo, PaLM-E, and GPT-4 embody this unified modeling paradigm by utilizing single-stream or modular transformer backbones and training with multimodal language modeling alongside cross-modal contrastive losses [**?**]. These models exhibit emergent generalist abilities across tasks including image captioning, visual question answering, and audio-visual scene interpretation, showcasing the effectiveness of large-scale, integrated multimodal training regimes.

Beyond canonical modalities, there is growing emphasis on incorporating non-verbal multimodal signals—such as postural behavior—in research contexts including collaborative work and education. Generative AI frameworks empowered by multimodal large language models (LLMs) have shown promising capability in extracting meaningful features from complex, noisy non-verbal data collected in naturalistic settings [6, 12**?**]. This integration broadens multimodal fusion applications and highlights the necessity of adaptable architectures that simultaneously process verbal and non-verbal streams with robustness and sensitivity.

Despite significant advances, key challenges persist. Effective modality alignment and fusion remain complex, requiring sophisticated architectures and objective functions that carefully balance modality-specific nuances while fostering shared semantic abstractions. Scalability issues also arise as increasing model size and multimodal data volume place higher demands on training resources. Furthermore, multimodal hallucination—where models generate plausible but factually incorrect multimodal associations—poses risks in sensitive domains, underscoring the importance of robust calibration methods and the integration of human-in-the-loop validation protocols [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. For example, Kasthuriarachchy and Senanayake [9] proposed a methodology that aggregates features from various layers of BERT to better capture the linguistic characteristics of noisy social media text. Their work introduced probing tasks specifically designed for Tweets and demonstrated that initial and intermediate transformer layers encode richer and more relevant linguistic information than the final layers. This approach significantly improved classification performance on noisy textual datasets, highlighting the utility of layer-wise representational specialization.

Cross-lingual semantic similarity tasks present additional challenges due to disparities in resource availability and structural differences across languages. Strategies leveraging multiple monolingual PLMs to independently embed sentences, followed by integration of these embeddings, have demonstrated effectiveness in capturing semantic alignments across languages. Chen et al. [10] empirically showed that utilizing pretrained English models can enhance NLP performance in low-resource languages by combining their embeddings with those from native-language models. Similarly, Zilong et al. [**?**] demonstrated a practical application for cross-lingual conversational sentence similarity between Korean and English, effectively improving both accuracy and computational efficiency by integrating multiple monolingual embeddings. These examples underscore the potential of ensemble monolingual approaches to reduce dependence on parallel corpora or expensive multilingual pretraining, benefitting multilingual and low-resource scenarios.

Fundamentally, large PLMs exhibit notable limitations including vulnerability to adversarial inputs, difficulties in compositional generalization, and challenges in interpretability [27]. Their brittleness when handling out-of-distribution or noisy linguistic phenomena highlights the insufficiency of purely statistical pattern recognition models for complex natural language understanding. Sublime [27] argues that addressing such brittleness requires hybrid neuro-symbolic architectures that integrate symbolic reasoning and explicit knowledge with learned representations. These approaches specifically target enhanced robustness and cognitive generalization in complex linguistic scenarios, seeking to overcome essential deficits in current neural models that impact interpretability and compositional capacity.

In summary, advancing robust AI systems capable of human-like language understanding requires synergizing advanced multimodal fusion techniques with strategies specialized for handling noisy and diverse textual inputs. The integration of semantic embeddings extracted from multi-layer PLMs, supported by cross-modal attention mechanisms and large-scale weakly supervised training, provides a promising architectural blueprint. Concrete case studies such as those by Kasthuriarachchy and Senanayake [9] for noisy social media text and Chen et al. [10] and Zilong et al. [?] for low-resource cross-lingual tasks demonstrate the effectiveness of these strategies in practice. Nonetheless, achieving robust generalization across noisy, multilingual, and multimodal inputs demands continued innovation to address challenges around representational alignment, model interpretability, and computational scalability [6, 9, 10, 12, 16, 24, 27? ? ? ? ? ?].

2.3 Applications of Multimodal AI and Large Language Models

Multimodal AI and large language models (LLMs) have seen extensive application across diverse domains, demonstrating substantial improvements in integrating and processing heterogeneous data types. Key application areas include healthcare, robotics, natural language understanding, and multimedia information retrieval. For instance, in healthcare, multimodal models leverage textual medical records, imaging data, and genomic information to enhance diagnostic accuracy and personalized treatment recommendations []. In robotics, the combination of language and sensory inputs facilitates more intuitive human-robot interactions and effective task execution []. Natural language understanding and generation benefit from multimodal cues such as visual context or speech signals to improve comprehension and response quality []. Furthermore, multimedia retrieval systems utilize multimodal embeddings to bridge modalities like images, text, and audio, achieving superior performance on benchmark datasets [].

These varied applications underscore the versatility of integrating modalities, where large language models serve as a robust backbone for processing and generating unified semantic representations. The synergy between large-scale pretrained language models and specialized modality encoders enables state-of-the-art performance on complex real-world tasks. Integration across modalities not only enriches information understanding but also facilitates cross-domain transfer learning, advancing benchmarks in multiple fields.

This integrative overview highlights the broad utility of multimodal AI and LLMs, emphasizing their role as foundational models driving innovation and performance enhancement across domains. The following sections build upon this foundation by exploring detailed architectures, training methodologies, and evaluation metrics that enable these advances.

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—including genetic, proteomic, clinical, imaging, and environmental information—multimodal frameworks effectively capture complex pathophysiological interactions that single-source models inadequately represent [?]. This comprehensive approach facilitates advances in 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 example of this integration is the CONCH system, which synergizes patient data with contextual clinical information to improve diagnostic accuracy and therapeutic guidance [?]. Multimodal datasets in ophthalmology exemplify this progress by 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 from 2,451 patients across international cohorts, significantly surpasses expert ophthalmologists by achieving an 83.9% top-five accuracy in predicting gene classes underlying rare inherited retinal diseases [23]. This success is primarily attributable to modality-specific Convolutional Neural Network (CoAtNet0) ensembles, which employ weighted cross-entropy loss functions and ensemble averaging to mitigate data imbalance, while Uniform Manifold Approximation and Projection (UMAP) visualizations reveal meaningful genotype-phenotype correlations that include genes unseen during training. Eye2Gene's interpretability, enhanced by attention maps highlighting image regions influencing predictions, facilitates clinical trust. Nonetheless, limitations remain, including coverage gaps toward rare genes (only 63 of approximately 281 known IRD genes included) and reliance predominantly on imaging data without additional clinical context [23].

Multimodal AI further advances surgical domains, especially within intraoperative environments where real-time recognition of surgical instruments can enhance workflow efficiency and patient safety. Evaluations of publicly available LLMs—including ChatGPT-4 and its visual-optimized variant ChatGPT-4o, as well as Google's Gemini—demonstrate promising category-level instrument recognition accuracies, with ChatGPT-4o reaching 89.1% [?]. However, fine-grained subtype identification remains substantially more challenging, dropping to approximately 33–39% accuracy. These results underscore the intrinsic difficulty of nuanced visual pattern recognition in surgical settings and suggest that hybrid retrieval-augmented generation frameworks that combine LLMs with domain-specific knowledge bases and data augmentation techniques are essential to improve performance further [?].

Underlying these technical advances are 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 [?]. Practical deployment is complicated by computational overheads and real-time operational demands. Trustworthy AI frameworks that dynamically regulate data access based on user roles and data sensitivity—integrating attribute-based and role-based access control with semantic sensitivity detection—represent promising approaches to balance privacy with information utility in healthcare LLM applications [2].

