

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, 14 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, limitations inherent in uni-modal training have driven the emergence of multimodal large language models (MLLMs) that synthesize heterogeneous inputs—including text, images, audio, video, and structured biomedical data—to enable richer, contextually grounded understanding and generation [4?].

This progression draws motivation from insights in cognitive neuroscience and human cognition, emphasizing the integrated, multisensory nature of conceptual representations in the brain [6?]. Behavioral and neural evidence indicates that human conceptual knowledge embeds both semantic and perceptual features systematically across modality-specific and associative cortical areas [12?]. For instance, recent representational similarity analysis (RSA) studies reveal that MLLMs naturally develop object and concept embeddings that closely mirror 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 advantages of grounding computational architectures in neural and behavioral data to improve model interpretability and functional convergence [?].

Beyond theoretical foundations, integrating multimodal data within PLMs has catalyzed major improvements across practical

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

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

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

domains. In healthcare, MLLMs effectively leverage diverse medical 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 this requires sophisticated fusion mechanisms such as modality-specific encoders and cross-modal attention, combined with training strategies addressing data heterogeneity and scalability [19]. Similarly, in social and collaborative learning, incorporating nonverbal signals like posture and environmental cues through generative MLLMs enhances interaction analysis granularity, facilitating more effective pedagogical interventions [?]. Collectively, these applications demonstrate the wide-ranging impact of multimodal PLMs in tackling complex real-world tasks that demand the integration and interpretation of diverse data streams [16].

From a technical viewpoint, the architecture of MLLMs has progressed from early fusion approaches to dynamic cross-modal attention and instruction tuning, reflecting increasingly advanced methods for capturing multimodal correlations and task adaptability [26]. To address efficiency challenges posed by the quadratic complexity of self-attention mechanisms, innovations such as sparse Mixture-of-Experts (MoE) models—exemplified by the Switch Transformer—enable scaling to trillion-parameter regimes via selective expert activation, balancing computational cost and performance [17]. Foundational work also confirms that Transformers possess the expressivity and Turing completeness necessary to model long-range dependencies critical for multimodal reasoning [?]. Despite these advances, challenges remain in aligning multimodal data, ensuring representation robustness, and improving interpretability, which motivate multi-level evaluation frameworks combining quantitative metrics with cognitive benchmarks [5].

Ethical and trust considerations are equally crucial, especially in sensitive areas such as healthcare and finance. Emerging frameworks that integrate dynamic trust profiling, adaptive information sensitivity detection, and privacy-preserving output controls offer essential safeguards for responsible AI deployment [18]. Advances in explainable AI (XAI) tailored for multimodal data fusion—particularly via graph neural networks that preserve causability—underscore the importance of transparent, human-centered explanations aligned with domain expertise [?]. These aspects emphasize the need to embed ethical, privacy, and interpretability concerns early in model design and pretraining, alongside efforts to enhance model alignment with human values and to mitigate undesired behaviors [7].

The extensive literature on multimodal and pretrained language models spans foundational cognitive parallels, technical innovations in architecture and efficiency, diverse application domains, and ethical imperatives. This survey organizes these multifaceted advances within a coherent framework, clarifying both the potential and ongoing challenges of 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 heterogeneous data incompatibility. Late fusion methods aggregate output decisions or high-level features from individual modality-specific models, allowing for flexible and independent processing pipelines. Hybrid fusion combines elements of both, integrating features at multiple stages to harness their respective advantages.

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.

In summary, multimodal fusion enhances representation learning by leveraging complementary information across modalities. The choice of fusion strategy depends on data characteristics, computational constraints, and target applications. Understanding these methods and their appropriate contexts is essential for developing effective multimodal systems.

