Journal of Artificial Intelligence and AI Ethics

SCIVOLVE

Research Article

Open Access Full Text Article

Multimodal AI: PaLM-E's Role within Vision–Language–Robotics and Future of Fine-Tuning

Zarif Bin Akhtar[1]*

[1]Department of Computing, Institute of Electrical and Electronics Engineers (IEEE), USA.

***Correspondence:**

Zarif Bin Akhtar
Department of Computing, Institute of Electrical and Electronics Engineers (IEEE), USA.

Received: February 20, 2026; **Published:** March 27, 2026.

How to cite this article:

Akhtar, Z.B. (2026), 'Multimodal AI: PaLM-E's role within Vision–Language–Robotics and Future of Fine-Tuning', *Journal of Artificial Intelligence and AI Ethics*, 1(1), pp. 1–10.

Abstract

This study explores the convergence of artificial intelligence (AI), robotics, and advanced language models, centering on the PaLM-E framework. By examining its adaptability and reasoning in varied robotic contexts, the work demonstrates how PaLM-E can interpret natural language instructions and translate them into precise, low-level robotic commands. The investigation also evaluates Parameter-Efficient Fine-Tuning (PEFT) strategies, including Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), tracing their development and highlighting their capacity to improve performance while reducing the number of trainable parameters. Beyond robotics, the research surveys notable generative AI systems such as GPT-3, GPT-4, Copilot, Bard, LLaMA, Stable Diffusion, Midjourney, and DALL-E assessing their versatility in producing text, code, images, and other outputs from natural language prompts. An overview of AI's historical progression is provided, from speculative concepts to modern, practical implementations, with emphasis on generative AI's rapid expansion in the 21st century. Real-world applications are examined across robotics, planning, business intelligence, and synthetic data generation, alongside an assessment of hardware and software deployment options, from local consumer systems to cloud-based infrastructures. The advantages of local deployment for privacy protection, intellectual property security, and freedom from external censorship are emphasized. Ethical considerations including issues of bias, misinformation, security, and societal implications are addressed, with proposed guidelines for responsible AI development and integration. Overall, the work highlights the deep interconnection between vision, language, and robotics, offering insights that may guide the next generation of generative AI research and applications.

Keywords: Artificial Intelligence (AI), Computer Vision, Deep Learning (DL), Generative Artificial Intelligence (GAI), Large Language Models (LLMs), Machine Learning (ML), Robotics

1. Introduction

Artificial intelligence (AI) continues to advance rapidly, with the integration of vision, language, and robotics emerging as a critical domain for creating systems capable of seamless, context-aware interaction with the physical world (Akhtar, 2024; Akhtar, 2025; Akhtar and Rawol, 2025). This study examines the PaLM-E architecture a multimodal model that extends the boundaries of perception and reasoning in robotic systems (Hamid, 2025; Jin, Yu and Grzybowski, 2025; Tharayil, Krishnapriya and Alomari, 2025). By assessing its performance in diverse robotic environments, evaluating its approach to vision-language reasoning, and analyzing its decision-making capabilities, we reveal how PaLM-E demonstrates remarkable adaptability, problem-solving capacity, and generalization across tasks.

In parallel, the research investigates the concept of Parameter-Efficient Fine-Tuning (PEFT), which enables large language models (LLMs) to adapt to specialized tasks while minimizing computational demands. Techniques such as Low-Rank Adaptation (LoRA) and its quantized variant (QLoRA) are discussed alongside methods including T-Few, AdaMix, and MEFT. These strategies are explored in terms of their potential to maintain or improve task performance while significantly reducing the number of trainable parameters, offering a pragmatic balance between efficiency and capability.

The discussion then shifts toward Generative Artificial Intelligence (Generative AI), focusing on its growing influence in domains spanning text, code, imagery, audio, and multimodal synthesis. Prominent transformer-based models such as GPT-3, Copilot, Bard, and LLaMA, as well as image-generation systems including Stable Diffusion, Midjourney, and DALL-E, are examined for their impact and adaptability (Fahad et al., 2025; Bewersdorff et al., 2025; Jacobs, 2025). These technologies are shaping industries from creative arts and journalism to healthcare, robotics, and financial analytics. Alongside their transformative potential, the work addresses pressing ethical considerations, including fairness, misinformation, and societal impact (Lee et al., 2025; Bravo, Rodriguez, Hidalgo and Angulo, 2025; Parvin, Joo, Jung and Mandal, 2025).

A historical perspective frames this analysis, tracing AI's evolution from early theoretical propositions in the mid-20th century to the diverse, high-performance generative models of today (Akhtar, no date; Akhtar and Rozario, 2025; Akhtar, 2024). The trajectory includes philosophical debates, pioneering computational experiments, and the modern surge in creative automation. Finally, the study surveys the software and hardware ecosystems supporting generative AI, from privacy-preserving local deployments to scalable cloud-based infrastructures.

By weaving together technical developments, historical context, and ethical discourse, this research presents a holistic view of generative AI's role in shaping future intelligent systems highlighting its potential, its challenges, and its significance for the next generation of AI-driven solutions.

2. Methods and Experimental Analysis

This study employs a multi-phase approach to evaluate the PaLM-E architecture within robotic environments, focusing on its adaptability, decision-making efficiency, and task performance. The first stage involves designing a diverse set of robotic tasks, ranging from simple manipulations to complex, long-horizon action sequences. To provide a consistent baseline for evaluation, a standardized testing framework is developed, simulating realistic environmental conditions to measure PaLM-E's responses and reasoning capabilities. In the integration stage, PaLM-E is combined with low-level language-to-action control policies, enabling natural language prompts to be converted into precise executable commands for robotic actuators. This stage is essential for assessing the model's readiness for real-world deployment and its resilience in dynamically changing settings. Robustness is further tested by introducing adversarial perturbations, while its ability to generalize is examined through tasks outside its training distribution. This enables a deeper understanding of its potential in transfer learning and adaptability to novel scenarios.