Looking forward, progress depends on curating multimodal biomedical datasets and fostering collaborative data-sharing frameworks that adhere strictly to privacy standards while enabling clinically validated AI systems. The development of pretrained

Table 2: Summary of Key Application Areas and Performance Benchmarks for Multimodal AI and Large Language Models

Application Domain	Modalities Involved	Representative Tasks	Benchmark Performance
Healthcare	Text, Imaging, Genomics	Diagnosis, Treatment Prediction	Improved accuracy over unimodal baselines
Robotics	Language, Visual, Sensor Data	Human-Robot Interaction, Control	Enhanced task completion and adaptability
Natural Language Understanding	Text, Visual	Question Answering, Dialogue Systems	Higher contextual understanding and relevance
Multimedia Retrieval	Text, Image, Audio	Cross-modal Search, Recommendation	State-of-the-art retrieval metrics

multimodal biomedical models incorporating domain-specific reasoning capabilities is crucial for creating scalable, generalizable AI solutions in medicine [**?**].

2.3.2 *Real-Time Safety and Autonomous Systems.*

Multimodal AI plays a critical role in real-time safety management for autonomous and semi-autonomous systems by integrating heterogeneous data sources to enhance situational awareness and enable timely interventions. This integration includes drone-acquired imagery, vehicular telemetry, and environmental sensor data, which are processed through convolutional neural networks combined with advanced sensor fusion techniques. Such an approach enables robust detection of traffic hazards such as congestion, accidents, and unsafe driving behaviors, achieving mean average precisions exceeding 90%, with an accuracy improvement of approximately 15% under adverse or complex environmental conditions [7].

The real-time decision-making frameworks employed blend rule-based reasoning with learning algorithms, ensuring safety alerts are generated with latencies below 200 milliseconds—an essential threshold for effective highway safety measures. This multimodal AI framework not only reduces accident risks significantly (up to 30%) but also leverages unmanned aerial vehicle (UAV) networks for dynamic and adaptive monitoring. Despite these advancements, challenges remain in managing limited data bandwidth, maintaining communication reliability within UAV and vehicular networks, and safeguarding privacy amid extensive data collection efforts [7]. Addressing the increasing complexity of traffic environments demands the incorporation of sophisticated predictive analytics and enhanced multi-agent coordination mechanisms, which constitute promising directions for future research aimed at proactive risk anticipation and mitigation in smart city infrastructures.

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), particularly in low-resource and linguistically diverse settings. For instance, integrating PLMs such as Chinese BERT into non-autoregressive (NAR) automatic speech recognition (ASR) models addresses the traditional trade-off between decoding speed and transcription accuracy. By enriching acoustic representations with the linguistic context provided by PLMs, these systems achieve character error rates (CERs) competitive with traditional autoregressive baselines (e.g., 6.9% vs. 6.5%) while maintaining lower real-time factors conducive to faster inference [24]. This approach effectively tackles challenges posed by tonal and homophonic features characteristic of Chinese speech without compromising computational efficiency.

Cross-lingual adaptation further demonstrates that pretrained English language models, when fine-tuned systematically, outperform native-language models trained from scratch on low-resource languages. This finding underscores the value of transfer learning in leveraging resource-rich English linguistic representations to improve NLP tasks in languages with limited data [10].

Moreover, geographic adaptation of PLMs through fine-tuning on curated region-specific corpora mitigates biases arising from training on predominantly North American and European English data. This geographically informed fine-tuning substantially enhances model performance, as evidenced by F1 score improvements exceeding 4 points and meaningful reductions in perplexity and error rates for region-specific lexical and syntactic phenomena [**?**]. Table 3 summarizes these gains across diverse English variants.

Together, these developments highlight the crucial role of linguistic diversity and geographic variation in designing robust and equitable NLP systems. Future research directions include integrating geographic factors with social and cultural dimensions to further improve fairness, robustness, and representation in pretrained language models.

2.3.4 *Text Generation and Emotion Recognition.*

Text generation frameworks powered by pretrained language models (PLMs) cover diverse paradigms including open-ended, conditional, and controllable generation tasks, each with unique challenges concerning output coherence, diversity, and ethical considerations [**?**]. Advanced reinforcement learning (RL) techniques, especially Reinforcement Learning from Human Feedback (RLHF) and Proximal Policy Optimization (PPO), have been employed to better align model outputs with human preferences, thereby enhancing multi-step reasoning and reducing undesirable biases [**?**]. Nonetheless, these approaches still face hurdles such as sample inefficiency, complex reward design, and safety issues. These limitations motivate the exploration of hybrid strategies that blend symbolic reasoning with hardware acceleration, aiming to achieve improved scalability, interpretability, and robustness.

In emotion recognition, integrating PLMs with deep neural architectures has markedly advanced the field beyond traditional single-label classification towards nuanced multi-label emotion detection. By extracting contextualized embeddings through fine-tuned PLMs and employing a sigmoid-activated output layer, these models substantially outperform conventional baselines, yielding macro F1-score improvements of 5–7% [22]. This methodology adeptly captures overlapping emotional states such as joy, sadness, and anger, and enhances model interpretability. Persistent challenges include mitigating data imbalance and discriminating semantically similar emotions, which encourage future research to

Table 3: Improvements in F1 scores from Geographic Adaptation of PLMs on Regional English Variants [?]

Region	Base Model F1	Adapted Model F1	Improvement
African English	72.4	77.1	+4.7
Indian English	70.9	75.8	+4.9
Caribbean English	68.3	72.6	+4.3

integrate multimodal inputs and develop explainability techniques to further boost performance and transparency.

Collectively, these developments highlight a comprehensive landscape where multimodal AI techniques and large language models significantly advance both predictive and generative tasks. They concurrently emphasize crucial ethical, privacy, and fairness considerations essential for responsible deployment. The broad applicability of these technologies across healthcare, autonomous systems, and natural language domains underlines their transformative impact on contemporary AI research and practical applications.

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

Explainable AI (XAI) emphasizes enhancing the transparency and interpretability of AI systems, which is essential for fostering user trust and supporting informed decision-making processes. Trustworthiness in AI involves multiple dimensions, including reliability, robustness, fairness, and accountability, ensuring systems perform reliably across diverse and unpredictable real-world scenarios. Ethical considerations address critical issues such as bias mitigation, privacy preservation, and the broader societal implications of AI deployment. The integration of XAI methods contributes significantly to trustworthiness by elucidating model decisions, enabling users and developers to identify and address potential biases or unfair outcomes. Collectively, these aspects establish a foundation for responsible AI development and deployment, guiding researchers and practitioners to create AI systems that are not only effective but also adhere to human values and societal norms, thereby promoting equitable and ethical technology adoption.

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 signifies a substantial paradigm shift away from reliance 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 instance, environmental monitoring applications that combine multimodal sensor inputs with explainable gradient boosting and attention mechanisms achieve not only significant accuracy gains but also notably enhanced interpretability metrics, as summarized in Table 4 [21]:

These findings underscore that combining interpretability tools with stringent data quality management, provenance tracking, and compliance auditing promotes transparency and user trust. Nonetheless, balancing increased model complexity with user comprehensibility remains challenging, particularly when handling high-dimensional heterogeneous data.