2.1 Multimodal Embeddings and Fusion Techniques

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

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

Table 1: Summary of Multimodal Fusion Methods and Their Applications

Fusion Method	Description	Applications
Early Fusion	Combines raw or low-level features before learning; captures joint feature space	Multimedia retrieval, medical image analysis
Late Fusion	Integrates decisions or high-level features from separate modality models	Emotion recognition, autonomous driving
Hybrid Fusion	Combines early and late fusion advantages via multi-stage integration	Multimodal sentiment analysis, healthcare monitoring

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

Beyond canonical modalities, the incorporation of non-verbal multimodal signals—such as postural behavior—in contexts like collaborative work and education has become increasingly prominent. Generative AI-driven frameworks using multimodal large language models (LLMs) have proven capable of extracting meaningful features from complex, noisy non-verbal data collected in naturalistic environments [6, 12?]. This integration not only broadens the applicability of multimodal fusion techniques but also underscores the necessity for adaptable architectures that process verbal and non-verbal information streams simultaneously.

Despite these promising developments, several challenges remain. Modality alignment and fusion complexity require sophisticated architectural designs and objective functions that balance the representation of modality-specific nuances while fostering shared semantic abstractions. Scalability issues also arise as both model size and multimodal data volume increase training demands. Moreover, the issue of multimodal hallucination—where models generate semantically plausible but factually incorrect associations across modalities—raises concerns in sensitive applications, emphasizing the need for robust calibration methods and human-in-the-loop validation approaches [24?].

2.2 Handling Noisy and Diverse Textual Inputs

The heterogeneity of textual inputs, particularly those originating from social media platforms like Twitter or from low-resource languages, imposes substantial challenges on representation learning frameworks. Textual noise manifests as deviations from standard language conventions, including irregular syntax, spelling variations, and code-switching, all of which undermine the assumptions underpinning pretrained language models (PLMs). To address these complications, specialized embedding methodologies have been developed that extract multi-layer latent features from models such as BERT, combining linguistic signals from diverse abstraction levels to create enriched sentence representations. This approach has been shown to improve downstream classification performance

on noisy textual datasets [9]. Importantly, initial and intermediate transformer layers tend to encode richer and more relevant linguistic information for noisy text than the final layers, indicating a layer-wise representational specialization that can be exploited for more robust noisy text understanding.

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. This approach notably reduces dependence on extensive parallel corpora or resource-intensive multilingual pretraining, making it particularly suitable for low-resource and multilingual contexts [10?]. Such methods facilitate improved semantic textual similarity by combining strengths from distinct monolingual models, thereby enhancing both efficiency and accuracy.

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. These shortcomings motivate the exploration of hybrid architectures that integrate symbolic reasoning and explicit knowledge with learned representations. Such neuro-symbolic approaches aim to bolster robustness and generalization capabilities in intricate linguistic scenarios, addressing critical cognitive faculties that current neural architectures lack.

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. 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.

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 demonstrate 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, significantly surpasses expert ophthalmologists by achieving an 83.9% top-five accuracy in predicting gene classes underlying rare inherited retinal diseases across diverse international cohorts [23]. This success is largely attributable to modality-specific Convolutional Neural Network (CoAtNet0) ensembles, which address data imbalance through weighted loss functions and ensemble averaging, while Uniform Manifold Approximation and Projection (UMAP) visualizations reveal meaningful genotype-phenotype correlations. Eye2Gene's interpretability, enhanced by attention maps highlighting image regions influencing predictions, supports clinical trust; however, limitations include coverage gaps towards rare genes and reliance predominantly on image-only data comparisons [23].

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

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 that must be safeguarded without compromising model utility. Strategies such as differential privacy, federated learning, and transparency frameworks are actively explored to mitigate bias and maintain confidentiality. However, practical deployment remains complex due to computational overheads and real-time operational demands [2?]. 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 directions to balance privacy with information utility in healthcare LLM applications [2].