The research also investigates Parameter-Efficient Fine-Tuning (PEFT) methods, with emphasis on Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA). A literature-driven overview traces the evolution of PEFT techniques, outlining both their advantages and associated limitations. LoRA implementation is analyzed with attention to the starting point preservation hypothesis, which underpins its capacity to reduce trainable parameters without degrading task accuracy. Subsequently, QLoRA is examined for its use of quantization to further enhance efficiency while retaining model fidelity. Beyond PaLM-E, the study surveys the broader field of Generative AI, evaluating prominent transformer-based models such as GPT-3, GPT-4, Copilot, Bard, LLaMA, Stable Diffusion, Midjourney, and DALL-E. These models are assessed for their ability to interpret natural language inputs and produce varied outputs across modalities. Historical insights into AI's progression from speculative beginnings to real-world applications are interwoven to contextualize the emergence of generative AI in the 21st century.

Experimental analysis also extends to application domains, including robotics, automated planning, and business intelligence. Particular attention is given to the role of generative AI in synthetic data generation and its implications for analytics and decision-making processes. Modalities examined include text, code, image, and multimodal generation, covering both unimodal and hybrid architectures.

The investigation incorporates a review of the software and hardware ecosystems supporting generative AI, from integration into consumer-grade devices to deployment on high-performance cloud infrastructure. Scalability, accessibility, and usability are compared, with a focus on the benefits of local execution such as enhanced privacy, intellectual property protection, and resistance to censorship or rate limits.

Finally, the methodology embeds an ethical evaluation of generative AI deployment. Issues such as data privacy, model bias, misinformation, and misuse (including deepfakes and synthetic media) are addressed, alongside proposed guidelines for responsible innovation. This ensures that technical performance assessments are complemented by societal impact considerations.

The combined experimental and analytical processes yield insights into the synergy between vision, language, and robotics, illustrating their collective role in expanding the capabilities of AI systems. The findings also inform projections for future model development, offering recommendations to address emerging challenges in this rapidly advancing field.

3. Background Research and Available Knowledge

Generative Artificial Intelligence (generative AI or GenAI) describes AI systems capable of creating original content such as text, images, audio, video, or other forms of media by leveraging generative models trained on large datasets. These models capture patterns, structures, and semantic relationships within the training data, enabling them to produce new outputs with similar characteristics. The early 2020s marked a significant leap in this field, driven by advances in transformer-based deep neural networks. These breakthroughs facilitated the emergence of systems capable of responding to natural language prompts, including large language model (LLM) chatbots and text-to-image synthesis tools (Akhtar, 2024; Akhtar, 2025; Akhtar and Rawol, 2025).

Generative AI has found widespread application across numerous sectors, from creative domains like art, literature, and screenwriting to technical areas such as software engineering, healthcare diagnostics, financial forecasting, gaming, marketing, and fashion design. Substantial investments from major technology companies including Microsoft, Google, and Baidu alongside smaller innovators, reflect the technology's growing strategic importance. At the same time, its rapid adoption has raised concerns over misuse, including the facilitation of cybercrime, the spread of disinformation, and the creation of deepfake media (Akhtar and Rawol, 2025; Hamid, 2025; Jin, Yu and Grzybowski, 2025; Tharayil, Krishnapriya and Alomari, 2025; Fahad et al., 2025; Bewersdorff et al., 2025; Jacobs, 2025; Lee et al., 2025; Bravo et al., 2025; Parvin et al., 2025; Akhtar, 2025).

The conceptual roots of AI trace back to the mid-20th century, with the field formally established as an academic discipline in 1956. Early notions of automated creativity date as far back as ancient Greece, evolving through mechanical and programmable automatons to today's sophisticated generative systems (Fahad et al., 2025; Bewersdorff et al., 2025; Jacobs, 2025; Lee et al., 2025;

Bravo et al., 2025; Parvin et al., 2025; Akhtar, no date; Akhtar and Rozario, 2025; Akhtar, 2024; Ma et al., 2025; Hawthorne, 2025). Alan Turing's seminal 1950 work posed foundational questions regarding machine intelligence, laying theoretical groundwork that continues to influence AI research. Over the decades, AI has experienced alternating waves of enthusiasm and challenge, leading to milestones such as early generative planning systems and, more recently, advanced generative models capable of highly complex, multimodal tasks (Nayak, 2025; Bland, 2025; Wu and He, 2025; Wang et al., 2025; Ullah et al., 2025; Hou et al., 2025; Ma et al., 2025; Yang et al., 2025; Areerob et al., 2025; Carvalhido, Cardoso and Cerqueira, 2025; Schouten et al., 2025).

Modern generative AI spans a broad range of modalities, including natural language, programming code, visual art, music, video production, molecular design, robotics, autonomous planning, and business analytics. While high-capacity LLMs such as GPT-4 and PaLM are typically deployed on large-scale data center infrastructure, smaller models with fewer parameters can operate on personal computers, embedded systems, and smartphones. Generative AI capabilities are now embedded in diverse products, from conversational agents like ChatGPT to development tools such as GitHub Copilot, with many frameworks also released as open-source software (Hou et al., 2025; Ma et al., 2025; Yang et al., 2025; Areerob et al., 2025; Carvalhido, Cardoso and Cerqueira, 2025; Schouten et al., 2025; Huang et al., 2025; Lu et al., 2025; Yuan, Li and Zhao, 2025; Ryu, Choi and Yoo, 2025; Yang et al., 2025).

Running generative AI models locally provides key advantages, including improved privacy, protection of intellectual property, and freedom from external rate limits or content restrictions. However, resource-intensive models with hundreds of billions of parameters generally require cloud-based access due to their computational demands. Generative AI stands as both a transformative force across industries and a technology that raises critical challenges, underscoring the need for ongoing research into its ethical, social, and technical implications.

4. Experimental Designs & Simulations

The experimental component of this research is structured into two primary domains: the evaluation of PaLM-E for robotic applications and the investigation of Parameter-Efficient Fine-Tuning (PEFT) strategies for Large Language Models (LLMs).

4.1. PaLM-E Robotic Evaluation

PaLM-E, developed by Google researchers, addresses one of the key challenges in robotics limited access to large-scale, high-quality datasets by combining multimodal sensor inputs with advanced language modeling. Unlike earlier approaches, PaLM-E integrates data directly from robotic agents, including visual inputs, robot state information, and neural scene representations, alongside the capabilities of the PaLM language model. This fusion produces a powerful vision-language architecture capable of performing tasks across various robots and operational settings.

The model merges PaLM with ViT-22B, one of Google's high-performance vision transformers, enabling proficiency in both visual interpretation and natural language reasoning. Its largest variant, PaLM-E-562B, surpasses prior models on the OK-VQA benchmark without task-specific fine-tuning, while preserving strong language capabilities. In its architecture, multimodal observations text, images, and state data are injected into a pre-trained language model, which generates textual outputs in an auto-regressive manner for decision-making or question answering.