Moreover, addressing noise 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, such as context-aware transformer frameworks, which help constrain model complexity and enhance explanation precision [13]. Transformer-based models leveraging context-aware self-attention mechanisms explicitly incorporate diverse contextual information (e.g., temporal, device, location) to capture dynamic feature interactions, thereby improving explanation fidelity while maintaining computational efficiency.

However, disentangling contributions of intertwined features while preserving interpretability remains a critical and unresolved challenge. This emphasizes the need for standardized benchmarks to rigorously assess explanation faithfulness and incorporate human-grounded evaluation methodologies.

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.

Table 4: Performance comparison demonstrating the effectiveness of explainable AI integrated with data governance in multimodal sensor classification [21].

Model	Accuracy (%)	Explanation Fidelity
Baseline Deep Neural Network (DNN)	87.2	0.65
Proposed XAI-Governed Model	92.1	0.87

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 remains problematic and requires continual adaptation of privacy parameters to comply with dynamic data governance policies. While machine learning–based trust modeling offers increased adaptability, emphasizing transparency and auditability is essential to prevent opaque, inscrutable decision-making processes [2]. Furthermore, systemic concerns persist regarding transparency, reproducibility, and intellectual property rights in black-box model access, particularly within Language-Models-as-a-Service (LMaaS) paradigms. Proprietary LMaaS models severely limit user visibility into internal operations [11], complicating benchmarking and evaluation efforts. Addressing these challenges requires standardized benchmarking protocols for black-box models, enhanced 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 rooted in their training data and architectural paradigms. Documented risks include inadvertent memorization of sensitive or private content, systemic biases reflecting underlying data distributions, propagation of toxic or false information, and model failure modes when exposed to adversarial inputs [3]. These vulnerabilities complicate efforts to deploy LLMs responsibly in sensitive or high-stakes contexts.

Addressing these complexities requires a mechanistic interpretability framework that surpasses superficial behavioral benchmarks. This framework entails detailed analyses of internal model representations and decision pathways to ensure fairness, transparency, and safety in model design and deployment [2, 3, 25, 28**?**]. For example, trust frameworks embedding dynamic user profiling and adaptive output control can regulate sensitive data disclosure according to contextual trust levels, as demonstrated in [2]. This approach integrates Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) with Information Sensitivity Detection using Named Entity Recognition enhanced by domain-specific dictionaries and semantic analysis, achieving privacy-aware responses with minimal latency overhead. In addition, interpretability efforts within multimodal biomedical AI emphasize the integration of diverse data types while preserving ethical considerations and privacy, leveraging advances in fusion strategies, privacy-preserving techniques, and rigorous clinical validation [**?**]. Similarly, multimodal explainable AI research focuses on aligning explanations with human cognition and mitigating biases through causal and counterfactual reasoning frameworks, addressing challenges unique to heterogeneous multimodal data [25]. Achieving such mechanistic rigor confronts practical hurdles including the high computational and data costs of interpretability analyses and balancing multitask learning objectives without sacrificing robustness [20].

Emerging research directions aim to improve computational efficiency, develop multimodal grounding for enhanced language understanding, and bolster robustness against adversarial and distributional shifts. Crucially, these technical advances must be embedded within ethical frameworks prioritizing inclusivity, transparency, and user trust. Cross-disciplinary insights—from linguistics, cognitive science, and social sciences—are vital to address biases and toxicity at their origins, ensuring that challenges are tackled through both technical innovation and principled social perspectives [3, 28]. Through such integrated approaches, melding technical progress with ethical guidelines, AI systems can evolve into genuinely trustworthy, reliable, and resilient technologies.

2.4.4 Early Integration of Human Preferences in Large Language Models. A critical advancement in improving large language model (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). The primary objective of this approach is to produce models that internalize human-aligned behaviors intrinsically, thereby reducing undesirable outputs like toxicity and hallucinations while maintaining or improving language modeling capabilities [4].

Recent work by Korbak et al. [4] introduces a multi-objective pretraining framework that incorporates pairwise human preference judgments directly into the model's learning objective. Specifically, this framework optimizes both the standard next-token prediction loss and a preference ranking loss simultaneously. The preference objective is based on a Bradley-Terry model likelihood, formulated as:

$$\mathcal{L}_{pref} = -\sum_{(x_i,x_j)} y_{ij}\log\sigma(s_i - s_j) + (1 - y_{ij})\log\sigma(s_j - s_i),$$

where s_i and s_j are the preference scores predicted by the model for inputs x_i and x_j, respectively; y_{ij} is an indicator variable representing the preferred instance; and σ denotes the sigmoid function. This formulation encourages the model to assign higher scores to more preferred outputs according to human judgments, effectively embedding aligned behavior inside the model parameters rather than treating it as an external corrective layer.

Human preference data collection protocols in this context are based on crowdsourced or expert comparisons of model-generated text pairs, labeled according to human judgments reflecting desirable qualities such as factual accuracy, helpfulness, and toxicity avoidance. Although currently limited by the scalability and cost of acquiring diverse, high-quality preference datasets, ongoing efforts aim to develop more efficient scalable protocols for preference collection that leverage active learning or synthetic augmentation [4].

Table 5 presents key experimental results comparing traditional language model pretraining, fine-tuning with preference data, and joint pretraining using preference integration. The joint pretraining approach achieves a notable reduction in toxicity rate and higher preference accuracy, with only a slight increase in perplexity, reflecting a balanced enhancement in both language modeling and alignment metrics.

Beyond text-only models, the paradigm of early human preference integration holds potential applicability to multimodal architectures, where various modes such as vision and language generate outputs subject to human evaluation. Extending preference-based pretraining objectives to these broader contexts could help align complex AI systems more deeply to human values and expectations. Additionally, hybrid frameworks that combine this early-stage preference embedding with subsequent selective fine-tuning or RLHF may further improve sample efficiency and robustness in alignment [4].

Despite promising advances, significant challenges remain, primarily related to the cost and scarcity of rich preference data and the inherent difficulty in balancing multiple learning objectives during large-scale pretraining. Future research directions include establishing scalable data collection pipelines, refining multi-objective training algorithms that mitigate conflicts or trade-offs between objectives, and exploring generalization of preference-infused models to downstream tasks and domains.

Overall, integrating human preferences early in LLM pretraining represents a foundational step toward building more trustworthy, safe, and aligned AI systems by embedding human values directly within model parameterization rather than relying on costly corrective interventions.

3 Behavioral, Cognitive, and Neuroscientific Insights

This section synthesizes key behavioral, cognitive, and neuroscientific findings that illuminate the alignment and divergence between artificial intelligence models—particularly large language models (LLMs) and multimodal large language models (MLLMs)—and human cognition. We focus on the ways these models capture human conceptual representations, linguistic capabilities, and output consistency, while highlighting challenges and implications for future AI development and ethical deployment.

3.1 Representational Similarity Analysis and Conceptual Alignment

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 modalities. Studies employing RSA demonstrate that embeddings generated by LLMs and 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 fusiform face area (FFA) [6**?**]. For instance, Du et al. [6] analyzed 4.7 million human similarity judgments on nearly 2,000 natural objects, deriving sparse, non-negative embeddings via the Sparse Positive Similarity Embedding (SPoSE) method. These embeddings predict human choice behavior and exhibit semantic clustering that parallels human conceptual structures, capturing semantic categories like animals and food as well as perceptual features such as hardness and texture. Multimodal models further enhance this alignment by representing spatial and color information, underscoring the importance of rich sensory data in forming internal representations analogous to those in the brain [**?**].