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

2.3.2 Real-Time Safety and Autonomous Systems. Multimodal AI 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].

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

Addressing the increasing complexity of traffic environments demands the incorporation of sophisticated predictive analytics and 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), notably in low-resource and linguistically diverse contexts. The incorporation of PLMs such as Chinese BERT into non-autoregressive (NAR) automatic speech recognition (ASR) models alleviates the classical trade-off between decoding speed and transcription accuracy. By enriching acoustic representations with linguistic context provided by PLMs, these systems attain character error rates competitive with traditional autoregressive baselines (e.g., 6.9% vs. 6.5%) while achieving lower real-time factors conducive to rapid inference [24]. Such approaches address inherent challenges in tonal and homophonic features characteristic of Chinese speech without sacrificing computational efficiency.

Cross-lingual adaptation studies reveal that pretrained English language models, when fine-tuned systematically, outperform native-language models trained from scratch on low-resource languages, demonstrating the potency of transfer learning in leveraging resource-rich linguistic representations to enhance non-English NLP tasks [10]. Moreover, geographic adaptation of PLMs via fine-tuning on curated regional corpora effectively mitigates biases introduced by training on predominantly North American and European English data. This geographically informed adaptation yields substantial improvements, with F1-score increases exceeding 4 points and measurable reductions in perplexity and error rates related to region-specific lexical and syntactic features [?].

Collectively, these advances highlight the critical importance of accounting for linguistic diversity and geographic variation to develop robust and equitable NLP systems. Future directions include combining geographic factors with social and cultural considerations to further enhance fairness, robustness, and representation in pretrained language models.

2.3.4 Text Generation and Emotion Recognition. Text generation frameworks powered by PLMs encompass varied paradigms including open-ended, conditional, and controllable generation tasks. Each paradigm presents distinct challenges regarding output coherence, diversity, and ethical constraints [?]. The integration of reinforcement learning (RL) techniques—particularly Reinforcement Learning from Human Feedback (RLHF) and Proximal Policy Optimization (PPO)—has facilitated the alignment of model outputs with human preferences, enhancing multi-step reasoning capabilities and mitigating undesirable biases [?]. Despite these advances,

challenges such as sample inefficiency, reward design complexity, and safety concerns remain, motivating the development of hybrid strategies that combine symbolic reasoning and hardware acceleration to improve scalability and robustness.

In the domain of emotion recognition, the combination of PLMs with deep neural architectures has extended capabilities beyond single-label classification to nuanced multi-label emotion detection. By leveraging contextualized embeddings and applying a sigmoid-activated output layer, these models significantly outperform traditional baselines, achieving macro F1-score improvements of 5–7% [22]. This approach effectively manages overlapping emotional states such as joy, sadness, and anger while offering enhanced interpretability. Nevertheless, persistent challenges include addressing data imbalance and distinguishing semantically similar emotions, which encourage future exploration into integrating multimodal inputs and explainability techniques to further enhance performance and transparency.

Together, these advances delineate a comprehensive landscape where multimodal AI and large language models markedly enhance predictive and generative tasks. They also foreground critical ethical, privacy, and fairness considerations indispensable for responsible deployment. Their broad applicability across healthcare, autonomous systems, and natural language domains underscores their transformative potential in modern AI research and applications.

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

Explainable AI (XAI) focuses on making AI systems transparent and understandable to users, which is critical for fostering trust and enabling informed decision-making. Trustworthiness in AI encompasses reliability, robustness, fairness, and accountability, ensuring systems behave as expected in diverse real-world settings. Ethical considerations involve addressing issues such as bias mitigation, privacy preservation, and the societal impact of AI deployment. Integrating XAI methods plays a crucial role in enhancing trustworthiness by providing insights into model decisions and facilitating the detection of potential ethical concerns. Together, these aspects form the foundation for responsible AI development and deployment, guiding both researchers and practitioners toward building AI systems that are not only powerful but also aligned with human values and societal norms.

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

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

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 shown in the following summary [21]:

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

Moreover, noise management and mitigation of hallucinations in large multimodal models require 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 improve 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, enhancing explanation fidelity while maintaining computational efficiency.

