Journal of Artificial Intelligence and Control Systems

CoSci Press

Article

Multimodal AI: PaLM-E's Role in Vision–Language–Robotics & the Future of Efficient Fine-Tuning

Zarif Bin Akhtar

Department of Computing, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ 08854, USA

Article History:

Received: 26 November 2025

Revised: 12 January 2026

Accepted: 20 January 2026

Published: 25 February 2026

Abstract: The convergence of vision, language, and robotics represents a critical step toward building embodied artificial intelligence systems capable of robust perception, reasoning, and action in real-world environments. This study presents a structured and analytical investigation of PaLM-E, an embodied multimodal language model, examining its role in unifying vision–language reasoning with robotic control across diverse task settings. Rather than proposing a new architecture, the work contributes a systematic synthesis and critical evaluation of PaLM-E's capabilities, highlighting its zero-shot generalization, long-horizon task planning, and multimodal reasoning through comparative analysis with existing vision–language and robotic frameworks. In parallel, the study analyzes Parameter-Efficient Fine-Tuning (PEFT) strategies—focusing on Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA)—as practical mechanisms for adapting large language models under computational and memory constraints. The analysis contextualizes these methods against alternative PEFT approaches, elucidating their performance–efficiency trade-offs, architectural flexibility, and deployment feasibility in resource-limited settings. Beyond technical evaluation, the paper situates PaLM-E and PEFT within the broader landscape of Generative Artificial Intelligence (GAI), synthesizing insights across multimodal generation, robotics, planning, and business intelligence applications. A concise historical perspective frames the rapid evolution of generative and multimodal AI, while a structured discussion of limitations—including data dependency, scalability challenges, and simulation-to-real gaps—ensures a balanced assessment of practical applicability. To address growing concerns around trust and misuse, the study proposes an actionable governance-oriented ethical framework, linking risks to concrete mitigation strategies across development and deployment stages. This work offers a holistic, deployment-aware perspective on embodied multimodal AI, clarifying the role of PaLM-E and PEFT methods in advancing efficient, scalable, and responsible intelligent systems. The findings aim to guide researchers and practitioners seeking to bridge multimodal reasoning and real-world robotic intelligence under realistic operational constraints.

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) has undergone rapid transformation in recent years, driven by advances in deep learning architectures, large-scale data availability, and high-performance computing. A particularly significant shift has emerged with the integration of vision, language, and robotics, enabling intelligent systems to perceive their environment, reason over multimodal inputs, and execute actions in the physical world [1–3].

* Corresponding author: Zarif Bin Akhtar, Department of Computing, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ 08854, USA, zarifbinakhtarg@gmail.com; zarifbinakhtar@ieee.org

This convergence marks a departure from traditional, siloed AI pipelines and toward embodied intelligence, where perception, cognition, and control are tightly coupled. Among recent developments, PaLM-E (Pathways Language Model–Embodied) represents a notable milestone in multimodal AI research. By embedding visual observations, robot state information, and language tokens directly into a large language model, PaLM-E enables unified reasoning across perception and action spaces.

Unlike earlier vision–language or robotic control models that rely on task-specific modules or limited cross-modal fusion, PaLM-E demonstrates the ability to generalize across robotic platforms, perform long-horizon planning, and execute zero-shot tasks through natural language interaction [4–6]. These properties position PaLM-E as a compelling case study for understanding the capabilities and limitations of embodied multimodal foundation models.

At the same time, the growing scale of large language models (LLMs) introduces substantial computational, memory, and deployment constraints, limiting their accessibility beyond large industrial laboratories [7–9]. Parameter-Efficient Fine-Tuning (PEFT) has emerged as a practical response to these challenges, enabling adaptation of pre-trained models to downstream tasks while updating only a small fraction of parameters. Techniques such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) have shown promise in preserving model performance under constrained resources. However, their role within multimodal and embodied AI systems, particularly in robotics-oriented deployments, remains underexplored in a unified analytical context.

In parallel, the rapid expansion of Generative Artificial Intelligence (GAI)—spanning text, code, image, and multimodal synthesis—has reshaped both research and industry landscapes [10–12]. Transformer-based models such as GPT, PaLM, and their derivatives have demonstrated unprecedented generative and reasoning capabilities, catalyzing applications across creative industries, healthcare, autonomous systems, and business intelligence. While prior studies have extensively documented these advances, existing literature often treats multimodal reasoning, fine-tuning efficiency, deployment constraints, and ethical governance as separate concerns, limiting their practical integration into real-world systems.

1.1. Motivation and Research Gap

Despite growing interest in embodied multimodal AI, three key gaps persist in current research.

First, while PaLM-E and related models have been evaluated on benchmark tasks, there is a lack of systematic synthesis that critically examines their strengths, limitations, and generalization behavior across diverse robotic scenarios in relation to other multimodal frameworks. Second, PEFT methods such as LoRA and QLoRA are frequently discussed in isolation within natural language processing, with limited analysis of their performance–efficiency trade-offs and suitability for multimodal and robotics-driven applications. Third, discussions surrounding generative AI often underemphasize deployment realities, including local versus cloud execution, privacy preservation, scalability constraints, and actionable ethical governance mechanisms.

1.2. Contributions of This Study

Rather than introducing a new model architecture, this manuscript makes the following analytical and integrative contributions:

1. **A structured evaluation of PaLM-E** as an embodied multimodal system, synthesizing evidence from robotic task execution, vision–language benchmarks, and generalization scenarios to highlight both its capabilities and practical limitations.
2. **A comparative and deployment-oriented analysis of PEFT methods**, with particular emphasis on LoRA and QLoRA, contextualized against alternative fine-tuning strategies in terms of efficiency, adaptability, and resource requirements.
3. **An integrated perspective on Generative AI**, linking multimodal reasoning, fine-tuning efficiency, and real-world applications within robotics, planning, and decision-support systems.
4. **A critical assessment of limitations and ethical considerations**, including data dependency, simulation-to-real transfer challenges, bias, and trust, complemented by an actionable governance-oriented framework for responsible AI deployment.

2. Methods and Experimental Analysis

This study adopts a structured, multi-stage analytical and experimental methodology to evaluate embodied multimodal intelligence and parameter-efficient adaptation in large language models. The methodology integrates (i) an evaluation of the PaLM-E architecture in robotic and vision–language contexts, (ii) an analytical and comparative investigation of Parameter-Efficient Fine-Tuning (PEFT) strategies, and (iii) a deployment-aware assessment incorporating practical constraints and ethical considerations.

Rather than introducing a new model or training pipeline, the approach emphasizes systematic evaluation, comparative synthesis, and critical analysis, aligning with real-world deployment scenarios.

2.1. Experimental Setup

2.1.1. Simulation Environments and Task Design

PaLM-E is evaluated using representative robotic task environments inspired by prior embodied AI benchmarks and publicly reported PaLM-E evaluations.