Mahner et al. [**?**] found that alignment varies by representation level and model architecture. Early layers in deep neural networks predominantly encode low-level visual features (e.g., shape and texture), while higher layers represent abstract semantic categories, mirroring hierarchical processing stages in human visual cognition. However, some representational dimensions unique to human cognition remain underrepresented in current models. These gaps reflect challenges arising from dataset biases and the complex integration of visual and semantic cues. Addressing these gaps requires future research focused on embedding richer multimodal context and augmenting behavioral data to close representational discrepancies [**? ? ?**]. Potential directions include the creation of datasets that better capture human-unique cognitive features, improved multimodal fusion techniques that integrate semantics and perception more effectively, and interdisciplinary frameworks combining cognitive neuroscience, psychology, and AI modeling.

3.2 Multimodal Context, Ethical Considerations, and Model Architecture

Enriched multimodal context involves both architectural innovations and the nature of training datasets and behavioral signals guiding representation learning. Empirical evidence indicates that training on multimodal datasets reflecting human perceptual and conceptual experiences substantially improves AI models' capacity to capture human-like representations [**? ?**]. For example, MLLMs in healthcare integrate diverse data types such as images, audio, and text, enhancing clinical decision-making despite challenges related to data fusion complexity, interpretability, and ethical concerns [**? ?**]. These challenges include computational demands, modality alignment, and handling bias, privacy, and safety issues.

Interdisciplinary integration not only enhances interpretability and robustness but also informs ethical frameworks essential for developing artificial general intelligence (AGI) aligned with human cognitive principles [**? ? ?**]. Ethical deployment mandates addressing biases, ensuring privacy, safety, and informed consent in AI

Table 5: Performance comparison of language models trained with and without early integration of human preferences [4].

Method	Perplexity	Toxicity Rate	Preference Accuracy
Standard LM pretraining	12.3	0.15	0.50
Fine-tuning on preferences	12.5	0.10	0.68
Joint pretraining with preferences	12.7	0.07	0.75

systems that reflect human cognition and societal norms. Future research should emphasize dynamic ethical alignment mechanisms, improved transparency in multimodal fusion, and the development of domain-specific guidelines to responsibly harness the power of MLLMs.

3.3 Linguistic and Cognitive Evaluations of Pretrained Language Models

Complementing neuroscientific insights, cognitive and linguistic evaluations of pretrained language models (PLMs) reveal both strengths and limitations across syntactic, semantic, and reasoning tasks prior to fine-tuning. Chang and Bergen [3] provide a comprehensive survey showing PLMs robustly handle fundamental syntactic rules (e.g., subject-verb agreement), semantic compositionality, analogical reasoning, and basic logical inference. Conversely, they struggle with complex syntactic phenomena, negation, implicit pragmatics, and multi-step inferencing, evidencing incomplete linguistic and conceptual comprehension. These linguistic limitations resonate with neuroscientific findings of incomplete semantic alignment, indicating current architectures and pretraining approaches inadequately capture the full complexity of human linguistic cognition.

Future work should aim to enhance models' understanding of nuanced linguistic and pragmatic phenomena through more sophisticated training objectives, enriched multimodal grounding, and learning from richer behavioral signals. Interdisciplinary approaches bridging linguistics, cognitive science, and machine learning are vital to push beyond current limitations.

3.4 Assessing and Improving Model Consistency

Systematic assessments of model consistency emphasize challenges in reliability and interpretability. Wang et al. [28] developed benchmark datasets evaluating pretrained models such as GPT-2, BERT, RoBERTa, and GPT-3 across factual, paraphrase, and negation consistency dimensions. Results show factual consistency often falls below 80%, undermining trustworthiness in critical applications. Detailed analyses attribute inconsistencies mainly to miscalibrated confidence rather than fundamental knowledge deficits.

Temperature scaling, a post hoc probabilistic calibration method, improves consistency metrics by up to 20% without altering model weights [28]. This highlights the importance of integrating uncertainty quantification and calibration mechanisms to enhance robustness and practical deployment. Future research directions include developing more advanced calibration techniques, incorporating uncertainty during training, and designing benchmarks that evaluate model consistency under real-world conditions.

In summary, convergent evidence indicates that while AI models increasingly approximate human-like conceptual embeddings and linguistic competence, significant limitations remain in semantic depth, multimodal integration, and output consistency. Overcoming these limitations requires interdisciplinary frameworks bridging behavioral science, neuroscience, computational modeling, and linguistic theory. Such integrative approaches are critical to advancing AI systems whose internal representations and outputs more faithfully reflect the complexity and richness of human cognition and communication, and to ensuring their responsible and ethical application [3, 28? ? ?].

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

Retrieval-augmented language models (LMs) enhance their natural language processing capabilities by leveraging external knowledge sources during inference, enabling more accurate and contextually relevant responses. A core challenge in these systems is managing retrieval noise—irrelevant or erroneous information retrieved—which can degrade model outputs. To mitigate this, techniques such as relevance re-ranking of retrieved documents, confidence-based retrieval filtering, and integrating retrieval results within robust attention mechanisms have been explored. For example, re-ranking helps prioritize higher-quality information, reducing noise impact on the LM's predictions.

Combining retrieval augmentation with other model advances, such as prompt tuning or multi-task learning, further strengthens adaptability and performance. Retrieval offers dynamically updated knowledge, while other adaptation methods refine model behavior across diverse tasks, enabling synergistic effects. A practical illustration involves first retrieving region-specific encyclopedic data and then applying geographic prompt tuning to produce locally nuanced language generation. This layered approach not only improves factual accuracy but also aligns stylistic and cultural relevance.

In geographic adaptation, language models are tailoring outputs to reflect specific regional language use, dialects, and local knowledge. Beyond purely technical adaptation, the social impact of geographic customization calls for careful consideration. Adapting to geographic context can help reduce biases, improve inclusivity, and foster better user trust by respecting linguistic diversity. However, risks include reinforcing stereotypes or marginalizing minority dialects if adaptations are not carefully designed and monitored. Thus, multidimensional adaptation frameworks that incorporate societal metadata—such as socio-economic status, urban versus rural distinctions, or cultural identifiers—are critical. These frameworks enable nuanced adaptation beyond geography alone, ensuring that models respond sensitively to complex social variables.

To illustrate multidimensional adaptation, consider augmenting a language model with metadata tags representing geography,

Table 6: Summary of Key Techniques, Challenges, and Directions in Aligning AI Models with Human Cognition

Aspect	Key Techniques & Findings	Challenges & Future Directions
Representational Similarity Analysis (RSA)	Sparse Positive Similarity Embedding, hierarchical feature representation across layers [6?]	Address human-unique representational dimensions, overcome dataset biases, enrich multimodal context [? ? ?]
Multimodal Large Language Models (MLLMs)	Integration of vision, audio, text; multimodal fusion and contrastive learning [? ?]	Data fusion complexity, computational demands, interpretability, ethical concerns (bias, privacy) [? ?]
Linguistic and Cognitive Evaluations of PLMs	Analysis of syntactic, semantic, reasoning capabilities [3]	Improve handling of complex syntax, negation, pragmatics, multi-step reasoning [3]
Consistency Assessments	Benchmark datasets for factual, paraphrase, negation consistency; probabilistic calibration [28]	Enhance output stability and trustworthiness via calibration and uncertainty quantification [28]
Ethical AI Deployment	Integration of cognitive principles, bias mitigation, privacy safeguards [? ? ?]	Develop frameworks respecting human cognition and societal values for AGI [? ? ?]

socio-cultural attributes, and temporal context during both retrieval and generation. This can guide retrieval components to prioritize regionally and socially relevant documents, while prompting mechanisms adjust output style accordingly. Such integration demands transparent ethical guidelines and evaluation protocols to balance personalization with fairness.