A notable strength of PaLM-E lies in its ability to transfer knowledge gained from large-scale vision-language training into robotic tasks. This shared representational framework allows it to handle multiple, distinct objectives without loss of performance. Experiments are conducted to measure its adaptability, robustness under environmental perturbations, and efficiency in executing both familiar and novel robotic tasks. Figures 1, 2 provide conceptual illustrations of this integration.

4.2. PEFT Model Fine-Tuning Experiments

The second branch of experimentation examines PEFT methodologies, with a focus on optimizing LLMs for specialized applications while minimizing computational and memory costs. Several approaches are explored, including T-Few, AdaMix, and MEFT, leading to a detailed assessment of Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA).

LoRA introduces trainable low-rank matrices into each layer of a Transformer during fine-tuning, significantly reducing the number of parameters that must be updated. This design minimizes storage demands and computational overhead while maintaining or enhancing model accuracy. The starting point preservation principle and the role of low-rank adapters in efficient task-switching are analyzed, highlighting LoRA's suitability for time-sensitive or resource-constrained deployments.

Building upon LoRA, QLoRA integrates quantization to further improve parameter efficiency. Experimental trials employ NF4 quantization and Double Quantization techniques, enabling large-scale models to be fine-tuned on limited hardware resources without sacrificing performance. Memory footprint reduction, scalability across different LLM architectures, and applicability to real-world tasks are assessed in detail.

4.3. Simulation and Evaluation Process

Step-by-step simulation workflows are designed to evaluate the adaptation of pre-trained models to targeted tasks. For PaLM-E, robotic simulation environments are configured to replicate realistic operational conditions, with varied task complexity and dynamic changes to test generalization capabilities. For PEFT techniques, controlled fine-tuning sessions are run across multiple datasets, enabling comparative analysis of performance metrics, parameter savings, and resource utilization. The combined experimentation provides a comprehensive understanding of PaLM-E's role in robotic intelligence and the efficiency gains achievable through PEFT methods such as LoRA and QLoRA. The findings offer practical guidance for researchers and practitioners aiming to deploy high-performance models under real-world computational constraints.

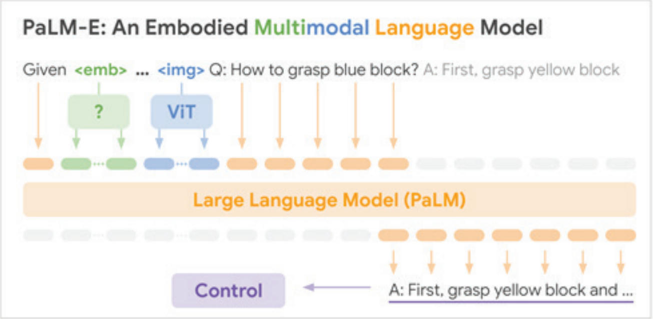

PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and ...

Figure 1. PaLM-E An Embodied Multimodal Language Model in action 1

5. GAI: A Deep Dive

Generative Artificial Intelligence (GAI) refers to a class of AI algorithms capable of producing a wide variety of content, including text, images, audio, code, simulations, and videos.

Recent advances have demonstrated the transformative potential of this technology in creative and professional domains. Among the most prominent examples is ChatGPT, developed by OpenAI, which functions as a versatile conversational agent capable of answering diverse questions and generating contextually relevant outputs. Upon its public release in November 2022, the tool achieved rapid adoption, amassing over one million users within just five days.

ChatGPT's versatility extends to generating computer code, essays, poetry, and more, illustrating its adaptability across both technical and creative tasks. Despite ongoing debates and concerns about AI's societal impact, tools like ChatGPT and image-generation systems such as DALL·E have the potential to reshape industries ranging from design and content creation to education and software development. However, the scale of its influence on employment, productivity, and workflows remains uncertain, underscoring the dual nature of the opportunities and challenges it presents.

To place GAI in context, it is important to distinguish it from the broader fields of Artificial Intelligence (AI) and Machine Learning (ML). AI broadly refers to systems designed to mimic human intelligence, while ML is a subset that enables models to learn patterns from data without explicit rule-based programming. GAI represents a specific advancement within ML, in which models do not simply classify existing data but can generate novel content on demand. Text-based GAI systems like ChatGPT are often trained through self-supervised learning, where vast volumes of text are used to predict subsequent words in a sequence. While ChatGPT has brought global attention to this capability, it follows a lineage of influential predecessors such as GPT-3 and BERT, both of which significantly advanced natural language processing.

The evolution from traditional supervised learning to self-supervised learning has played a crucial role in enhancing model quality and versatility. Building GAI models is resource-intensive, requiring substantial computational power, expertise, and data resources typically concentrated within large technology companies such as OpenAI, DeepMind, and Meta. For instance, GPT-3's training involved roughly 45 terabytes of text data, reflecting the scale of investment necessary to achieve such capabilities.

The outputs of GAI models can closely resemble human-created work, spanning essays, programming scripts, visual art, music, and multimedia content. Their quality depends heavily on the breadth and relevance of the training data, as well as alignment with the intended application. The combination of pattern recognition and controlled randomness in these models allows them to produce varied and often creative results. This flexibility has made GAI valuable in industries such as information technology, marketing, and healthcare, where it can automate tasks like drafting technical documents, refining source code, or producing tailored marketing content.

Nevertheless, these systems carry inherent risks. While outputs may appear convincing, they can also contain factual inaccuracies, exhibit bias, or generate inappropriate material. Such flaws may lead to reputational damage or legal complications. Mitigation strategies include curating high-quality training data, developing smaller domain-specific models to limit scope, and employing human oversight to review outputs prior to deployment.

Given the relatively recent emergence of large-scale GAI, its long-term societal, ethical, and regulatory implications are still evolving. Organizations implementing these tools must monitor emerging legislation, ethical guidelines, and technological safeguards to ensure responsible adoption. Ultimately, while GAI offers extraordinary potential to enhance productivity and creativity, realizing these benefits responsibly will require deliberate, well-informed, and ongoing management of associated risks.