However, disentangling contributions of intertwined features while preserving interpretability remains a critical and unresolved challenge, emphasizing the need for standardized benchmarks that 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.

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, its deployment must remain transparent and auditable 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 models in LMaaS severely limit user visibility into internal operations [11], complicating benchmarking and evaluation efforts. Addressing these challenges demands 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 originating from their training data and architectural paradigms. Documented risks include inadvertent

Table 4: Performance comparison demonstrating the effectiveness of explainable AI 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

memorization of sensitive or private content, systemic biases reflecting the underlying data distributions, the propagation of toxic or false information, and model failure modes when exposed to adversarial inputs [3]. These vulnerabilities significantly complicate efforts to deploy LLMs responsibly in sensitive or high-stakes contexts.

To address these complexities, a mechanistic interpretability framework that goes beyond superficial behavioral benchmarks is essential. This framework involves 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 instance, trust frameworks incorporating dynamic user profiling and adaptive output control can secure sensitive data disclosure based on context and user trust levels [2]. In addition, interpretability efforts for multimodal biomedical AI highlight the importance of integrating diverse data types while maintaining ethical considerations and privacy [?]. Similarly, multimodal explainable AI research strives to align explanations with human cognition and mitigate biases through causal and counterfactual reasoning [25]. Achieving this level of rigor confronts practical hurdles, including the high computational and data costs of mechanistic analyses and the difficulty of balancing multitask learning objectives without compromising robustness [20].

Emerging research emphasizes improving computational efficiency, developing multimodal grounding for richer language understanding, and enhancing robustness against adversarial and distributional shifts. Equally important is embedding these technical advances within ethical frameworks that prioritize inclusivity, transparency, and user trust. Cross-disciplinary insights from linguistics, cognitive science, and social sciences are critical for addressing bias and toxicity at their origin, ensuring that challenges are approached both technically and socially [3, 28]. Only through such integrated approaches—melding technical innovation with principled ethical guidelines—can AI systems become truly trustworthy, reliable, and resilient.

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

Formally, this approach optimizes the likelihood under a Bradley-Terry model applied to human preference pairs, promoting the

internalization of aligned behavior directly within model parameters instead of as an external corrective process. Specifically, the preference loss is defined 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 preference scores predicted by the model for inputs x_i and x_j , respectively, y_{ij} indicates the preferred instance, and σ is the sigmoid function. Empirical evidence shows that such jointly pretrained models outperform those fine-tuned on preference data post-pretraining, achieving lower rates of undesirable outputs without sacrificing language modeling quality. This paradigm shift suggests that early preference integration reinforces intrinsic model trustworthiness and reduces dependency on costly, potentially unstable fine-tuning methods.

Table 5 summarizes key comparative results from recent experiments [4], illustrating improvements in perplexity, toxicity reduction, and preference accuracy for jointly pretrained models relative to alternatives.

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

3 Behavioral, Cognitive, and Neuroscientific Insights

This section aims to synthesize 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 how these models capture human conceptual representations, linguistic capabilities, and output consistency, while discussing challenges and implications for future AI development and ethical deployment.

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 have demonstrated 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 example, Du et al. [6] analyzed 4.7 million

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

Method	Perplexity	Toxicity Rate	Preference Accuracy
LM only	12.3	0.15	0.50
Fine-tuned preferences	12.5	0.10	0.68
Joint pretraining	12.7	0.07	0.75

human similarity judgments on nearly 2,000 natural objects, deriving sparse, non-negative embeddings via the 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 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 [?].

However, the extent of alignment varies by representation level and model architecture. Mahner et al. [?] found that early layers in deep neural networks predominantly encode low-level visual features (e.g., shape and texture), while higher layers represent abstract semantic categories. This hierarchical pattern reflects processing stages in human visual cognition. Nonetheless, some representational dimensions exclusive to human cognition remain underrepresented in current models. These gaps highlight challenges arising from dataset biases and the complexity of integrating visual and semantic cues, emphasizing the need for models embedding richer multimodal context alongside augmented behavioral data [? ? ?].