In summary, recent advances point toward cohesive strategies combining robust retrieval-augmented architectures and sophisticated geographic and societal adaptation mechanisms. These efforts are essential to harness the full potential of language models in producing accurate, context-aware, and socially responsible language generation.

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 computes a retrieval query vector $q_t = W_q h_t$ derived from the decoder's hidden state h_t, which is then used to attend over a memory bank composed of past hidden states. This memory bank is projected into keys $k_i = W_k m_i$ and values $v_i = W_v m_i$, where m_i are stored memory vectors. The model retrieves salient information via scaled dot-product attention weights $\alpha_{t,i} = \frac{\exp(q_t^\top k_i/\sqrt{d})}{\sum_j \exp(q_t^\top k_j/\sqrt{d})}$, computing the retrieved vector $r_t = \sum_i \alpha_{t,i} v_i$ that is integrated into the token prediction. 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, achieving a perplexity of 13.7 compared to 15.3 for Transformer-XL and 17.8 for fixed-window transformers [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, which can propagate errors into generated outputs and undermine quality. Additionally, scalable indexing strategies are required for multi-document retrieval to extend the model's applicability beyond single documents. Naïve memory storage approaches prove prohibitively expensive at large scales, prompting the need for advanced memory selection techniques. Integration with external knowledge bases also remains a critical direction for future research to further enhance retrieval quality and grounding [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 tokenizing continuous numerical properties into sequences that preserve decimal ordering, thereby instilling an inductive bias favoring numerical proximity. It employs numerical encodings and an alternating training scheme combining permutation language modeling, property prediction, and conditional generation objectives. 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. It achieves superior performance on benchmark datasets such as MoleculeNet, characterized by high novelty (over 99

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. For instance, adaptation raised African English F1 scores from 72.4 to 77.1, Indian English from 70.9 to 75.8, and Caribbean English from 68.3 to 72.6 [?]. 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 not only advances technical performance but also promotes inclusive language technologies that recognize and respect diverse linguistic identities [?].

4.3 Synthesis and Outlook

Recent advancements in retrieval-augmented transformer architectures, multitask scientific foundation models, and geographic adaptation strategies collectively signify a critical shift toward more context-aware, precise, and equitable language modeling paradigms. These innovations move beyond traditional static and monolithic designs, enabling dynamic and multifaceted systems that facilitate nuanced understanding across diverse domains and demographic contexts. This evolution highlights not only technical progress but also a growing commitment to addressing real-world challenges through improved adaptability, fairness, and inclusivity. As large-scale language models continue to integrate heterogeneous data sources and adapt to varied environments, future research is poised to further enhance their robustness and applicability, fostering more reliable and socially responsible AI systems.

5 Challenges and Future Directions

This survey aims to provide a comprehensive overview of the current landscape in multimodal AI, focusing on the integration of diverse data modalities to achieve more robust, flexible, and human-like intelligence. The objectives are to identify the primary technical and conceptual challenges faced by this field and to critically evaluate the prevailing methods addressing these issues. By clarifying these goals, we establish a clear scope that guides the subsequent discussion.

The field of multimodal AI faces several significant and interrelated challenges that hinder its advancement. First, the effective fusion of heterogeneous data sources—such as vision, language, audio, and sensor data—remains difficult due to differing data structures, representation scales, and temporal dependencies. For instance, aligning spoken language with corresponding visual scenes demands models that can capture complex, fine-grained correlations across modalities.

Second, there is the challenge of data scarcity and imbalance: multimodal datasets are often limited in size or biased toward certain modalities or contexts, which affects model generalization and fairness. Moreover, interpreting and explaining multimodal models' decisions are complicated by their high-dimensional and cross-modal nature, raising concerns about transparency and trustworthiness.

Third, multimodal architectures must handle variability in data quality, missing modalities, and dynamic modality availability in real-world scenarios. Designing models that gracefully degrade or adapt in such conditions is an open problem.

Lastly, ethical and societal implications, including privacy concerns, bias amplification, and accountability, require systematic frameworks to guide responsible multimodal AI development and deployment.

This section dissects these core challenges with concrete examples and contrasts competing methodological approaches where relevant. For instance, we compare early fusion techniques, which integrate raw features, with late fusion strategies that combine modality-specific decisions. We also examine emerging approaches such as attention-based cross-modal transformers, evaluating their strengths and limitations in handling multimodal interactions.

Going forward, advancing multimodal AI demands interdisciplinary research to improve theoretical foundations, data resource development, evaluation metrics, and ethical guidelines. Taking a holistic view that bridges technical innovation with societal impact will be essential to unlock the full potential of multimodal intelligence.

5.1 Integration of Heterogeneous Modalities

Effectively combining information from diverse data types—such as text, images, audio, and video—remains a fundamental challenge in multimodal learning. Each modality exhibits distinct structural and representational characteristics, requiring models to bridge semantic gaps and contextual differences for meaningful fusion. A key difficulty lies in aligning semantic content across modalities, for example, mapping the abstract concepts expressed in text to the visual or auditory features present in images and audio streams. This necessitates advanced strategies for learning joint representations that are both rich in informative content and robust enough to generalize across varied tasks and domains. Approaches often involve designing shared embedding spaces or leveraging attention mechanisms to capture cross-modal interactions while preserving modality-specific nuances. Despite progress, achieving seamless integration that fully exploits complementary information remains an open research frontier.

5.2 Scalability and Efficiency

Multimodal models often consist of large and complex architectures that incur significant computational costs, rendering them resource-intensive for practical deployment. Real-world applications, such as real-time language translation involving video feeds, require models that are both lightweight and maintain high accuracy. Addressing this challenge involves developing strategies to balance model performance with computational efficiency, including model compression, efficient architecture design, and adaptive computation. These remain active areas of research aimed at enabling scalable and efficient multimodal systems suitable for deployment in resource-constrained environments.

5.3 Data Scarcity and Quality

While single-modal datasets are abundant, high-quality, large-scale, and well-annotated multimodal datasets remain limited. For example, datasets that include paired speech and gesture data with detailed semantic annotations are rare, which significantly hinders

Table 7: Performance improvements from geographic adaptation of pretrained language models on underrepresented English variants.

Region	Base Model F1	Adapted Model F1	Improvement
African English	72.4	77.1	+4.7
Indian English	70.9	75.8	+4.9
Caribbean English	68.3	72.6	+4.3

progress in this area. Moreover, ensuring the precise temporal alignment and consistency of multimodal data involves complex challenges, such as synchronizing diverse data streams and maintaining annotation quality across modalities. Addressing these issues is critical to advancing robust multimodal research and applications.

5.4 Interpretability and Explainability

Understanding how multimodal models make decisions is essential, especially for applications in healthcare and autonomous systems. Unlike unimodal models, the complexity of interactions between modalities complicates interpretability. Developing techniques that provide clear explanations about how individual modalities contribute to final predictions remains an important goal. Current methods focus on disentangling modality-specific contributions and revealing cross-modal influences to offer more transparent and trustworthy insights. Such approaches aim to enhance user trust and aid debugging by clarifying the decision-making process within complex multimodal architectures.