6. GAI: A Techspertive Retrospect

Generative Artificial Intelligence (GAI), often associated with large language models (LLMs), represents a specialized branch of machine learning designed to produce coherent, human-like language. One recent example is Bard, an experimental platform enabling users to collaborate with generative AI driven by an advanced LLM.

To appreciate GAI's significance, it helps to frame it within the broader scope of artificial intelligence. Much of today's AI takes the form of machine learning, where systems typically neural networks acquire skills by analyzing vast datasets. These networks can perform diverse tasks, such as identifying objects in images or predicting the next word in a sentence. Language models, a distinct type of neural network, are refined through large-scale training and evaluated through rigorous testing. By learning patterns from extensive textual input, these models can anticipate subsequent words with increasing sophistication as their training data grows.

Practical applications of language models are already embedded in everyday tools Google's Smart Compose and Smart Reply in Gmail being prime examples and also serve as the foundation for Bard. What makes GAI unique is its ability to generate entirely new material rather than simply replicating learned examples. LLMs, as a subset of GAI, can construct original combinations of words that read as naturally as human writing. The scope of GAI extends beyond text, encompassing capabilities to create images, audio, and even video content. The influence of such technology on creative disciplines is profound. Its introduction has been likened to the arrival of the drum machine in music an innovation that altered the creative landscape without replacing human artistry. By automating repetitive or labor-intensive tasks, GAI can free creators to focus on high-level, imaginative work while serving as a catalyst for new forms of expression.

However, the technology also presents challenges. In educational settings, for example, it may prompt debates over how learning achievements are assessed. This underscores the importance of responsible development. Companies such as Google have articulated AI Principles and built internal oversight structures aimed at minimizing harm, mitigating bias, and reducing toxic outputs.

Looking forward, the outlook for GAI remains optimistic. When deployed thoughtfully, it has the potential to reshape creative workflows, enhance productivity, and inspire novel approaches to problem-solving. Rather than replacing human ingenuity, GAI can amplify it empowering individuals to explore uncharted ideas and tackle emerging challenges with fresh insight.

7. Case Studies Analysis: Machine Learning Mysteries

A recent collaborative study conducted by researchers from MIT, Google Research, and Stanford University delves into the phenomenon of in-context learning within large language models (LLMs) such as OpenAI's GPT-3 and GPT-4. In-context learning describes the remarkable ability of a model to perform unfamiliar tasks after being shown only a handful of examples without undergoing additional training on new datasets.

The investigation centered on theoretical models resembling today's LLMs, with the goal of uncovering how they acquire capabilities without parameter updates. The researchers propose that within the vast architecture of transformer-based networks like GPT-3, there may exist smaller, more specialized linear models. These embedded models could be "trained" on-the-fly using simple learning strategies, all while leaving the main network's parameters untouched. By studying a transformer designed specifically for in-context learning, the team discovered that such a system can effectively "encode" a linear model into its hidden states. Interestingly, this process occurs in the earliest layers of the network. Once in place, the larger model can run and refine the smaller one using contextual information already available, allowing it to handle new tasks dynamically.

The lead authors highlight the efficiency of this approach, noting that it sidesteps the need for complex re-engineering or exhaustive domain-specific data gathering. They argue that in-context learners may not simply be recalling patterns from prior training but are instead demonstrating the capacity to genuinely learn new skills from limited examples. These findings challenge the notion that LLMs merely memorize their training data. Instead, the research positions in-context learning as a powerful capability one that could inform the design of more resource-efficient AI systems capable of adapting to novel problems without costly retraining cycles. As such, this work advances our understanding of modern LLMs and their potential for tackling increasingly sophisticated machine learning challenges.

8. GAI: Creativity Perspectives

From a creative work standpoint, generative AI particularly large language and image-based foundation models has begun reshaping the way content is produced, refined, and distributed across industries. These models open possibilities for automating content creation, enhancing quality, diversifying formats, and tailoring outputs to specific audiences or domains. State-of-the-art systems such as OpenAI's GPT-3 and GPT-4 are capable of generating an extensive range of outputs, including written text, imagery, and even video. Their development requires immense datasets and substantial computational resources; however, once trained, they can be adapted for specialized use cases with comparatively modest datasets. This adaptive capability underscores the continued importance of human participation not only in designing effective prompts but also in critically reviewing and refining AI-generated material.

Marketing and brand communications have emerged as early beneficiaries of generative AI integration. Tools such as Jasper, a GPT-3–powered platform tailored for marketing, can produce blog articles, social media content, and other audience-focused materials. In advertising, image-generation models like DALL·E 2 have already been leveraged by brands such as Heinz and Nestlé to produce creative, visually distinctive campaigns. Similarly, in software development, GPT-3's Codex has demonstrated the ability to translate natural language instructions into code snippets. Experiments by Deloitte have reported productivity gains of around 20% in coding tasks when using Codex, indicating significant potential for accelerating development workflows.

Conversational AI is another domain experiencing improvement through generative models. By enhancing context awareness and dialogue coherence, these systems can serve as more effective chatbots and virtual assistants. Nonetheless, persistent challenges such as mitigating biased outputs or harmful language remain, requiring ongoing refinements in content filtering and model governance.

Generative AI also holds promise in organizational knowledge management. For example, Morgan Stanley is collaborating with OpenAI to fine-tune GPT-3 and GPT-4 for delivering specialized insights in wealth management. Such targeted applications highlight the potential of these systems to act as sophisticated knowledge repositories when trained on domain-specific data.

However, with these opportunities come complex ethical and legal considerations. The proliferation of deepfake content, uncertainties surrounding content ownership, and debates over intellectual property rights underscore the need for proactive policy development. As generative AI matures, its capacity to create diverse forms of media including articles, emails, imagery, and code may become a routine aspect of professional and creative workflows.

While these advancements suggest significant transformations ahead, the broader societal and creative implications remain dynamic and, in some cases, unpredictable. From a personal standpoint, the evolution of generative AI is poised not only to enhance existing creative processes but also to introduce entirely new paradigms of knowledge work bringing with it both exciting opportunities and novel challenges.

9. GAI: Accountability, Ethics, Trust, Public Interests, Proactive Managements

The evolving influence of generative AI within artificial intelligence signals its potential to transform numerous aspects of professional and personal life. Unlike traditional AI systems, generative AI particularly large language models can respond to prompts in natural language with outputs that closely resemble human expression, making the technology accessible even to individuals without deep technical expertise.