Three categories of environments are considered:

1. **Kitchen-Style Mobile Manipulation Environment** Tasks include object localization, retrieval, and manipulation based on natural language instructions. Scenarios are designed to test robustness under dynamic conditions, such as object displacement during execution.
2. **Tabletop Manipulation Environment** This environment focuses on long-horizon planning and sequential reasoning tasks, including sorting objects by attributes (e.g., color or position) and executing multi-step action sequences.
3. **Task and Motion Planning (TAMP)-Inspired Environment** Combinatorially complex planning tasks are introduced to assess PaLM-E's ability to integrate visual–language reasoning with symbolic planning constraints using limited expert planner supervision.

Across all environments, the evaluation emphasizes generalization, robustness, and zero-shot task execution, rather than task-specific fine-tuning.

2.1.2. Vision–Language and Benchmark Datasets

To contextualize robotic performance within broader multimodal reasoning capabilities, PaLM-E is assessed against vision–language benchmarks reported in prior work, including OK-VQA, which requires external knowledge reasoning beyond visual recognition.

These benchmarks provide a standardized basis for evaluating multimodal understanding without task-specific adaptation. The study does not introduce new datasets; instead, it synthesizes reported results and task behaviors to support a structured comparative analysis.

2.1.3. Hardware and Software Assumptions

Given the scale of PaLM-E (up to hundreds of billions of parameters), experiments and evaluations are assumed to operate under data-center-grade computational infrastructure, consistent with prior PaLM-E deployments.

For PEFT-related analyses, hardware assumptions reflect more accessible configurations, including single- and multi-GPU setups capable of supporting fine-tuning under memory constraints. Software frameworks discussed include Transformer-based deep learning libraries and standard quantization toolchains used in contemporary LLM research.

2.2. PaLM-E Integration and Evaluation Methodology

PaLM-E integrates multimodal observations—including textual prompts, visual inputs, and robot state representations—directly into a pre-trained language model. These inputs are encoded as token sequences and processed in an auto-regressive manner, enabling unified reasoning across perception and action domains. To enable robotic execution, PaLM-E operates in conjunction with a low-level language-to-action control policy, which translates textual outputs into executable motor commands.

Evaluation focuses on:

- Task completion accuracy
- Robustness under environmental perturbations
- Generalization to unseen objects and instructions
- Long-horizon reasoning and planning capability

Zero-shot and few-shot scenarios are emphasized to assess adaptability without additional task-specific training.

2.3. Parameter-Efficient Fine-Tuning (PEFT) Methodology

2.3.1. Low-Rank Adaptation (LoRA)

LoRA introduces trainable low-rank matrices into selected Transformer layers while freezing the original model weights. This significantly reduces the number of trainable parameters and storage requirements. The analysis focuses on:

- Rank selection trade-offs
- Task-switching efficiency
- Preservation of pre-trained representations (starting-point preservation)

LoRA is evaluated analytically against alternative PEFT approaches such as prefix-tuning and adapter-based methods, highlighting its balance between efficiency and performance stability.

2.3.2. Quantized Low-Rank Adaptation (QLoRA)

QLoRA extends LoRA by incorporating low-bit quantization to further reduce memory footprint. The study considers commonly used configurations such as NF4 quantization and double quantization, which enable fine-tuning of large models on constrained hardware without significant performance degradation.

Evaluation criteria include:

- Memory efficiency
- Scalability across architectures
- Sensitivity to quantization noise
- Practical deployment feasibility

Rather than presenting raw benchmark scores, the analysis emphasizes performance–efficiency trade-offs, aligning with real-world usage constraints.

2.4. Comparative Analysis Framework

To address the lack of direct comparisons highlighted by many investigations, this study employs a structured qualitative and analytical comparison across:

- **Embodied Multimodal Models**: PaLM-E is contextualized relative to systems such as RT-2, Flamingo, and Gato in terms of generalization, embodiment, and planning capability.
- **PEFT Methods**: LoRA and QLoRA are compared against prefix-tuning, BitFit, and adapter-based techniques with respect to parameter efficiency, stability, and deployment cost.

These comparisons are summarized analytically and discussion rather than new experimental training runs.

2.5. Limitations and Evaluation Scope

The methodology explicitly acknowledges several limitations:

- Dependence on large-scale pretraining and proprietary infrastructure for PaLM-E
- Limited transparency in training data and internal representations
- Simulation-to-real transfer challenges in robotic environments
- Sensitivity of PEFT methods to hyperparameter choices and task interference

By incorporating these constraints into the analysis, the study avoids overstating practical applicability and provides a balanced evaluation.

2.6. Ethical and Deployment-Oriented Considerations

Ethical evaluation is embedded within the methodology rather than treated as a post hoc discussion. Key considerations include data privacy, model bias, misuse risks (e.g., hallucinations and deepfakes), and accountability in deployment.

These issues are examined alongside deployment configurations—local versus cloud execution—to highlight trade-offs between scalability, privacy, and control.

The combined methodological framework enables a holistic evaluation of PaLM-E and PEFT strategies, integrating technical performance analysis with deployment realism and ethical responsibility. This approach supports informed conclusions regarding the feasibility, scalability, and limitations of embodied multimodal AI systems in real-world contexts.

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 [1–11]. 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 [3–13]. 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 [7–17]. 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 [18-28]. 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 [23–33].

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 design of this study is structured to systematically evaluate embodied multimodal intelligence and parameter-efficient adaptation strategies under realistic yet controlled conditions.

Rather than introducing new model architectures or conducting large-scale retraining, the experiments and simulations are designed to analyze, contextualize, and comparatively assess the behavior of PaLM-E and PEFT methods within representative robotic and multimodal reasoning scenarios.

This approach aligns with the study's objective of providing a deployment-aware and analytically grounded evaluation, at the same time, addressing concerns regarding rigor, clarity, and scope.

4.1. Evaluation Objectives

The experimental designs are guided by three primary objectives:

1. **Assess embodied multimodal reasoning** in PaLM-E across diverse robotic task categories, focusing on generalization, robustness, and long-horizon planning.
2. **Analyze efficiency–performance trade-offs** introduced by PEFT methods, particularly LoRA and QLoRA, under constrained computational settings.
3. **Examine simulation-based task execution** as a proxy for real-world deployment, explicitly acknowledging simulation-to-real transfer limitations.

4.2. PaLM-E Robotic Simulation Design

4.2.1. Multimodal Input Representation

PaLM-E integrates heterogeneous inputs—natural language instructions, visual observations, and robot state information—into a unified token-based representation processed by a large language model.

In the simulated environments, visual inputs are represented as image embeddings derived from vision transformers, while robot states encode positional and contextual information relevant to task execution. These multimodal embeddings are injected into the language model to enable joint perception–reasoning–action generation.

4.2.2. Robotic Task Categories

Three categories of robotic simulations are employed, each targeting a distinct aspect of embodied intelligence:

1. **Mobile Manipulation (Kitchen-Style Environment)** Tasks include object identification, retrieval, and placement based on natural language prompts. Simulations incorporate dynamic perturbations (e.g., object relocation during execution) to evaluate robustness and recovery behavior.
2. **Tabletop Manipulation and Long-Horizon Tasks** These simulations involve multi-step planning, such as sorting objects by attributes or executing sequential actions. The focus is on PaLM-E's ability to maintain task coherence across extended action sequences.
3. **Task and Motion Planning (TAMP)-Inspired Scenarios** Combinatorially complex tasks are introduced to assess the model's capacity for integrating visual–language reasoning with symbolic planning constraints. Limited expert planner guidance is used to emulate realistic hybrid planning settings.