5.5 Robustness and Generalization

Models must perform reliably under noisy or missing modalities to be practically useful in real-world scenarios. For example, a video classification system should continue to function adequately even if the audio track is corrupted or absent. Achieving robustness involves explicitly addressing challenges such as modality dropout, where one or more modalities might be partially or entirely unavailable during inference. Additionally, models must generalize well across domain shifts that occur due to changes in environment, sensor characteristics, or data distribution variations. Ensuring consistent performance across these varied scenarios is critical for deploying multimodal systems in dynamic and unpredictable settings.

5.6 Future Research Directions

To advance the field, researchers should focus on several key areas. First, designing unified frameworks capable of flexibly handling multiple modalities is crucial; such frameworks should dynamically adapt to the presence or absence of specific inputs to improve robustness and applicability across diverse scenarios. Second, there is a need for creating comprehensive benchmarks that encompass diverse and challenging multimodal tasks, accompanied by standardized evaluation protocols to facilitate fair comparison and reproducibility. Third, investigating transfer learning approaches that leverage the strengths of large unimodal models for multimodal tasks can yield significant performance gains while reducing training costs. Fourth, developing interpretability tools specifically tailored for multimodal architectures is essential to understand complex interactions across modalities and foster trust in these systems. Finally, exploring energy-efficient model designs suitable for deployment on edge devices will enable practical use in resource-constrained environments, addressing the growing demand for real-time and privacy-preserving applications.

Summary of Key Challenges

This table succinctly encapsulates the fundamental challenges inherent in multimodal AI research and development, linking each challenge with representative examples and clear research objectives. Addressing these challenges through focused strategies will be essential to advancing multimodal AI toward more effective, interpretable, and real-world deployable systems.

5.7 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? ? ? ?]. The challenge is exacerbated by underrepresentation and bias linked to geographic, cultural, and societal factors, reinforcing the critical need for expanding datasets with balanced and diverse representation. Ethical and fairness considerations further demand the development of equitable benchmarking protocols that evaluate model performance inclusively across different demographic and linguistic groups [7? ? ?].

Moreover, in specialized domains such as healthcare and biomedical applications, multimodal AI benefits from integrating heterogeneous data types—ranging from medical imaging and genomic information to clinical notes and environmental data—yet faces pronounced challenges in data harmonization, privacy preservation, and interpretability [? ?]. These domain-specific complexities highlight the importance of curated high-quality datasets alongside privacy-preserving approaches such as federated learning and differential privacy to foster trustworthy and effective models.

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, 8? ? ?]. Recent innovations such as sparse attention mechanisms, kernel-based approximations, and memory/recurrence strategies aim to alleviate these limitations [8]. For

Table 8: Summary of Major Challenges in Multimodal AI

Challenge	Description	Example	Research Objective
Integration	Combining heterogeneous data types with different structures and modalities	Aligning text semantics with visual features	Develop joint embedding methods that effectively capture cross-modal relationships
Scalability	Managing high computational cost and memory usage in complex multimodal systems	Real-time multimodal translation systems requiring low latency	Design efficient architectures that balance accuracy with computational speed and resource constraints
Data Scarcity	Limited availability of large-scale, well-annotated datasets spanning multiple modalities	Lack of paired speech-gesture semantic datasets for training	Create novel datasets and advanced data augmentation techniques to improve generalization
Interpretability	Challenges in explaining the decision processes involving multiple interdependent modalities	Understanding model outputs in critical domains such as healthcare diagnostics	Develop robust explainability and interpretability techniques that accommodate cross-modal reasoning
Robustness	Ensuring model performance despite noisy, incomplete, or missing modality data during inference	Video classification tasks dealing with missing or corrupted audio input	Build models resilient to missing, noisy, or corrupted inputs to maintain reliable performance

example, Sparse Mixture-of-Experts models like the Switch Transformer employ expert routing, activating only subsets of parameters per input token, thus significantly reducing computational overhead while facilitating scaling to trillion-parameter models [5, 26]. These models combine architectural sparsity with load balancing losses to maintain model expressiveness and prevent expert underutilization. Nevertheless, they require careful hyperparameter tuning and introduce new challenges concerning training stability and efficiency.

Beyond architectural innovations, alternative approaches such as adopting non-Euclidean geometric spaces — e.g., hyperbolic embeddings — show promise in encoding complex semantic and syntactic relationships more succinctly, potentially improving model efficiency without increasing parameter or data requirements [5]. Addressing the intertwined issues of data scarcity and computational constraints therefore necessitates an integrated strategy: curating diverse, representative and privacy-aware datasets, adopting efficient transformer variants and hybrid architectures, and leveraging domain-specific optimization and cross-modal learning techniques. Such holistic approaches are essential to enhance both the performance and scalability of multimodal and multilingual models deployed in real-world heterogeneous environments.

5.8 Interpretability, Ethics, and Safety

Ensuring interpretability and ethical compliance in multimodal AI systems is an urgent and complex challenge, particularly as these models increasingly influence sensitive sectors such as healthcare, finance, and safety-critical domains [2, 25**?**]. The inherent opacity of large models complicates understanding their internal decision-making processes, which calls for 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 provide integrated explanations across modalities, utilizing approaches like causal inference and counterfactual reasoning to harmonize model rationales with human cognitive expectations [2, 25]. However, the heterogeneity of multimodal data and challenges introduced by fusion layers create substantial obstacles in producing explanations that are both faithful and unbiased [16].

Ethical considerations impose rigorous requirements to mitigate bias, preserve privacy, secure informed consent, and comply with evolving regulatory frameworks [2, 25**?**]. Emerging frameworks that embed dynamic trust mechanisms have shown promise by balancing information disclosure against privacy preservation, employing adaptive controls guided by user trust profiles and assessments of data sensitivity [3]. For example, trust profiling combines role-based and attribute-based access control to assign dynamic trust scores based on credentials, behavior, and context, while sensitivity detection leverages domain-specific semantic analyses, such as Named Entity Recognition enriched with specialized dictionaries, to identify sensitive information with high accuracy [2]. Adaptive output control mechanisms modulate the detail of model responses using techniques including differential privacy, redaction, and summarization to balance information utility and privacy [2]. Despite such advances, real-world deployment necessitates robust safeguards against misuse, fairness violations, and harmful outcomes, underscoring the importance of continuous monitoring and transparent accountability mechanisms [3**?**]. Therefore, progress in ethical safeguards and interpretability research must advance in parallel with technical model developments to realize responsible AI systems aligned with societal values.

5.9 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 using representational similarity analysis demonstrate alignment of multimodal embeddings with neural representations localized in category-selective brain regions, underscoring the potential for cognitive-inspired architectures to enhance semantic and perceptual grounding [**?**]. Complementing these efforts, trustworthiness frameworks embedding dynamic, context-aware privacy controls within models address critical ethical and security challenges associated with sensitive data handling, especially in high-stakes domains such as healthcare [2].

In parallel, sparse mixture-of-experts models, such as Switch Transformers, have introduced efficient scalability for trillion-parameter architectures by activating only a single expert per token, significantly reducing memory and computational overhead while preserving or improving performance in cross-lingual and zero-shot tasks [26]. Switching mechanisms optimize expert utilization through learned gating and load balancing, enabling stable training and superior results compared to dense counterparts. These advances point to practical pathways for scaling large models without prohibitive resource demands.