Enriched multimodal context involves both architectural considerations and the nature of training datasets and behavioral signals guiding representation learning. Empirical evidence suggests that training on multimodal datasets that reflect human perceptual and conceptual experiences improves AI models' capacity to capture human-like representations [? ?]. For instance, multimodal large language models in medical domains integrate diverse data types such as images, audio, and text to improve clinical decision-making, although challenges related to data fusion, interpretability, and ethical concerns remain [? ?]. Importantly, this 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 [? ? ?]. For example, ethical deployment requires addressing biases and ensuring privacy, safety, and informed consent in AI systems that reflect human cognition and social norms.

Complementing neuroscientific insights, cognitive and linguistic evaluations of pretrained language models (PLMs) reveal their respective strengths and limitations in syntactic, semantic, and reasoning tasks prior to fine-tuning. Chang and Bergen [3] provide a comprehensive survey showing that PLMs robustly handle fundamental syntactic rules (e.g., subject-verb agreement), semantic compositionality, analogical reasoning, and basic logical inference. Conversely, these models struggle with complex syntactic phenomena, negation, implicit pragmatics, and multi-step inferencing, indicating incomplete linguistic and conceptual comprehension. These linguistic limitations resonate with neuroscientific findings

regarding incomplete semantic alignment, suggesting current architectures and pretraining approaches inadequately capture the full complexity of human linguistic cognition.

Systematic assessments of model consistency further underscore challenges in reliability and interpretability. Wang et al. [28] developed benchmark datasets to evaluate pretrained models such as GPT-2, BERT, RoBERTa, and GPT-3 across factual, paraphrase, and negation consistency. Results show factual consistency below 80% in many cases, challenging trustworthiness in critical applications. Detailed analyses attribute inconsistencies mainly to miscalibrated confidence rather than fundamental knowledge deficits. Temperature scaling—a post hoc probabilistic calibration technique—improves consistency metrics by up to 20% without changing model weights [28]. This highlights the importance of integrating uncertainty quantification and calibration mechanisms to enhance robustness and practical deployment.

The following table summarizes key techniques, challenges, and future directions informed by behavioral, cognitive, and neuroscientific insights in AI model development:

In summary, these convergent lines of evidence illustrate 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 necessitates interdisciplinary frameworks that bridge behavioral science, neuroscience, computational modeling, and linguistic theory. Such integrative approaches are critical for advancing AI systems whose internal representations and outputs more faithfully reflect the complexity and richness of human cognition and communication, and for ensuring their responsible and ethical application [3, 28? ? ?].

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

4.1 Retrieval-Pretrained Transformer Architectures

Recent developments in transformer architectures address the inherent limitations of fixed-length context windows in conventional models by incorporating retrieval mechanisms that dynamically query relevant contextual information. The Retrieval-Pretrained Transformer (RPT) exemplifies such innovation. At each decoding step, RPT 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

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 [67]	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 [??]

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 advances technical performance and 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 an important transition toward more context-aware, precise, and equitable language modeling paradigms. These innovations move beyond traditional static and monolithic designs, enabling dynamic, multifaceted systems that achieve nuanced understanding across a variety of domains and demographic contexts. This progression marks a pivotal evolution in both the design and practical utility of large-scale language models, highlighting their growing capacity to address diverse real-world challenges with improved adaptability and fairness.

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

5 Challenges and Future Directions

The field of multimodal AI presents several significant challenges that must be addressed to unlock its full potential. Here, we clarify key difficulties with concrete examples and provide a structured overview to aid understanding.