In enterprise settings, interest in leveraging generative AI continues to grow, yet this expansion brings heightened attention to ethical, trust, and governance concerns. A critical question emerges: can business users confidently rely on AI-generated outputs? Addressing this requires a clear understanding of the risks, including the possibility of inaccurate or fabricated ("hallucinated") content, which may undermine decision-making. Biases embedded in training data can further distort outputs, raising the risk that users might place undue trust in results that are incomplete, misleading, or incorrect.

Attribution presents another pressing issue. Since generative AI outputs can mirror patterns in their training data, there is a potential for plagiarism and intellectual property infringement. Balancing accurate attribution with necessary human oversight becomes a legal, ethical, and reputational challenge for organizations.

Transparency and explainability are essential to mitigating these risks. Because many end-users lack technical expertise, there is a need for clear, accessible explanations of how generative AI systems function, coupled with organization-wide AI literacy initiatives. This fosters better risk awareness and equips users to critically evaluate AI-generated content.

Accountability remains a core pillar in this discussion. Establishing robust oversight mechanisms ensures that responsibility for generative AI outputs can be traced back to both the system's creators and the enterprise deploying it. Maintaining human involvement in critical review processes safeguards against over-reliance on automated outputs, reinforcing the importance of human judgment, contextual awareness, and ethical scrutiny.

Ultimately, effective governance of generative AI requires proactive strategies that integrate trust, transparency, and accountability into every stage of deployment. This includes implementing ethical guidelines, ensuring rigorous oversight, and fostering a culture of informed use linking AI-generated outcomes directly to the stakeholders responsible for them in the broader, rapidly evolving AI landscape.

10. GAI: Challenges & Future Directions

The ongoing enthusiasm surrounding generative AI often overshadows the reality that its greatest value lies in targeted, domain-specific applications rather than as an all-encompassing solution. While tools such as ChatGPT have captured global attention, their most meaningful impact will likely emerge in specialized contexts, where they can deliver unique ways of accessing and interpreting highly specific information.

Recent developments, such as the creation of ChatGPT plugins by various companies, illustrate this shift. A general-purpose chatbot may not excel in every task, but when applied to focused areas such as travel planning for platforms like Expedia it can provide a competitive advantage in industries where information discovery is central. This raises questions about whether generative AI will disrupt major players like Google or instead represent an "iPhone moment" that subtly reshapes user expectations and behaviors. A likely future involves organizations deploying large language model (LLM)-powered tools that learn from their own data and services, fueling the first wave of business transformation. OpenAI's move to open a waiting list for companies to access ChatGPT plugins signals recognition of the technology's commercial potential. This points to a near-term proliferation of products and interfaces underpinned by generative AI capabilities.

It is important, however, to acknowledge that OpenAI is not the sole gatekeeper of this technology, nor is ChatGPT the only option available. A growing ecosystem includes open-source and self-hosted LLMs, which allow enterprises to integrate AI directly with proprietary data while maintaining privacy and control.

Another notable trend is the shift toward domain-specific language models. By fine-tuning general-purpose LLMs on targeted datasets, organizations can create powerful, specialized information retrieval tools. Potential use cases range from managing product catalogs to enhancing internal knowledge systems demonstrating that generative AI's most practical applications may be highly tailored rather than universally broad.

From this perspective, the future of AI appears less intimidating and more functional. As generative AI becomes embedded in specific domains, it will shed its image as an all-knowing entity and instead be recognized as a context-aware assistant.

GitHub Copilot exemplifies this transition serving as a problem-solving aid for developers, grounded in their existing expertise and workflows. True success for generative AI will come when it is seamlessly integrated into specialized environments, with its limitations understood and its strengths applied pragmatically.

11. Results and Findings

PaLM-E, Google's embodied multimodal language model, demonstrates outstanding performance across diverse robotic settings and vision–language benchmarks. Its evaluation spanned three distinct robotic scenarios each incorporating real-world robots and encompassed tasks such as visual question answering (VQA), image captioning, and general natural language understanding.

When integrated into a robotic control system, PaLM-E operates in tandem with a low-level language-to-action policy that converts textual instructions into executable robot actions. In a kitchen environment, for instance, the model successfully directed a mobile robot to locate and retrieve a bag of chips, maintaining robustness even when the object was intentionally returned to its drawer mid-task. In another case, the model responded to the instruction to fetch an unfamiliar green block by generating a plan that extended beyond the robot's prior training experience.

In a separate tabletop robotic setup, PaLM-E handled complex, long-horizon tasks such as sorting colored blocks into designated corners by analyzing visual inputs and producing detailed sequences of text-based actions. This capability surpassed that of earlier models, especially in multi-step precision tasks. Notably, the model exhibited zero-shot generalization, as demonstrated by successfully pushing red blocks toward a coffee cup despite the absence of prior training on that exact task.

A third evaluation environment, inspired by task and motion planning (TAMP), tested PaLM-E's ability to address combinatorially complex planning challenges. By leveraging visual–language knowledge transfer and incorporating a relatively small amount of expert TAMP planner data, the model delivered effective and generalizable solutions.

From a benchmarking standpoint, PaLM-E also proved competitive with leading vision–language-only architectures. It achieved a state-of-the-art score on the demanding OK-VQA dataset without the need for task-specific fine-tuning. The largest variant, PaLM-E-562B, demonstrated exceptional visual comprehension and extensive world knowledge.
Among its most advanced features are visual chain-of-thought reasoning and multi-image inference enabling it to decompose problem-solving into sequential reasoning steps and to synthesize information from multiple images, even though it was originally trained using single-image prompts. Figures 3, 4, 5 and Table 1 illustrate these capabilities in greater detail, offering visual insight into the model's reasoning and task execution processes.