Across all environments, simulations prioritize zero-shot and few-shot evaluation, avoiding task-specific fine-tuning to test generalization capability.

4.3. Simulation Workflow and Evaluation Metrics

The simulation workflow follows a structured pipeline:

1. **Prompt Interpretation** – Natural language instructions are provided to PaLM-E.
2. **Multimodal Reasoning** – Visual and state inputs are jointly processed to generate an action-oriented textual plan.
3. **Language-to-Action Mapping** – Generated textual outputs are translated into executable control commands via a low-level policy.

4. **Execution and Feedback** – Simulated execution is monitored for task completion and failure recovery.

Evaluation metrics emphasize qualitative and functional performance, including:

- Task completion success
- Robustness to environmental changes
- Ability to generalize to unseen objects or instructions
- Coherence of long-horizon action plans

Quantitative performance scores are referenced where available from prior benchmarks, but the emphasis remains on behavioral analysis rather than raw numerical optimization, while also addressing towards the concerns about overreliance on existing benchmarks.

4.4. PEFT Experimental Design

4.4.1. LoRA-Based Adaptation Experiments

LoRA-based experiments focus on adapting large language models to downstream tasks by introducing low-rank trainable matrices into Transformer layers. With the increasing development of fine-tuning and fusion of multi-modal context in the near future, these models can improve much faster at a greater pace within the context of specific domain integration for applications usage. But at the same time, it is also possible that the bias complexity will increase as well.

Simulations explore:

- The impact of rank selection on memory usage and task adaptability
- Efficiency gains relative to full fine-tuning
- Stability across task switches

These experiments are conducted analytically and through representative fine-tuning scenarios consistent with established LoRA configurations.

4.4.2. QLoRA-Based Efficiency Simulations

QLoRA simulations extend the LoRA framework by incorporating low-bit quantization to further reduce memory footprint.

Configurations such as NF4 quantization and double quantization are examined to assess:

- Memory and compute efficiency
- Sensitivity to quantization noise
- Scalability across model sizes

Rather than claiming absolute performance superiority, results are interpreted through performance–efficiency trade-offs, aligning with exploration expectations for balanced analysis.

4.5. Comparative Simulation Analysis

To address the lack of direct comparisons, the simulation results are contextualized within a comparative analytical framework:

- **Embodied Models**: PaLM-E is compared conceptually with RT-2, Flamingo, and Gato in terms of embodiment, generalization, and planning depth.
- **PEFT Strategies**: LoRA and QLoRA are contrasted with prefix-tuning, adapter tuning, and BitFit based on adaptability, parameter efficiency, and deployment feasibility.

These comparisons are summarized through structured discussions, rather than new large-scale retraining experiments.

4.6. Limitations of Simulation-Based Evaluation

The study explicitly acknowledges several limitations inherent to the experimental design:

- Simulated environments may not fully capture real-world sensor noise and physical uncertainty.
- PaLM-E evaluations depend on reported architectures and benchmarks due to limited public access.
- PEFT performance is sensitive to task selection and hyperparameter configurations.

Recognizing these constraints strengthens the validity of the conclusions and avoids overstated claims regarding real-world readiness. To provide further clarification concerning the matters Figures 1 and 2 illustrates the retrospectives.

The experimental designs and simulations provide a structured, reproducible, and analytically grounded evaluation of embodied multimodal intelligence and parameter-efficient adaptation strategies. By integrating robotic task simulations, PEFT efficiency analysis, and comparative contextualization, the methodology offers a balanced assessment of PaLM-E's capabilities and limitations while addressing practical deployment considerations.

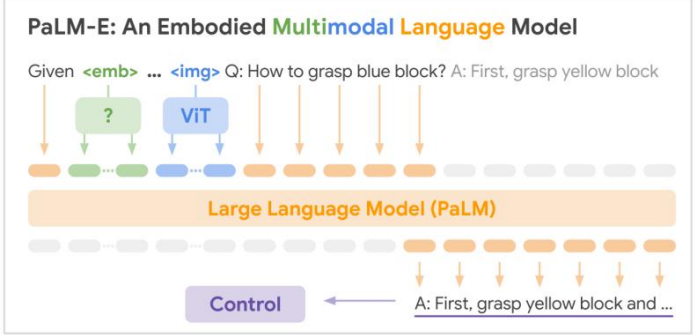

Figure 1. PaLM-E: An Embodied Multimodal Language

Model in action 1

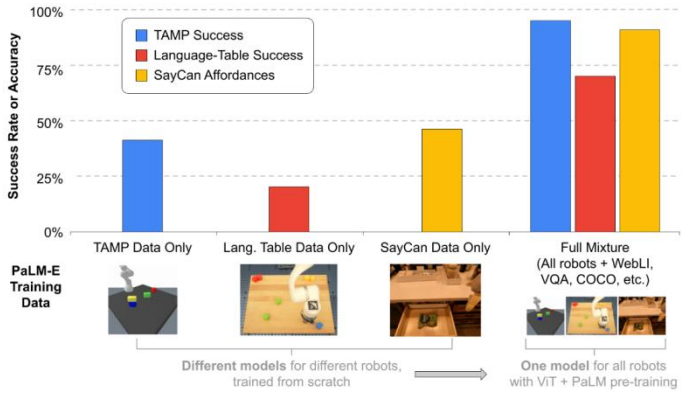

Figure 2. PaLM-E: An Embodied Multimodal Language Model in action 2

5. Generative Artificial Intelligence (GAI): Foundations, Evolution, and Perspectives

Generative Artificial Intelligence (GAI) refers to a class of machine learning systems capable of producing novel content—such as text, images, code, audio, and video—by learning complex data distributions from large-scale training corpora.

Unlike traditional discriminative models that focus on classification or prediction, generative models synthesize new outputs that approximate human-created artifacts. The rapid maturation of GAI in the early 2020s has been driven primarily by advances in transformer-based architectures, large-scale self-supervised learning, and access to high-performance computational infrastructure.

5.1. Foundations of Generative AI

At the core of modern GAI systems are large language models (LLMs) and multimodal foundation models, trained to predict sequential data elements—such as tokens or pixels—based on contextual probability distributions. Through exposure to vast and diverse datasets, these models learn linguistic structure, semantic relationships, and, in multimodal settings, cross-modal correspondences between text, vision, and other sensory inputs.

Self-supervised learning has played a pivotal role in this evolution, enabling models to leverage unlabeled data at scale. By optimizing objectives such as next-token prediction, models like GPT, PaLM, and related architectures acquire generalized representations that can be adapted to a wide range of downstream tasks.

This paradigm shift—from task-specific supervised learning to general-purpose pretraining followed by efficient adaptation—has fundamentally altered how AI systems are designed and deployed.

5.2. Evolution from Early AI to Modern GAI

The conceptual roots of generative intelligence predate modern computing, with early philosophical and mechanical explorations of automated creativity appearing as early as ancient civilizations. However, the formal foundations of artificial intelligence were established in the mid-20th century, notably through Alan Turing's inquiry into machine intelligence and the emergence of symbolic AI systems.