Additionally, models incorporating dynamic retrieval mechanisms enable effective long-range context modeling beyond fixed-length inputs, supporting improved zero-shot retrieval-augmented generation and multi-document integration, which enhances coherence and factual consistency in downstream applications [17]. The Retrieval-Pretrained Transformer exemplifies this by leveraging self-retrieval over past hidden states with attention-based weighted combinations, yielding better long-range perplexity and downstream task performance. Challenges remain in memory indexing scalability and mitigating retrieval noise, yet such retrieval-augmented models lay the groundwork for adaptive, document-level understanding and generation.

Research into temporal dynamics and the development of lightweight, transformer-based frameworks tailored for real-time applications also address critical operational constraints in dynamic environments. For instance, end-to-end transformer frameworks for facial action unit detection utilize global self-attention to bypass traditional landmark dependence, achieving state-of-the-art accuracy with improved robustness and computational efficiency [14]. Such frameworks demonstrate balance between accuracy and resource constraints needed for deployment in real-time, interactive systems.

Overall, these advances highlight a trend towards building AI systems that are not only more capable across diverse modalities and languages but also more interpretable, efficient, ethically aligned, and cognitively grounded.

5.10 Integration and Multidisciplinary Collaboration

The progression of scalable sparse transformer architectures, exemplified by innovations such as the Switch Transformer, necessitates integration with advanced frameworks addressing explainability, trustworthiness, and privacy to foster transparent and secure AI systems [26]. For instance, trust mechanisms embedded within large language models dynamically manage sensitive data disclosure based on user trust profiles and data sensitivity, employing modules like Role-Based and Attribute-Based Access Control alongside semantic sensitivity detection to ensure privacy compliance while preserving utility [2]. This framework incorporates dynamic trust scoring through User Trust Profiling, semantic and domain-specific sensitivity detection using Named Entity Recognition enhanced with contextual analysis, and adaptive output control strategies including differential privacy, redaction, and summarization to balance information utility with privacy [2]. Such synergy enables models that operate efficiently at scale while adhering to ethical standards and legal regulations.

Given the complexity of both technical challenges and ethical considerations, collaboration across diverse sectors—including AI researchers, healthcare practitioners, ethicists, linguists, and security specialists—is essential to tailor AI deployments appropriately and safely across domains [2, 3, 26**?**]. Multidisciplinary engagement supports the development of domain-specific standards and best practices, such as multimodal data integration strategies in biomedical applications that enhance personalized medicine while ensuring privacy and interpretability [**?**]. These strategies leverage innovative fusion techniques for diverse data types—genetic, proteomic, imaging, clinical, and environmental—alongside privacy-preserving approaches like federated learning and homomorphic encryption, addressing challenges of data harmonization, interpretability, and scalability [**?**]. Additionally, guidelines are imperative for managing language model behavior to mitigate biases, misinformation, and privacy risks, informed by comprehensive behavioral and mechanistic analyses that cover linguistic, semantic, and ethical dimensions [3]. Collectively, these collaborative efforts promote responsible AI adoption in critical sectors such as healthcare, finance, and education, ensuring robust, interpretable, and ethically aligned deployments.

5.11 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 biomedical AI, for instance, integrating multimodal data types such as biobank records, medical imaging, wearable sensors, and multi-omics sequencing facilitates a holistic understanding of human health and disease [**?**]. Innovative transformer architectures enable effective fusion of heterogeneous biomedical data, advancing personalized medicine, digital clinical trials, and remote patient monitoring while addressing challenges of data harmonization, interpretability, privacy, and scalability [**?**].

Within biomolecular AI, promising directions include coupling structural bioinformatics with class II Human Leukocyte Antigen (HLA) prediction models—specifically transformer-based peptide-HLA binding predictors—which enhance vaccine design and immunotherapy development [**?**]. Automated mutation optimization pipelines utilizing transformer attention to improve binding affinity predictions open avenues for iterative experimental validation and model refinement [**?**]. This integration illustrates the capability of multimodal approaches to push the technical boundaries of domain-specific applications while maintaining focus on ethical imperatives and equitable data representation.

Moreover, in domains such as surgical instrument recognition, multimodal AI demonstrates potential for improving workflow efficiency and patient safety through robust category-level instrument identification, though challenges remain in fine-grained subtype detection requiring expanded datasets and specialized training [**?**]. Similarly, multimodal AI frameworks integrating visual, telemetry, and environmental sensor data can enhance real-time highway safety management, delivering dynamic hazard detection and timely interventions [7].

In natural language processing and speech recognition, multimodal and retrieval-augmented transformer models show advances in handling long-range dependencies and improving recognition accuracy in low-resource and multilingual scenarios, benefiting from pretrained language models augmented with domain- and region-specific adaptations [10, 17, 24**?**]. Geographic adaptation of pretrained language models has demonstrated measurable improvements—for example, gains of 4–5 F1 points in African, Indian, and Caribbean English variants—highlighting the importance of including culturally and regionally diverse data [**?**]. Additionally, cross-lingual adaptation using pretrained English models has proven advantageous in low-resource languages, further underscoring the impact of transfer learning and data diversity [10]. In speech recognition, integrating pretrained language models enhances accuracy and efficiency, notably in non-autoregressive frameworks for languages such as Chinese [24]. Such domain-tailored multimodal AI underscores the importance of combining architectural innovation with rigorous dataset curation and geographic-cultural inclusivity to achieve robust, generalizable, and ethically sound AI systems.

5.12 PLM-Specific Innovations and Challenges

Pretrained language models (PLMs) face significant challenges in processing noisy and informal textual inputs such as social media content. Recent studies have demonstrated that leveraging layered BERT-based representations, which capture diverse linguistic features across different model layers, can effectively improve the understanding of non-standard language use [9]. Specifically, initial and intermediate BERT layers have been shown to better encode linguistic characteristics of noisy texts, enhancing classification performance on tasks involving such inputs. Kasthuriarachchy and Senanayake [9] introduced five new probing tasks for Tweets to benchmark noisy text comprehension, evidencing that sentence vectors derived from early and middle BERT layers outperform existing vectors for noisy text classification regardless of sentence length.

In the domain of emotion recognition, models integrating multimodal signals with PLMs have advanced multi-label emotion classification by fine-tuning contextual embeddings extracted from pretrained models to handle overlapping emotions. This approach has yielded notable improvements in detecting nuanced emotional states such as joy, sadness, and anger [22]. Jabreel et al. [22] reported a 5-7% macro F1-score improvement over baselines, with effectiveness in distinguishing similar emotions despite persistent challenges related to data imbalance and limited interpretability. Future research directions include exploring transfer learning, explainability, and multimodal integration to further enhance performance and robustness.

To mitigate the inherent black-box nature of deep learning models, hybrid symbolic-connectionist methods are emerging. These integrate symbolic reasoning with PLMs to augment robustness, controllability, and interpretability in natural language generation and reasoning tasks, addressing issues related to model opacity [**?**]. Li et al. [**?**] emphasize the importance of combining symbolic knowledge structures with PLMs to improve controllable generation, fairness, and ethical considerations while balancing model size and computational demands.