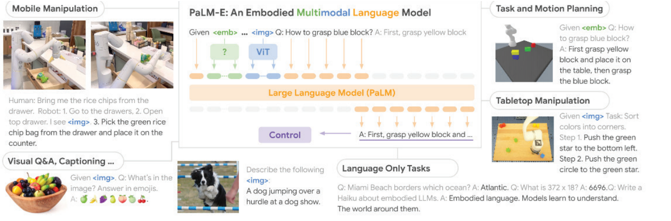

Mobile Manipulation
Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.
PaLM-E: An Embodied Multimodal Language Model
Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block
?
ViT
Large Language Model (PaLM)
Control
A: First, grasp yellow block and ...
Task and Motion Planning
Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.
Tabletop Manipulation
Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.
Visual Q&A, Captioning ...
Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍏🍒.
Describe the following <img>: A dog jumping over a hurdle at a dog show.
Language Only Tasks
Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696.Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

Figure 3. PaLM-E Robotic Environmental Performing Actions

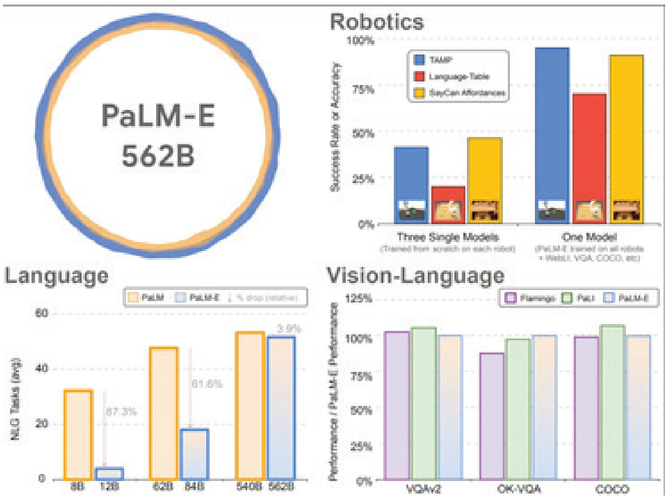
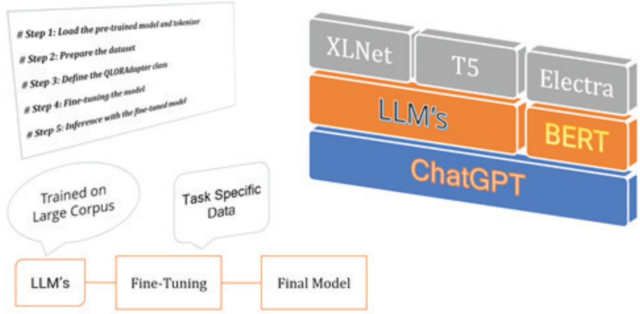


Figure 4. An overview visualization for PaLM-E Performing Actions



```
# Step 1: Load the pre-trained model and tokenizer
# Step 2: Prepare the dataset
# Step 3: Define the QLORAdapter class
# Step 4: Fine-tuning the model
# Step 5: Inference with the fine-tuned model
```

Figure 5. The Experimental Simulation Processing

TABLE 1. PaLM-E Robotic Environmental Performing Actions Evaluations

Evaluation Setting	Task(s)	Key Capability	Notable Outcomes
Kitchen Robot Environment	Object retrieval (e.g., bag of chips)	Robust object recognition & action execution	Successfully retrieved object despite interruptions (object returned to drawer mid-task)
Kitchen Robot Environment	Retrieval of unseen objects (e.g., green block)	Generalization beyond training data	Generated a viable action plan for unfamiliar objects
Tabletop Robot Environment	Sorting blocks by color into specific corners	Long-horizon task planning & precision	Produced accurate, sequential text-based actions for multi-step tasks
Tabletop Robot Environment	Zero-shot task execution (e.g., pushing red blocks to a coffee cup)	Zero-shot generalization	Adapted to novel instructions without prior training
TAMP-Inspired Robot Environment	Complex combinatorial planning tasks	Knowledge transfer from visual–language models	Solved planning problems using limited expert TAMP data
Benchmark Evaluation (OK-VQA)	Visual question answering	State-of-the-art vision–language integration	Achieved highest reported score without task-specific fine-tuning
General Model Capabilities	Multi-image inference	Cross-image reasoning	Integrated information from multiple images despite single-image training
General Model Capabilities	Visual chain-of-thought reasoning	Step-by-step inference	Decomposed problem-solving into logical reasoning stages

12. Discussions and Future Directions

In addressing the evolving landscape of Parameter-Efficient Fine-Tuning (PEFT), several key questions emerge, offering clarity on its goals, mechanisms, and potential impact in the field of Natural Language Processing (NLP).

Q1. What is the primary objective of parameter-efficient fine-tuning?

The central aim of PEFT is to adapt pre-trained language models for specific downstream tasks while significantly reducing the computational and memory demands compared to conventional full-parameter fine-tuning. This approach enables scalability, making advanced models more accessible for a wider range of applications and research environments.

Q2. How does Quantized Low-Rank Adaptation (QLoRA) improve parameter efficiency?

QLoRA integrates quantization into the low-rank adaptation process, allowing model weights to be represented in lower precision formats. This integration preserves model performance while drastically improving memory efficiency, eliminating the need for more complex or resource-intensive quantization strategies.

Q3. What are the main advantages of Low-Rank Adaptation (LoRA)?

LoRA offers several benefits:

- Reduced parameter overhead, requiring only a small fraction of trainable parameters.
- Efficient task-switching, enabling rapid adaptation to multiple tasks without retraining the entire model.
- Minimal impact on inference latency, preserving deployment efficiency. These strengths make LoRA an attractive solution for a broad range of NLP applications.

Q4. How can researchers benefit from PEFT techniques?

PEFT techniques empower researchers to fine-tune large-scale language models on task-specific datasets using limited computational resources. This facilitates experimentation with complex architectures without prohibitive hardware requirements, thus accelerating innovation and broadening participation in cutting-edge NLP research.

Q5. Which models can benefit from QLoRA?

QLoRA's versatility extends to a variety of architectures, including RoBERTa, DeBERTa, GPT-2, and GPT-3. This adaptability ensures that a diverse range of pre-trained models can be optimized efficiently for different application domains.

Future Directions

The continued refinement of PEFT methods, including LoRA and QLoRA, is likely to focus on further reducing resource requirements while enhancing generalization capabilities. Integration with emerging training paradigms, such as self-supervised and multimodal learning, presents promising avenues for extending their applicability beyond NLP into areas such as computer vision, robotics, and domain-specific AI systems. Additionally, community-driven development of open-source PEFT implementations will play a pivotal role in democratizing access to these techniques, ensuring their benefits are broadly realized across both academia and industry.

13. Conclusions

This study highlights PaLM-E as a landmark step toward unifying vision, language, and robotics within a single, generally capable model. By leveraging knowledge transfer between vision–language systems and robotic control, PaLM-E demonstrates the feasibility of addressing diverse tasks traditionally treated as separate domains.