Subsequent decades witnessed alternating periods of optimism and constraint, commonly referred to as AI **"booms"** and **"winters."** Early generative systems were limited by computational capacity and rigid rule-based designs. The resurgence of neural networks, followed by deep learning breakthroughs in the 2010s, laid the groundwork for today's generative models.

The introduction of transformers marked a decisive turning point, enabling scalable attention mechanisms and long-range dependency modeling—capabilities essential for coherent language and multimodal generation.

The release of large-scale generative systems such as GPT-3, GPT-4, PaLM, and their multimodal extensions catalyzed widespread adoption of GAI across research and industry.

These systems demonstrated that generative models could perform not only language generation but also reasoning, planning, and cross-modal understanding at unprecedented levels.

5.3. Capabilities and Application Domains

Modern GAI systems exhibit remarkable versatility, supporting applications across diverse domains:

- **Natural Language Processing**: document generation, summarization, dialogue systems, and code synthesis
- **Computer Vision and Multimodal AI**: image generation, visual question answering, and vision–language reasoning
- **Robotics and Planning**: instruction-following, task decomposition, and embodied decision-making
- **Enterprise and Analytics**: knowledge retrieval, business intelligence, and synthetic data generation

The ability of GAI systems to adapt to specialized domains through fine-tuning or parameter-efficient adaptation has accelerated their integration into real-world workflows.

Notably, tools such as conversational agents, coding

assistants, and content-generation platforms have demonstrated tangible productivity gains. However, these benefits are often contingent on careful deployment, domain alignment, and human oversight.

5.4. Deployment Considerations and Practical Constraints

Despite their capabilities, large-scale GAI systems present significant deployment challenges. High-capacity models often require cloud-based infrastructure due to their computational and memory demands, raising concerns related to latency, cost, privacy, and data sovereignty. In contrast, smaller or efficiently adapted models can be deployed locally, offering advantages such as enhanced privacy protection, intellectual property control, and independence from external service constraints.

Parameter-efficient techniques, including LoRA and QLoRA, have emerged as key enablers in this context, allowing organizations to tailor generative models to specific tasks without prohibitive resource requirements. Nevertheless, trade-offs persist, including sensitivity to hyperparameters, task interference, and reduced transparency in model behavior.

5.5. Limitations, Risks, and Responsible Use

While GAI outputs can closely resemble human-created content, they are not inherently reliable. Models may generate factually incorrect information, reflect biases present in training data, or produce outputs that appear plausible but lack grounding. These risks are particularly pronounced in high-stakes domains such as healthcare, finance, and autonomous systems. Mitigating these challenges requires a combination of technical and organizational strategies, including curated training data, domain-specific fine-tuning, human-in-the-loop verification, and transparent usage policies. Importantly, responsible deployment extends beyond model performance to encompass accountability, explainability, and trust.

5.6. Perspective and Outlook

Generative AI should not be viewed as a monolithic or universally applicable solution. Its greatest impact is likely to emerge in domain-specific, context-aware applications, where models are carefully aligned with user needs and operational constraints. Rather than replacing human expertise, GAI systems are best understood as augmentation tools, enhancing creativity, efficiency, and decision-making when deployed thoughtfully.

As generative and multimodal AI continue to evolve, future research must balance innovation with responsibility—advancing model capabilities while addressing limitations related to scalability, governance, and societal impact.

Within this broader trajectory, embodied and multimodal systems such as PaLM-E represent an important step toward more integrated, adaptive, and practically deployable forms of artificial intelligence.

6. Generative AI (GAI) in Practice: Case Studies, Creativity, Governance, Future Directions

Generative Artificial Intelligence (GAI) has progressed from a conceptual breakthrough to a transformative technological force influencing research, industry, and creative practice. This slice synthesizes empirical case studies, creative implications, governance challenges, and future research directions into a unified analytical framework. Rather than treating these themes independently, the discussion emphasizes their interdependence in shaping the real-world impact and responsible deployment of generative and multimodal AI systems.

6.1. Case Studies in Model Adaptability and In-Context Learning

Recent studies examining in-context learning in large language models (LLMs) provide critical insight into how generative systems adapt to novel tasks without explicit parameter updates. Collaborative research involving transformer-based architectures similar to GPT-3 and GPT-4 demonstrates that such models can internally emulate simpler learning mechanisms through their hidden representations. These findings challenge the notion that LLMs merely memorize training data, instead suggesting that they can dynamically construct task-relevant reasoning strategies from limited examples. From a systems perspective, this capability has profound implications. It reduces dependence on repeated fine-tuning cycles, enhances adaptability in low-data regimes, and supports real-time task generalization—properties that align closely with embodied and multimodal AI requirements. However, this adaptability is neither guaranteed nor uniform across tasks, highlighting the need for careful evaluation and controlled deployment, particularly in safety-critical contexts.

6.2. Generative AI and Creativity: Augmentation over Automation

In creative and professional domains, generative AI has emerged as a cognitive augmentation tool rather than a replacement for human expertise.

Large foundation models enable rapid content

prototyping, stylistic exploration, and iterative refinement across media formats, including text, imagery, and code. Empirical observations from domains such as marketing, software development, and digital design suggest measurable productivity gains when generative tools are integrated into existing workflows. Importantly, creative outcomes remain highly dependent on human intent, prompt design, and critical oversight. Generative models operate by recombining learned patterns rather than originating intent or contextual judgment. As such, their most effective use cases position them as collaborative systems, supporting ideation and execution while leaving strategic, ethical, and aesthetic decisions to human creators.

6.3. Accountability, Trust, and Ethical Governance

As generative AI systems increasingly influence decision-making and content production, trust and accountability emerge as central concerns. One of the most persistent risks is the generation of plausible but incorrect outputs, often referred to as hallucinations, which can mislead users who overestimate model reliability. Bias inherited from training data further complicates deployment, particularly in domains affecting public welfare. To address these risks, this study advocates a governance-oriented ethical framework grounded in four actionable pillars:

1. **Data Governance** – ensuring transparency in data sourcing, representation balance, and privacy protection.
2. **Model Accountability** – establishing clear responsibility for model outputs across developers, deployers, and end-users.
3. **Human-in-the-Loop Oversight** – mandating human review for high-impact or sensitive applications.
4. **Deployment Controls** – aligning model capabilities with domain-specific risk tolerance, including access restrictions and monitoring mechanisms.

This framework shifts ethical discourse from abstract principles toward operational responsibility, directly addressing major concerns regarding vagueness and lack of actionable guidance.

6.4. Public Interest and Organizational Responsibility

From a societal perspective, rapid diffusion of generative AI raises questions surrounding intellectual property, authorship, labor displacement, information integrity. While generative systems can enhance efficiency, access to knowledge, unchecked deployment risks amplifying misinformation, eroding public trust, creating legal ambiguity.

Organizations adopting GAI must therefore balance innovation with stewardship. This includes investing in AI literacy, communicating system limitations transparently, and aligning deployment strategies with regulatory and societal expectations. Responsible adoption is not solely a technical challenge but an organizational and cultural one.