Reducing dependence on large annotated datasets remains a critical area of innovation. Automated prompt construction techniques, including neural prompt synthesis and zero-shot prompting, have shown great promise. Zhao et al. [**?**] introduced NPPrompt, a method that automatically mines and synthesizes external task-related knowledge into coherent prompts without manual design. NPPrompt applies a combination of BM25, dense vector search, and BERT-based encoding to retrieve and rank relevant knowledge snippets, effectively framing tasks for PLMs and enabling substantial performance gains over baseline zero-shot methods—achieving results comparable to few-shot learning. This fully automated pipeline alleviates the brittleness of manual prompt engineering, enhancing generalization across diverse tasks and domains with minimal labeled data. Despite dependencies on external data quality and increased computational cost, such approaches represent significant progress toward scalable, adaptable, and efficient PLM deployment.

In summary, the evolving landscape of multimodal and multilingual AI is shaped by interconnected challenges including data scarcity, computational demands, ethical concerns, architectural innovation, and domain-specific complexities. Overcoming these obstacles will require continuous integration of technical advancements, cognitive insights, and ethical considerations through multidisciplinary collaboration and innovative methodologies. Such an integrative approach is essential to fully unlock the transformative potential of large-scale AI systems across a wide range of applications.

6 Conclusions

This survey has systematically traced the state-of-the-art developments and ongoing challenges in three intertwined areas of AI research: transformer architectures, multimodal large language models (MLLMs), and foundational pretrained language models (PLMs). Distinctively, we have connected architectural innovations, application-driven advances, and interdisciplinary insights to provide a comprehensive perspective that integrates technical depth with societal implications.

Transformer architectures have fundamentally transformed large-scale learning paradigms, epitomized by sparse Mixture-of-Experts designs such as the Switch Transformer. By routing each token to a single expert, this approach sustains extreme parameter scaling while markedly reducing computational costs, enabling practical deployment of trillion-parameter models [26]. The rigorous proof of transformers' Turing completeness under idealized conditions further affirms their universal computational potential, highlighting their versatility across AI tasks [**?**]. Nevertheless, the quadratic complexity bottleneck of self-attention motivates a rich vein of research into "X-formers" that leverage sparse attention patterns, kernel methods, and memory mechanisms to facilitate scalability and efficiency without sacrificing expressivity [8]. In computer vision, transformers have progressively displaced convolutional neural networks by capturing global context and long-range dependencies, though challenges of data efficiency and high training costs remain paramount [18].

MLLMs represent a crucial paradigm shift from unimodal to truly integrative AI, fusing vision, language, audio, and other modalities within unified frameworks. We have synthesized foundational principles such as modality-specific encoders, cross-modal interaction modules, and joint pretraining strategies that underpin state-of-the-art models [16]. Empirical successes on tasks including image captioning, visual question answering, and audio-visual speech recognition illustrate these models' potential to transcend text-only limitations [16]. Yet substantial obstacles persist, notably limited multimodal data availability, complex alignment requirements, and high computational demands, spurring ongoing innovations in self-supervised learning, parameter-efficient fine-tuning, and the establishment of standardized cross-modal evaluation benchmarks [16]. Specialized applications in healthcare and biomedicine harness heterogeneous data streams comprising clinical imaging, multi-omics, and wearable sensors to enable personalized medicine and real-time monitoring, underscoring transformative potential alongside heightened concerns about privacy, interpretability, and regulatory compliance [8**?**].

A unique contribution of this survey is its bridging of AI with cognitive neuroscience, elucidating how MLLMs' multimodal embeddings capture semantic categories and perceptual features in ways resonant with neural activity patterns in category-selective brain areas [19]. This semantic alignment advances AI interpretability and human-likeness but remains incomplete, as current models fall short in replicating nuanced, context-dependent human conceptual understanding [1]. We highlighted cutting-edge explainability approaches—such as multimodal explainable AI (MXAI) and graph neural network-based causal explanation frameworks—that enhance transparency and causal interpretability across modalities [21**?**]. This convergence of computational methods and neuroscientific insights charts a promising roadmap toward more interpretable, cognitively aligned AI systems.

We have also underscored ethical and robustness considerations as central themes linking architecture with societal impact. Integrating dynamic trust-aware frameworks that adapt data disclosure to user profiles and information sensitivity is especially critical for regulated domains like healthcare and finance [2]. Complementary explainability and governance approaches foster accountability, balancing model complexity with interpretability to build user trust [16, 21]. Persistent challenges—including hallucinations, bias, privacy concerns, and ethical alignment gaps—require sustained interdisciplinary research, informed policy development, and transparent model design [17**?**].

The practical ramifications of these advances are evident across multiple sectors: healthcare benefits from improved diagnostics and personalized treatment facilitated by multimodal integration [8]; education gains arise from enhanced analysis of non-verbal cues enabling collaborative learning [18]; transportation safety is augmented by sensor fusion supporting timely interventions [7]; multilingual NLP improvements stem from geographically adapted PLMs mitigating biases and improving regional performance [10**?**]; and resource-efficient adaptation leverages few-shot and zero-shot learning, reinforced by retrieval-augmented prompting techniques to transfer knowledge across domains [3**?**].

In Table 9, we summarize key insights, challenges, and future research directions spanning these domains.

Fully realizing the transformative potential of these technologies demands collaborative efforts across AI research, cognitive science, ethics, and domain expertise. Robust evaluation frameworks must systematically address not only task performance but also interpretability, fairness, and trustworthiness [22]. Transparent disclosure of data provenance and clear communication of model limitations, coupled with embedding human preferences into model training, will enhance alignment and help mitigate undesirable outputs such as hallucinations and toxicity [4**?**]. Addressing open challenges in scalability, multimodal data fusion, contextual reasoning, and domain adaptation will propel AI toward systems that are not only powerful and efficient but also ethically aligned and human-centric [4**?**].

In conclusion, this survey's distinctive integration of transformer technical advancements, multimodal language model breakthroughs, and cognitive neuroscience perspectives frames a comprehensive roadmap for AI research poised to reinvent multiple societal domains. Coupled with principled ethical frameworks and sustained innovation, this confluence promises robust, transparent, and ultimately beneficial AI systems that resonate with human values and cognitive architectures.

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

Table 9: Summary of Key Insights, Challenges, and Future Directions from the Survey

Domain	Key Insights	Challenges	Future Directions
Transformer Architectures	Sparse MoE enables trillion-scale models with efficiency [26]; Turing completeness proofs highlight universal computation [?]	Quadratic attention complexity; data efficiency; training costs [8, 18]	Hybrid sparse-kernel methods; hardware/software co-design; lightweight vision transformers [8, 18]
Multimodal LLMs	Principled multimodal fusion; robust zero/few-shot generalization [16]; impactful biomedical applications [?]	Data scarcity; alignment; computational expense; privacy and regulatory challenges [16?]	Self-supervised multimodal pretraining; parameter-efficient fine-tuning; robust evaluation benchmarks [16]
Cognitive-AI Integration	Semantic and perceptual embeddings align with neural representations [19]; MXAI advances interpretability [21?]	Partial modeling of human context-dependent reasoning and semantic nuances [1]	Richer behavioral datasets; graph causal explanations; multimodal cognitive alignment [1?]
Ethics and Robustness	Trust-aware AI frameworks; explainability governance [2, 16, 21]	Bias, hallucinations, privacy, alignment gaps [17?]	Interdisciplinary research; transparent model design; ethical regulations [17?]
Application Domains	Improved diagnostics, personalized medicine, safety, regional NLP, resource-efficient transfer [7, 8, 10, 18?]	Domain-specific data limitations; context adaptation; interpretability [8?]	Context-aware models; domain transfer; zero/few-shot with retrieval-enhanced prompts [3?]

[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