The model's ability to integrate multimodal inputs, maintain language proficiency at scale, and adapt to complex, long-horizon robotic scenarios underscore its potential as a foundation for future multimodal AI systems.

In parallel, the exploration of Parameter-Efficient Fine-Tuning (PEFT) methods, particularly Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), offers practical solutions to the computational and memory constraints that hinder the deployment of large language models (LLMs).

These approaches reduce the number of trainable parameters and memory footprint while preserving, and in some cases enhancing, task performance. Their adaptability across architectures such as RoBERTa, DeBERTa, GPT-2, and GPT-3 reinforces their utility in diverse NLP contexts.

The combined insights from PaLM-E's multimodal integration and the efficiency gains from PEFT techniques point toward a future where advanced AI models can be trained, adapted, and deployed at scale without prohibitive resource demands.

Such advancements promise to accelerate real-world adoption across sectors including robotics, healthcare, accessibility, and human–computer interaction. As research in these areas progresses, continued refinement of both multimodal model architectures and fine-tuning strategies will be pivotal in shaping the next generation of capable, resource-efficient AI systems.

Supplementary information

The various original data sources some of which are not all publicly available, because they contain various types of private information. The available platform provided data sources that support the findings and information of the research investigations are referenced where appropriate.

Acknowledgments

The author would like to acknowledge and enthusiastically thank the GOOGLE Deep Mind Research with its associated pre-prints access platforms. This research was deployed and utilized under the various platforms and provided by GOOGLE Deep Mind which is under the support of the GOOGLE Research and the GOOGLE Research Publications under GOOGLE Gemini platform. Using their provided platform of datasets and database files with digital software layouts consisting of free web access to a large collection of recorded models that are found in research access and its related open-source software distributions which is the implementation and simulation of analytics for the proposed research which was undergone and set in motion. There are many datasets, data models which are resourced and retrieved from a wide variety of GOOGLE service domains. All the data sources and various domains from which data has been included and retrieved for this research are identified, mentioned and referenced where appropriate. However, various original data sources some of which are not all publicly available, because they contain various types of private information. The available platform provided data sources that support the findings and information of the research investigations are referenced where appropriate.

Declarations

Funding: No Funding was provided for the conduction of this research.

Conflict of interest/Competing interests: There are no Conflict of Interest or any type of Competing Interests for this research.

Ethics approval: The author declares no competing interests for this research.

Consent to participate: The author has read and approved the manuscript and have agreed to its publication.

Consent for publication: The author has read and approved the manuscript and have agreed to its publication.

Availability of data and materials: The various original data sources some of which are not all publicly available, because they contain various types of private information. The available platform provided data sources that support the findings and information of the research investigations are referenced where appropriate.

Code availability: Mentioned in details within the Acknowledgements section.

Authors' contributions: Described in details within the Acknowledgements section.

References

Akhtar, Z.B. (2024) 'Unveiling the evolution of generative AI (GAI): a comprehensive and investigative analysis toward LLM models (2021–2024) and beyond', *Journal of Electrical Systems and Information Technology*, 11(1), 22. Available at: https://doi.org/10.1186/s43067-024-00145-1.

Akhtar, Z.B. (2025) 'Beyond perception: a comprehensive investigation into the advancements, challenges & ethical dimensions of AI and computer vision', *Real-World AI Systems*, 1(1), pp. 1–27. Available at: https://doi.org/10.64797/rwas.v1i1.9577.

Akhtar, Z. and Rawol, A. (2025) 'Harnessing artificial intelligence (AI) towards the landscape of big earth data: methods, challenges, opportunities, future directions', *Journal of Geography and Cartography*, 8(1), 10224. Available at: http://dx.doi.org/10.24294/jgc10224

Hamid, S. (2025) ‘Integrating artificial intelligence and multimodality in language education: a systematic review of emerging trends and practices’, *Journal of Social & Organizational Matters*, 4(2), pp. 400–416. Available at: https://doi.org/10.56976/jsom.v4i2.253

Jin, K., Yu, T. and Grzybowski, A. (2025) ‘Multimodal artificial intelligence in ophthalmology: applications, challenges, and future directions’, *Survey of Ophthalmology*, 71(1), pp. 158-167. Available at: https://doi.org/10.56976/jsom.v4i2.253

Tharayil, S.M., Krishnapriya, M.A. and Alomari, N.K. (2025) ‘How multimodal AI and IoT are shaping the future of intelligence’, in *Internet of Things and Big Data Analytics for a Green Environment. Chapman and Hall/CRC*, pp. 138–167. Available at: https://doi.org/10.1201/9781032656830

Fahad, S.A., Zhengkui, D.W., Chet, N.P., Wong, N., Ng, A.B. and See, S. (2025) ‘Advancements and applications of multimodal large language models: integration, challenges, and future directions’, in *Yafooz, W.M., Al-Gumaei, Y. (eds) AI-Driven: Social Media Analytics and Cybersecurity. Springer Nature Switzerland*, pp. 309–336. Available at: https://doi.org/10.1007/978-3-031-80334-5_19

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E. et al. (2025) ‘Taking the next step with generative artificial intelligence: the transformative role of multimodal large language models in science education’, *Learning and Individual Differences*, 118, 102601. Available at: https://doi.org/10.1016/j.lindif.2024.102601

Jacobs, C. (2025) ‘Examining multimodal AI resources in medical education: the role of immersion, motivation, and fidelity in AI narrative learning’, *JMIR Medical Education*, 11(1), e72190. Available at: https://doi.org/10.2196/72190

Lee, G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M. et al. (2025) ‘Multimodality of AI for education: towards artificial general intelligence’, *IEEE Transactions on Learning Technologies*, 18, pp. 666 – 683. Available at: https://doi.org/10.1109/TLT.2025.3574466

Bravo, L., Rodriguez, C., Hidalgo, P. and Angulo, C. (2025) ‘A systematic review on artificial intelligence-based multimodal dialogue systems capable of emotion recognition’, *Multimodal Technologies and Interaction*, 9(3), 28. Available at: https://doi.org/10.3390/mti9030028

Parvin, N., Joo, S.W., Jung, J.H. and Mandal, T.K. (2025) ‘Multimodal AI in biomedicine: pioneering the future of biomaterials, diagnostics, and personalized healthcare’, *Nanomaterials*, 15(12), 895. Available at: https://doi.org/10.3390/nano15120895