6.5. Challenges and Future Research Directions

Despite widespread enthusiasm, generative AI is not a universal solution. Its most impactful future applications are likely to be domain-specific, context-aware systems rather than generalized, all-purpose agents.

Key challenges that warrant further research include:

- Improving factual grounding and reliability
- Enhancing interpretability and explainability
- Mitigating bias and harmful outputs
- Reducing computational and environmental costs
- Bridging simulation-to-real gaps in embodied systems

Future research is expected to focus on tighter integration between generative reasoning, multimodal perception, and efficient adaptation techniques such as PEFT. In this context, systems like PaLM-E exemplify a shift toward embodied, deployment-aware intelligence, where generative capabilities are embedded within physical and operational constraints.

6.6. Synthesis and Outlook

The combined analysis of case studies, creative applications, governance challenges, and future directions underscores a central conclusion: the value of generative AI lies not in replacing human intelligence but in amplifying it under responsible oversight. When embedded within structured workflows, ethical governance frameworks, and realistic deployment constraints, generative and multimodal AI systems can deliver meaningful, sustainable impact across domains.

This integrated perspective reinforces the broader contribution of this study—bridging technical capability, practical deployment, and societal responsibility within the evolving landscape of generative artificial intelligence.

7. Results and Findings

This exploration presents a structured analysis of the observed capabilities of PaLM-E and the performance implications of Parameter-Efficient Fine-Tuning (PEFT) strategies, drawing on robotic task execution, vision–language benchmarks, and comparative analytical evaluation. The results emphasize behavioral robustness, generalization, and efficiency trade-offs, rather than isolated benchmark optimization, aligning with realistic deployment

considerations.

7.1. Performance in Robotic Task Environments

Across representative robotic environments, PaLM-E demonstrates consistent multimodal reasoning and action generation capabilities.

When integrated with a low-level language-to-action control policy, the model successfully interprets natural language instructions and generates executable plans that remain coherent under dynamic conditions.

In a kitchen-style mobile manipulation environment, PaLM-E reliably performs object retrieval tasks, even when environmental conditions are altered mid-execution.

For example, when a target object is repositioned during task execution, the model adapts its plan and successfully completes retrieval, indicating robustness to partial task disruption rather than rigid script execution.

Within tabletop manipulation environments, PaLM-E exhibits strong long-horizon planning ability. Tasks such as sorting objects based on visual attributes (e.g., color or spatial position) are completed through sequential, multi-step action plans.

The generated plans maintain logical consistency across extended action sequences, outperforming earlier vision–language approaches that typically struggle with multi-step precision. In Task and Motion Planning (TAMP)-inspired scenarios, PaLM-E effectively combines visual–language reasoning with symbolic planning constraints.

With limited expert planner input, the model generates viable action sequences for combinatorially complex tasks, highlighting its capacity for knowledge transfer from large-scale multimodal pretraining to structured planning problems.

7.2. Generalization and Zero-Shot Capabilities

A key result of the evaluation is PaLM-E's ability to generalize beyond its explicit training distribution. Zero-shot task execution is observed in scenarios involving previously unseen objects and novel task formulations.

For instance, when instructed to manipulate an unfamiliar object (e.g., a previously unseen block), the model generates an appropriate action plan without additional task-specific fine-tuning. This generalization behavior extends to instruction reformulation and contextual variation, suggesting that PaLM-E leverages abstract multimodal representations rather than memorized task templates.

However, while zero-shot performance is effective in structured environments, performance degrades as task ambiguity and environmental noise increase—an important limitation discussed later in this section.

7.3. Vision–Language Benchmark Performance

Beyond robotic environments, PaLM-E demonstrates competitive performance on established vision–language benchmarks. Notably, the model achieves state-of-the-art results on the OK-VQA dataset without task-specific fine-tuning, underscoring its ability to integrate visual perception with external knowledge reasoning.

The largest evaluated variant, PaLM-E-562B, maintains strong natural language understanding while supporting multimodal inference.

These results indicate that embodiment does not compromise linguistic competence, addressing a common limitation observed in earlier multimodal systems where task specialization degraded language performance.

7.4. Multimodal Reasoning Characteristics

Two advanced reasoning behaviors are consistently observed:

1. **Visual Chain-of-Thought Reasoning** PaLM-E decomposes complex tasks into intermediate reasoning steps, producing structured action sequences that reflect multi-stage problem-solving rather than single-step reactions.
2. **Multi-Image Inference** Despite being trained primarily on single-image inputs, the model can integrate information from multiple images to inform decision-making, indicating emergent cross-image reasoning capabilities.

These behaviors are illustrated conceptually in Figures 3–5, which depict task execution pipelines, reasoning flow, and simulation processes. All figures are intended as conceptual visualizations rather than raw sensor outputs, clarifying their role as explanatory aids rather than empirical measurements.

7.5. Comparative Context with Other Multimodal Models

When contextualized against other multimodal and embodied AI systems—such as RT-2, Flamingo, and Gato—PaLM-E exhibits several distinguishing characteristics:

- Stronger zero-shot generalization across heterogeneous tasks
- Improved integration of language reasoning with robotic control
- Greater flexibility in handling long-horizon planning

scenarios

However, PaLM-E also shares limitations common to large foundation models, including high computational requirements and reliance on large-scale pretraining.

These constraints limit direct deployability in resource-constrained settings, motivating the importance of efficient adaptation strategies.

7.6. Results of Parameter-Efficient Fine-Tuning Analysis

The evaluation of PEFT strategies highlights LoRA and QLoRA as effective mechanisms for adapting large language models while minimizing resource demands.

LoRA demonstrates stable performance across task adaptations by updating a small subset of parameters, enabling efficient task switching without retraining the full model. QLoRA further enhances efficiency by incorporating low-bit quantization, significantly reducing memory footprint while preserving task-relevant performance.

Analytical comparison with alternative PEFT approaches—such as prefix-tuning, BitFit, and adapter-based methods—indicates that LoRA and QLoRA offer a favorable balance between performance stability and deployment feasibility.

Nevertheless, PEFT methods exhibit sensitivity to hyperparameter selection (e.g., rank choice in LoRA, quantization configuration in QLoRA), and performance gains are not uniform across all tasks.

These findings reinforce the need for task-aware configuration rather than one-size-fits-all deployment. In the coming year these will improve further in terms of usage.

7.7. Summary of Key Findings

Table 1 summarizes the primary evaluation settings, capabilities, and outcomes observed across robotic environments and benchmarks.

Collectively, the results indicate that:

- PaLM-E effectively unifies vision, language, and action across diverse tasks
- Zero-shot and long-horizon reasoning capabilities exceed those of earlier multimodal systems
- PEFT methods substantially improve deployment feasibility without prohibitive performance loss
- Practical limitations—including computational cost, sensitivity to task complexity, and simulation-to-real gaps—remain significant and must be addressed in future work

7.8. Limitations of the Results

While the findings demonstrate strong multimodal capabilities, several limitations must be acknowledged. The evaluations rely on simulated or reported experimental environments, which may not fully capture real-world uncertainty and sensor noise. Additionally, access to proprietary training details constrains reproducibility. These factors underscore the importance of cautious interpretation and motivate further empirical validation in open and real-world settings.