5.1 Integration of Heterogeneous Modalities

Effectively combining information from diverse data types—such as text, images, audio, and video—remains a core challenge. For instance, aligning semantic content from text with visual features in images requires models to capture complex relationships across modalities that have fundamentally different structures and representations. This complexity often leads to difficulties in learning joint representations that are both informative and generalizable.

5.2 Scalability and Efficiency

Multimodal models typically involve large architectures with high computational costs, making them resource-intensive. Real-world applications, such as real-time language translation with video feeds, demand lightweight yet accurate models. Developing methods that balance performance with efficiency is an open research question.

5.3 Data Scarcity and Quality

While single-modal datasets are abundant, high-quality, large-scale, and well-annotated multimodal datasets are limited. For example, datasets that include paired speech and gesture data with semantic annotations are rare, hindering progress in this area. Moreover, ensuring the alignment and consistency of multimodal data is often non-trivial.

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.

5.5 Robustness and Generalization

Models must perform reliably under noisy or missing modalities. For example, a video classification system should still function adequately if the audio track is corrupted or absent. Addressing modality dropout and domain shifts to ensure robustness across varied scenarios is critical.

5.6 Future Research Directions

To advance the field, researchers should focus on: - Designing unified frameworks that flexibly handle multiple modalities, adapting to the presence or absence of specific inputs. - Creating benchmarks with diverse and challenging multimodal tasks, along with standardized evaluation protocols. - Investigating transfer learning approaches that leverage large unimodal models for multimodal tasks. - Improving interpretability tools tailored for multimodal architectures. - Exploring energy-efficient model designs for deployment in edge devices.

Summary of Key Challenges

This table encapsulates the principal challenges and aligns them with concrete research objectives to guide future endeavours. Addressing these issues with targeted strategies will push multimodal AI toward more practical and transformative applications.

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? ? ? ?]. Moreover, insufficient dataset diversity exacerbates biases and underrepresentation, especially concerning geographic and societal factors, thus highlighting the critical need for expanding datasets with balanced representation and the development of equitable benchmarking protocols to ensure fairness and inclusivity [7? ? ? ?].

In parallel, the computational scalability associated with transformer-based architectures remains a vital bottleneck. The quadratic time and memory complexities of standard self-attention mechanisms hinder efficient training and inference on long sequences and multimodal inputs [5, 8? ? ? ?]. Recent innovations, including sparse transformers and kernel-based attention approximations, have emerged to mitigate these challenges. For instance, Sparse Mixture-of-Experts models such as the Switch Transformer employ expert routing, activating only subsets of parameters per input token, significantly reducing computational overhead while facilitating scaling to trillion-parameter models [5, 26]. These advances combine architectural sparsity with load balancing losses to maintain model expressiveness and prevent expert underutilization [26]. Despite promising results, such models require careful tuning and pose new challenges in training stability and efficiency. Therefore, addressing data scarcity and computational constraints demands integrated

Table 8: Summary of Major Challenges in Multimodal AI

Challenge	Description	Example	Research Objective
Integration	Combining heterogeneous data types with different structures	Aligning text semantics with visual features	Develop joint embedding methods capturing cross-modal relations
Scalability	High computational and memory requirements	Real-time multimodal translation systems	Design efficient architectures balancing accuracy and speed
Data Scarcity	Limited large-scale, well-annotated multimodal datasets	Lack of paired speech-gesture semantic datasets	Create new datasets and data augmentation techniques
Interpretability	Difficulty in explaining multimodal decision processes	Understanding model output in healthcare diagnostics	Develop explainability techniques for cross-modal reasoning
Robustness	Handling noisy or missing modalities during inference	Video classification missing audio data	Build models resilient to missing or corrupted inputs

strategies involving the curation of diverse, high-quality datasets, adoption of efficient transformer variants, and cross-modal optimization techniques to enhance both model performance and scalability.

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, while sensitivity detection leverages domain-specific semantic analyses to identify sensitive information for adaptive output control [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. 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 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 [?]. 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].