Akhtar, Z.B. (2025) ‘Artificial intelligence within medical diagnostics: a multi-disease perspective’, *Artificial Intelligence in Health*, 2(3), pp. 44-62. Available at: https://doi.org/10.36922/aih.5173

Akhtar, Z.B. and Rozario, V.S. (2025) ‘AI perspectives within computational neuroscience: EEG integrations and the human brain’, *Artificial Intelligence and Applications*, 3(2), pp. 145-160. Available at: https://doi.org/10.47852/bonviewaia52024174

Akhtar, Z.B. (2024) ‘The design approach of an artificial intelligent (AI) medical system based on electronical health records (EHR) and priority segmentations’, *Journal of Engineering*, 2024, pp. 1–10. Available at: https://doi.org/10.1049/tje2.12381

Ma, Y., Ye, W., Cui, C., Zhang, H., Xing, S., Ke, F. et al. (2025) ‘Position: prospective of autonomous driving–multimodal LLMs world models embodied intelligence AI alignment and mamba’, in *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1010–1026. Available at: https://openaccess.thecvf.com/content/WACV2025W/LLVMAD/papers/Ma_Position_Prospective_of_Autonomous_Driving_-_Multimodal_LLMs_World_Models_WACVW_2025_paper.pdf

Hawthorne, H. (2025) ‘Advancing artificial intelligence through multimodal learning and cross-disciplinary integration’, *International Journal of Computer Science and Engineering Research and Development*, 15(2), pp. 41–46.

Nayak, B. (2025) ‘The evolution and architecture of multimodal AI systems’, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(1), pp. 1007–1017. Available at: https://doi.org/10.32628/CSEIT251112108

Bland, T. (2025) ‘Author’s reply: examining multimodal AI resources in medical education: the role of immersion, motivation, and fidelity in AI narrative learning’, *JMIR Medical Education*, 11, e72336. Available at: https://doi.org/10.2196/72336

Wu, X. and He, A. (2025) ‘Multimodal information fusion and artificial intelligence approaches for sustainable computing in data centers’, *Pattern Recognition Letters*, 189, pp. 17–22. Available at: https://doi.org/10.1016/j.patrec.2024.12.006

Wang, H., Zhou, M., Jia, X., Wei, H., Hu, Z. and Li, W. et al. (2025) ‘Recent progress on artificial intelligence-enhanced multimodal sensors integrated devices and systems’, *Journal of Semiconductors*, 46(1), 011610. Available at: http://dx.doi.org/10.1088/1674-4926/24090041

Ullah, E., Baig, M.M., Waqas, A., Rasool, G., Singh, R., Shandilya, A. et al. (2025) ‘Multimodal generative AI for anatomic pathology—a review of current applications to envisage the future direction’, *Advances in Anatomic Pathology*, 10–1097. Available at: https://doi.org/10.1097/pap.0000000000000498

Hou, C., Huang, T., Hu, K., Ye, Z., Guo, J. and Zhou, H. (2025) ‘Artificial intelligence-assisted multimodal imaging for the clinical applications of breast cancer: a bibliometric analysis’, *Discover Oncology*, 16(1), 537. Available at: https://doi.org/10.1007/s12672-025-02329-1

Ma, R., Cheng, Q., Yao, J., Peng, Z., Yan, M., Lu, J. et al. (2025) ‘Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses’, *NPJ Digital Medicine*, 8(1), 64. Available at: https://doi.org/10.1038/s41746-025-01461-0

Yang, X.Y., Li, Y.M., Wang, J.Y., Yuheng, J., Yi, Z. and Chen, M. (2025) ‘Utilizing multimodal artificial intelligence to advance cardiovascular diseases’, *Precision Clinical Medicine*, pbaf016. Available at: https://doi.org/10.1093/pcmedi/pbaf016

Areerob, K., Nguyen, V.Q., Li, X., Inadomi, S., Shimada, T., Kanasaki, H. et al. (2025) ‘Multimodal artificial intelligence approaches using large language models for expert-level landslide image analysis’, *Computer-Aided Civil and Infrastructure Engineering*, 40(19), pp. 2900-2921. Available at: https://doi.org/10.1111/mice.13482

Carvalhido, F., Cardoso, H.L. and Cerqueira, V. (2025) ‘Stress-testing of multimodal models in medical image-based report generation’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28), pp. 29251–29252. Available at: https://doi.org/10.1609/aaai.v39i28.35203

Schouten, D., Nicoletti, G., Dille, B., Chia, C., Vendittelli, P., Schuurmans, M. et al. (2025) 'Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications', *Medical Image Analysis*, 105, 103621. Available at: https://doi.org/10.1016/j.media.2025.103621

Huang, S.C., Jensen, M., Yeung-Levy, S., Lungren, M.P., Poon, H. and Chaudhari, A.S. (2025) 'A systematic review and implementation guidelines of multimodal foundation models in medical imaging', *Research Square*, rs-3. Available at: https://doi.org/10.21203/rs.3.rs-5537908/v1

Lu, J., Yang, W., Xiong, Z., Xing, C., Tafazolli, R., Quek, T.Q. and Debbah, M. (2025) 'Generative artificial intelligence-enhanced multimodal semantic communication in internet of vehicles: system design and methodologies', *IEEE Vehicular Technology Magazine*, 20(2), pp. 71-82. Available at: https://doi.org/10.1109/MVT.2025.3545399

Yuan, Y., Li, Z. and Zhao, B. (2025) 'A survey of multimodal learning: methods, applications, and future', *ACM Computing Surveys*, 57(7), pp. 1–34. https://doi.org/10.1145/3713070

Osamor, J., Ashawa, M., Shahrabi, A., Phillip, A. and Iwend, C. (2025) 'The evolution of phishing and future directions: A review', in *Proceedings of the International Conference on Cyber Warfare and Security*, 20(1), pp. 361–368. https://doi.org/10.34190/iccws.20.1.3366

Ryu, S.Y., Choi, J.Y. and Yoo, T.K. (2025) 'Automated detection of retinal artery occlusion in fundus photography via self-supervised deep learning and multimodal interpretability using a multimodal AI chatbot', *Medical & Biological Engineering & Computing*, 63(9), pp. 2679-2691. Available at: https://doi.org/10.1007/s11517-025-03353-7