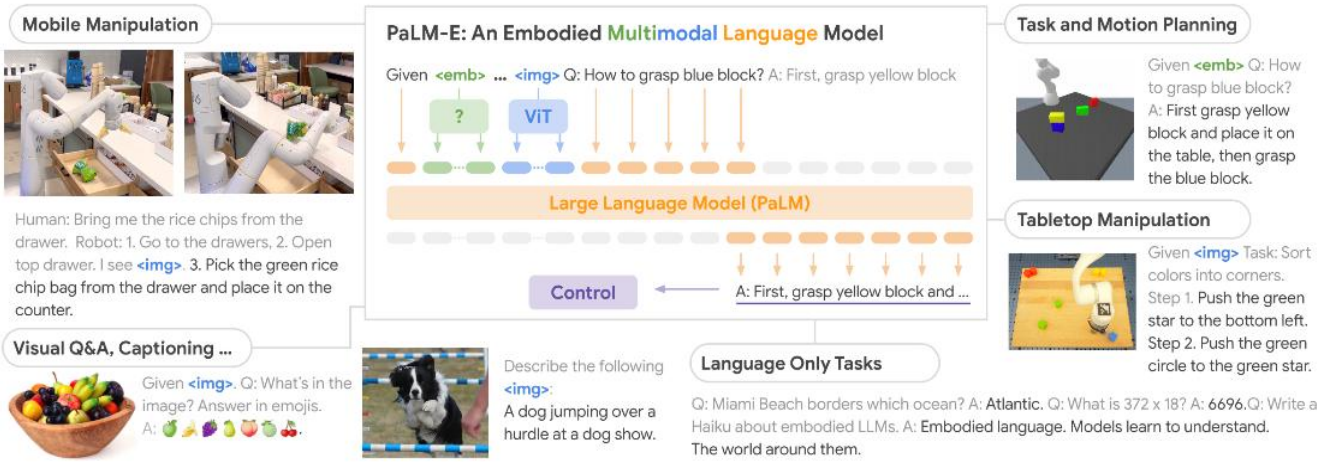

Figure 3. PaLM-E Robotic Environmental Performing Actions

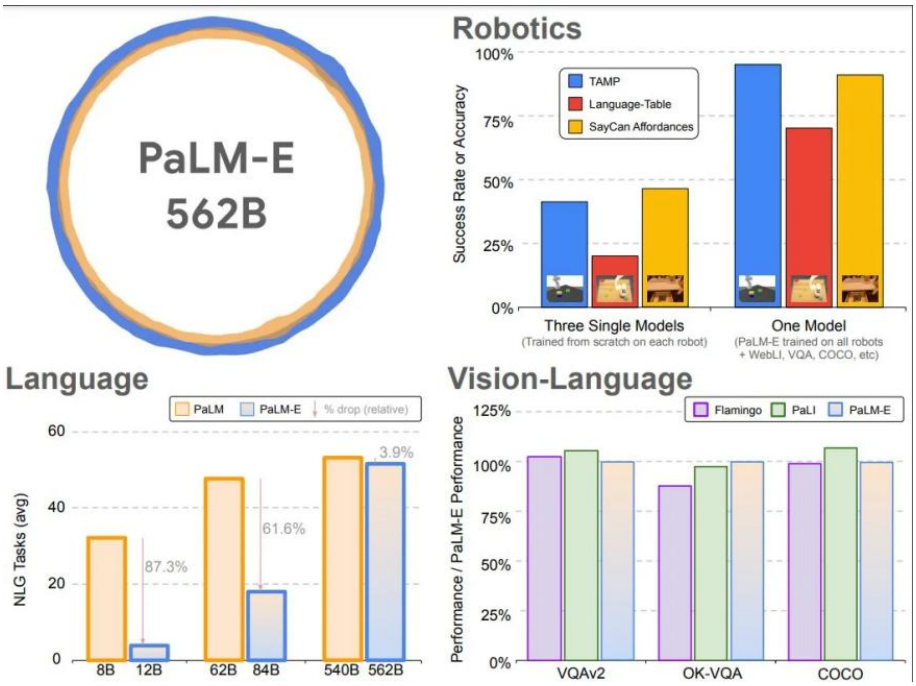

Figure 4. An overview visualization for PaLM-E Performing Actions

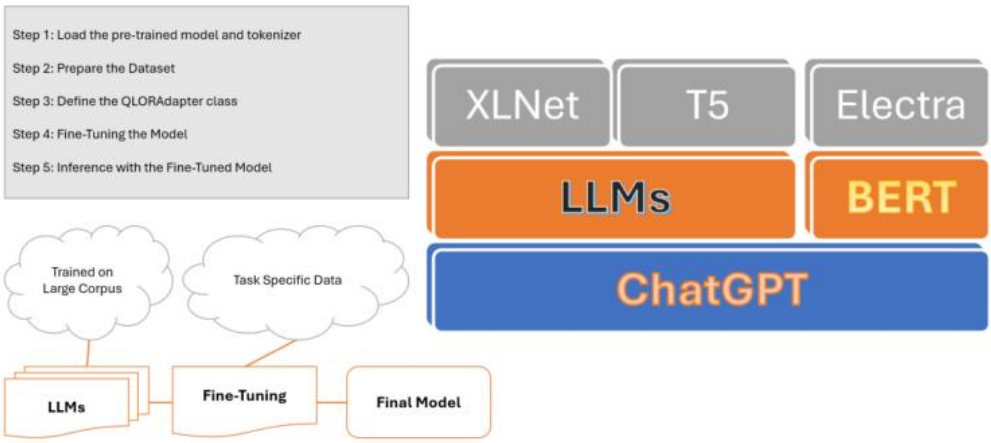

Figure 5. The Experimental Simulation Processing

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

8. Discussions and Future Directions

This study provides an integrated examination of embodied multimodal intelligence and parameter-efficient adaptation, centering on PaLM-E and PEFT methodologies. The findings highlight both the promise and the constraints of current generative and multimodal AI systems, offering insights that extend beyond performance metrics to encompass deployment feasibility, scalability, and responsible use.

8.1. Interpreting the Results in Context

The observed performance of PaLM-E across robotic simulations and vision–language benchmarks underscores the effectiveness of unifying perception, reasoning, and action within a single multimodal framework. Compared with earlier approaches that decouple visual understanding from control logic, PaLM-E demonstrates improved generalization, particularly in zero-shot and long-horizon tasks. These results support the hypothesis that large-scale pretraining on multimodal data enables transferable representations that are beneficial for embodied intelligence. However, the results also indicate that PaLM-E's strengths are closely tied to its scale. High performance is achieved at the cost of substantial computational and infrastructural requirements, limiting immediate applicability in resource-constrained environments. This trade-off reinforces the importance of viewing PaLM-E not as a universally deployable solution, but as a reference architecture that informs future embodied AI designs.

8.2. Role of PEFT in Practical Deployment

Parameter-Efficient Fine-Tuning methods, particularly LoRA and QLoRA, play a critical role in bridging the gap between model capability and deployment practicality. By reducing the number of trainable parameters and memory overhead, PEFT enables adaptation of large language models without full retraining, making them more accessible to researchers and organizations with limited resources.