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]. Research into temporal dynamics and the development of lightweight, transformer-based frameworks tailored for real-time applications, such as facial action unit detection, address critical operational constraints and requirements of dynamic environments, balancing accuracy and computational efficiency [14].

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, and aligned with human cognition.

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 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 [?], or guidelines managing language model behavior to mitigate biases, misinformation, and privacy risks [3]. Collectively, these collaborative efforts promote responsible AI adoption in critical sectors such as healthcare, finance, and education.

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 ref27.

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?]. 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 show 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 found to better encode linguistic characteristics of noisy texts, enhancing classification performance on tasks involving such inputs. Complementing this, emotion recognition models that integrate multimodal signals with PLMs have advanced multi-label emotion classification, achieving notable improvements in detecting nuanced emotional states such as joy, sadness, and anger [22]. These models typically fine-tune contextual embeddings extracted from pretrained models to handle overlapping emotions; however, challenges remain regarding data imbalance, differentiation of similar emotions, and limited interpretability. To address inherent black-box issues in deep learning, hybrid symbolic-connectionist approaches are emerging. These approaches aim to enhance robustness, controllability, and interpretability in natural language generation and reasoning tasks by integrating symbolic reasoning with PLMs, thus mitigating issues related to deep models' opacity [?].

Reducing dependence on large annotated datasets is another critical area of innovation. Automated prompt construction techniques, including neural prompt synthesis and zero-shot prompting, have shown promise in this regard. A notable example is NPPrompt, which automatically mines and synthesizes external task-related knowledge into coherent prompts without manual design, achieving substantial performance gains over baseline zero-shot methods and nearing few-shot learning levels [?]. By retrieving and ranking relevant knowledge snippets through a combination of BM25, dense vector search, and BERT-based encoding, NPPrompt frames tasks comprehensibly for PLMs, thereby enhancing generalization across diverse tasks and domains with minimal labeled data. Although dependent on the quality of external data and introducing computational costs, such methods represent a significant step toward more scalable, adaptable, and efficient PLMs.

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 activating only one expert per token, this approach sustains extreme parameter scaling while markedly reducing computational costs, enabling practical deployment of trillion-parameter models [26]. The theoretical demonstration of transformers' Turing completeness further affirms their universal computational potential, highlighting their versatility across AI tasks [?]. Nevertheless, the quadratic complexity bottleneck of self-attention spurs 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 the vision domain, transformers have successfully displaced convolutional neural networks by capturing global context and long-range dependencies, though challenges of data efficiency and 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 in diverse 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, particularly regarding limited multimodal data availability, alignment complexity, and high computational demand, motivating ongoing innovations in self-supervised learning, parameter-efficient fine-tuning, and cross-modal evaluation benchmarks [16]. In specialized fields like healthcare and biomedicine, MLLMs harness heterogeneous data streams encompassing clinical imaging, omics, and wearable sensors to enable personalized medicine and real-time monitoring, underscoring transformative applications 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 that resonate with neural activity patterns in category-selective brain areas [19]. This semantic alignment advances AI's 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 and 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 data 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 issues, and ethical alignment—require ongoing 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 treatments through multimodal integration [8]; education gains from enhanced analysis of non-verbal cues for collaborative learning [18]; transportation safety is augmented by sensor fusion enabling timely interventions [7]; multilingual NLP sees gains from geographically adapted PLMs mitigating biases and improving regional performance [10?]; and resource-efficient adaptation leverages few-shot and zero-shot learning bolstered by retrieval and prompt synthesis 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 model limitations, combined with embedding human preferences into model training, will enhance alignment and 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 to deliver 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/>

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 [?]	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; strong zero/few-shot generalization [16]; impactful biomedical applications [?]	Data scarcity; alignment; computational expense; privacy and regulatory issues [16?]	Self-supervised multimodal pretraining; parameter-efficient finetuning; 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?]

volumes/5/issues/4

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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