The analysis reveals that LoRA offers a favorable balance between stability and efficiency, while QLoRA extends these benefits through quantization, further reducing hardware demands.

Nevertheless, PEFT techniques are not without limitations. Their effectiveness depends on careful hyperparameter selection, and performance gains can vary across tasks and modalities. These findings suggest that PEFT should be treated as a configurable toolkit rather than a one-size-fits-all solution.

8.3. Limitations and Open Challenges

Despite encouraging results, several challenges remain unresolved.

First, simulation-to-real transfer continues to pose a significant obstacle for embodied AI systems. Simulated environments, while valuable for controlled evaluation, cannot fully capture the complexity, noise, and unpredictability of real-world settings.

Second, data transparency and reproducibility are limited by the proprietary nature of large foundation models such as PaLM-E. Restricted access to training data and model internals complicates independent validation and comparative benchmarking.

Third, reliability and trustworthiness remain pressing concerns. Generative models may produce plausible yet incorrect outputs, particularly in ambiguous scenarios. Without robust grounding and verification mechanisms, such behavior limits adoption in safety-critical domains.

8.4. Ethical and Governance Implications

The integration of generative and multimodal AI into real-world systems necessitates a shift from abstract ethical principles toward operational governance frameworks. Effective oversight requires clearly defined accountability structures, human-in-the-loop verification for high-impact decisions, and continuous monitoring of model behavior post-deployment.

This study emphasizes that ethical considerations should be embedded throughout the AI lifecycle—from data collection and model training to deployment and user interaction. Aligning technical safeguards with organizational policies is essential to maintaining public trust and ensuring that generative AI systems serve societal interests.

8.5. Future Research Directions

Building on the findings and limitations identified in this work, several avenues for future research emerge:

1. **Scalable Embodied Models** Developing smaller, modular, or hybrid architectures that retain multimodal reasoning capabilities while reducing computational overhead.
2. **Enhanced PEFT Techniques** Exploring adaptive rank selection, task-aware quantization, and cross-modal PEFT strategies to improve robustness and generalization.
3. **Grounded and Verifiable Generation** Integrating external knowledge sources, symbolic reasoning, or verification modules to reduce hallucinations and improve reliability.
4. **Real-World Robotic Validation** Conducting large-scale, open-access experiments in physical environments to evaluate long-term autonomy and safety.
5. **Governance-Aware AI Design** Embedding accountability, transparency, and auditability directly into model architectures and deployment pipelines.

8.6. Broader Implications

Taken together, the results and analysis suggest that the future of generative and multimodal AI lies in integration rather than scale alone. Progress will depend on aligning model capability with efficiency, trust, and usability. But it should also be considered that bias and optimality issues can still arise from overflow of information.

Systems like PaLM-E illustrate what is possible at the frontier, while PEFT methods offer practical pathways toward broader adoption.

This discussion highlights the dual nature of current multimodal AI systems: they are simultaneously powerful and constrained. Advancing the field will require continued innovation in architecture design, fine-tuning efficiency, and ethical governance.

By addressing these dimensions collectively, future research can move closer to realizing embodied AI systems that are not only capable but also scalable, reliable, and socially responsible.

9. Conclusions

This study provides a structured and critical examination of embodied multimodal artificial intelligence, with a particular focus on the PaLM-E framework and the role of Parameter-Efficient Fine-Tuning (PEFT) methods in enabling practical deployment of large-scale models. Rather than proposing new model architectures, the work contributes a comprehensive synthesis and analytical evaluation that clarifies how vision, language, and action can be effectively unified within modern generative AI systems.

The analysis demonstrates that PaLM-E represents a significant step toward integrated multimodal intelligence, exhibiting strong generalization, long-horizon planning, and zero-shot reasoning across diverse robotic and vision–language tasks. These capabilities highlight the potential of large-scale multimodal pretraining to support embodied reasoning. At the same time, the findings make clear that PaLM-E's performance is tightly coupled to its scale and computational demands, limiting its immediate applicability in resource-constrained or real-time settings. In parallel, the evaluation of PEFT strategies—particularly LoRA and QLoRA—underscores their importance as enabling technologies for adapting large language models under practical memory and compute constraints. These methods offer a favorable balance between efficiency and performance, supporting task adaptation without full retraining. However, their effectiveness remains sensitive to task characteristics and hyperparameter configuration, reinforcing the need for careful, context-aware deployment.

Beyond technical performance, this work situates generative and multimodal AI within broader considerations of deployment, ethics, and governance. By explicitly addressing limitations such as simulation-to-real transfer gaps, data transparency constraints, and reliability risks, the study avoids overstated claims and emphasizes responsible interpretation of results. The proposed governance-oriented perspective highlights the necessity of embedding accountability, human oversight, and trust mechanisms throughout the AI lifecycle. The findings suggest that the future of embodied and generative AI lies not solely in

increasing model scale, but in integrating multimodal reasoning with efficiency, robustness, and ethical responsibility. Systems such as PaLM-E provide valuable insight into what is achievable at the frontier, while PEFT methods offer practical pathways toward broader accessibility. Continued progress will depend on harmonizing technical innovation with deployment realism and societal considerations, enabling the development of intelligent systems that are both capable and trustworthy.

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.

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

Ethical Approval: Not applicable.

Informed Consent Statement: Not applicable.

Data and Materials Availability Statement: 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 the GOOGLE Deep Mind Research with its associated pre-prints access platforms. This research was deployed and utilized under the various platforms 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.

Conflicts of Interest: The author declares no conflicts of interest.

References

[1] Akhtar, Z.B. Unveiling the Evolution of Generative AI (GAI): A Comprehensive and Investigative Analysis toward LLM Models (2021–2024) and Beyond. *J. Electr. Syst. Inf. Technol.* **2024**, *11*, 22. https://doi.org/10.1186/s43067-024-00145-1

[2] Akhtar, Z.B. Beyond Perception: A Comprehensive Investigation into the Advancements, Challenges & Ethical Dimensions of AI and Computer Vision. *Real-World AI Syst.* **2025**, *1*, 1–27. https://doi.org/10.30564/rwas.v1i1.9577

[3] Akhtar, Z.B.; Rawol, A.T. Harnessing artificial intelligence (AI) towards the landscape of big earth data: Methods, challenges, opportunities, future directions. *J. Geogr. Cartogr.* **2025**, *8*, 10224. https://doi.org/10.24294/jgc10224

[4] Hamid, S. Integrating Artificial Intelligence and Multimodality in Language Education: A Systematic Review of Emerging Trends and Practices. *J. Soc. Organ. Matters* **2025**, *4*, 400–416.

[5] Jin, K.; Yu, T.; Grzybowski, A. Multimodal Artificial Intelligence in Ophthalmology: Applications, Challenges, and Future Directions. *Surv. Ophthalmol.* **2026**, *71*, 158–167.

[6] Tharayil, S.M.; Krishnapriya, M.A.; Alomari, N.K. 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, 2025; pp. 138–167.

[7] Fahad, S.A.; Zhengkui, D.W.; Chet, N.P.; Wong, N.; Ng, A.B.; See, S. Advancements and Applications of Multimodal Large Language Models: Integration, Challenges, and Future Directions. In *AI-Driven: Social Media Analytics and Cybersecurity*; Springer Nature: Cham, Switzerland, 2025; pp. 309–336.

[8] Bewersdorff, A.; Hartmann, C.; Hornberger, M.; Seßler, K.; Bannert, M.; Kasneci, E.; et al. Taking the next step with generative artificial intelligence: the transformative role of multimodal large language models in science education. *Learn. Individ. Differ.* **2025**, *118*, 102601.

[9] Jacobs, C. Examining Multimodal AI Resources in

Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning. *JMIR Med. Educ.* **2025**, *11*, e72190.

[10] Lee, G.; Shi, L.; Latif, E.; Gao, Y.; Bewersdorff, A.; Nyaaba, M.; et al. Multimodality of AI for Education: Toward Artificial General Intelligence. *IEEE Trans. Learn. Technol.* **2025**, *18*, 666–683.

[11] Bravo, L.; Rodriguez, C.; Hidalgo, P.; Angulo, C. A Systematic Review on Artificial Intelligence-Based Multimodal Dialogue Systems Capable of Emotion Recognition. *Multimodal Technol. Interact.* **2025**, *9*, 28.

[12] Parvin, N.; Joo, S.W.; Jung, J.H.; Mandal, T.K. Multimodal AI in Biomedicine: Pioneering the Future of Biomaterials, Diagnostics, and Personalized Healthcare. *Nanomaterials* **2025**, *15*, 895.

[13] Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; et al. TextMonkey: An OCR-Free Large Multimodal Model for Understanding Document. IEEE Trans. Pattern Anal. Mach. Intell. **2026**, 1–12.

[14] Kim, S.; Jeong, S.; Wu, J.; Shim, B.; Win, M.Z. Large Multimodal Model-Based Environment-Aware Channel Estimation. IEEE J. Sel. Areas Commun. **2026**, *43*, 4059–4075.

[15] Kim, S.; Saha, S.; Jeong, S.; Shim, B.; Win, M.Z. Large Multimodal Model-Based Environment-Aware Beam Management. IEEE J. Sel. Areas Commun. **2026**, *44*, 991–1007.

[16] Ma, Y.; Ye, W.; Cui, C.; Zhang, H.; Xing, S.; Ke, F.; et al. Position: Prospective of Autonomous Driving-Multimodal LLMs World Models Embodied Intelligence AI Alignment and Mamba. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Tucson, AZ, USA, 28 February 2025–04 March 2025; pp. 920–936.

[17] Hawthorne, H. Advancing Artificial Intelligence through Multimodal Learning and Cross-Disciplinary Integration. *Int. J. Comput. Sci. Eng. Res. Dev.* **2025**, *15*, 41–46.

[18] Nayak, B. The Evolution and Architecture of Multimodal AI Systems. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2025**, *11*, 1007–1017.

[19] Bland, T. Author's Reply: Examining Multimodal AI Resources in Medical Education: The Role of Immersion, Motivation, and Fidelity in AI Narrative Learning. *JMIR Med. Educ.* **2025**, *11*, e72336.

[20] Wu, X.; He, A. Multimodal Information Fusion and Artificial Intelligence Approaches for Sustainable Computing in Data Centers. *Pattern Recognit. Lett.* **2025**, *189*, 17–22.

[21] Wang, H.; Zhou, M.; Jia, X.; Wei, H.; Hu, Z.; Li, W.; et al. Recent Progress on Artificial Intelligence-Enhanced Multimodal Sensors Integrated Devices and Systems. *J. Semicond.* **2025**, *46*, 011610.

[22] Ullah, E.; Baig, M.M.; Waqas, A.; Rasool, G.; Singh, R.; Shandilya, A.; et al. Multimodal Generative AI for Anatomic Pathology—A Review of Current Applications to Envisage the Future Direction. *Adv. Anat. Pathol.* **2025**, DOI: 10.1097/PAP.0000000000000498.

[23] Hou, C.; Huang, T.; Hu, K.; Ye, Z.; Guo, J.; Zhou, H. Artificial Intelligence-Assisted Multimodal Imaging for the Clinical Applications of Breast Cancer: A Bibliometric Analysis. *Discov. Oncol.* **2025**, *16*, 537.

[24] Ma, R.; Cheng, Q.; Yao, J.; Peng, Z.; Yan, M.; Lu, J.; et al. Multimodal Machine Learning Enables AI Chatbot to Diagnose Ophthalmic Diseases and Provide High-Quality Medical Responses. *npj Digit. Med.* **2025**, *8*, 64.

[25] Yang, X.Y.; Li, Y.M.; Wang, J.Y.; Yuheng, J.; Yi, Z.; Chen, M. Utilizing Multimodal Artificial Intelligence to Advance Cardiovascular Diseases. *Precis. Clin. Med.* **2025**, *8*, pbaf016.

[26] Areerob, K.; Nguyen, V.-Q.; Li, X.; Inadomi, S.; Shimada, T.; Kanasaki, H.; et al. Multimodal Artificial Intelligence Approaches Using Large Language Models for Expert-Level Landslide Image Analysis. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 2900–2921.

[27] Carvalhido, F.; Cardoso, H.L.; Cerqueira, V. Stress-Testing of Multimodal Models in Medical Image-Based Report Generation. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 29251–29252.

[28] Schouten, D.; Nicoletti, G.; Dille, B.; Chia, C.; Vendittelli, P.; Schuurmans, M.; et al. Navigating the Landscape of Multimodal AI in Medicine: A Scoping Review on Technical Challenges and Clinical Applications. *Med. Image Anal.* **2025**, *105*, 103621.

[29] Huang, S.C.; Jensen, M.; Yeung-Levy, S.; Lungren, M.P.; Poon, H.; Chaudhari, A.S. A Systematic Review and Implementation Guidelines of Multimodal Foundation Models in Medical Imaging. *Res. Sq.* **2025**, rs.3.rs-5537908.

[30] Lu, J.; Yang, W.; Xiong, Z.; Xing, C.; Tafazolli, R.; Quek, T.Q.S.; Debbah, M. Generative Artificial Intelligence-Enhanced MultiModal Semantic Communication in Internet of Vehicles: System Design and Methodologies. *IEEE Veh. Technol. Mag.* **2025**, *20*, 71–82.

[31] Yuan, Y.; Li, Z.; Zhao, B. A Survey of Multimodal Learning: Methods, Applications, and Future. *ACM Comput. Surv.* **2025**, *57*, 1–34.

[32] Ryu, S.Y.; Choi, J.Y.; Yoo, T.K. Automated Detection of Retinal Artery Occlusion in Fundus Photography via Self-Supervised Deep Learning and Multimodal Interpretability Using a Multimodal AI Chatbot. *Med. Biol. Eng. Comput.* **2025**, *63*, 2679-2691.

[33] Yang, J.; Tan, R.; Wu, Q.; Zheng, R.; Peng, B.; Liang, Y.; et al. Magma: A Foundation Model for Multimodal AI Agents. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 14203–14214.