Article

Generative artificial intelligence (GAI): From large language models (LLMs) to multimodal applications towards fine tuning of models, implications, investigations

Zarif Bin Akhtar

Department of Computing, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ 08854-4141, USA; zarifbinakhtarg@gmail.com, zarifbinakhtar@ieee.org

CITATION

Akhtar ZB. Generative artificial intelligence (GAI): From large language models (LLMs) to multimodal applications towards fine tuning of models, implications, investigations. Computing and Artificial Intelligence. 2025; 3(1): 1498.
https://doi.org/10.59400/cai.v3i1.1498

ARTICLE INFO

Received: 3 July 2024
Accepted: 17 October 2024
Available online: 4 November 2024

Abstract: This research explores the transformative integration of artificial intelligence (AI), robotics, and language models, with a particular emphasis on the PaLM-E model. The exploration aims to assess PaLM-E's decision-making processes and adaptability across various robotic environments, demonstrating its capacity to convert textual prompts into very precise robotic actions. In addition, the research investigates Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), providing a historical overview of PEFT and highlighting their significance in enhancing task performance while reducing the number of trainable parameters. The broader scope of Generative AI is examined through an analysis of influential models like GPT-3, GPT-4, Copilot, Bard, LLaMA, Stable Diffusion, Midjourney, and DALL-E. These models' abilities to process natural language prompts and generate a wide range of outputs are thoroughly investigated. The research traces the historical evolution of AI, from its roots in science fiction to its practical applications today, with a focus on the rise of Generative AI in the 21st century. Furthermore, the research delves into the various modalities of Generative AI, covering applications in text, code, images, and more, and assesses their real-world impact on robotics, planning, and business intelligence. The implications of synthetic data generation for business analytics are also explored. The research inspects within both software and hardware landscapes, comparing local deployment on consumer-grade hardware along with cloud-based services, and underscores the benefits of local model deployment in terms of privacy protection, intellectual property security, and censorship resistance. Ethical considerations are central to this research, addressing concerns related to privacy, security, societal impact, biases, and misinformation. The research proposes ethical guidelines for the responsible development and deployment of AI technologies. Ultimately, this work reveals the deep interconnections between vision, language, and robotics, pushing the boundaries of AI capabilities and providing crucial insights for future AI model development and technological innovation. These findings are intended to guide the field through the emerging challenges of the rapidly evolving Generative AI landscape.

Keywords: artificial intelligence (AI); computer vision; deep learning (DL); generative artificial intelligence (GAI); large language models (LLMs); machine learning (ML); models fine tuning; robotics

1. Introduction

In the rapidly advancing field of artificial intelligence (AI), the convergence of vision, language, and robotics is emerging as a critical area of exploration, driving the development of intelligent systems capable of interacting with the world in more holistic and meaningful ways [1–3].

This research investigates the PaLM-E model, a pioneering effort in multimodal AI designed to bridge the gap between perception, language understanding, and robotic control. PaLM-E's ability to navigate complex, real-world scenarios and its adaptability across various tasks position it as a significant advancement in the integration of AI with robotics. The exploration extends its focus to large language models (LLMs) and introduces the concept of Parameter-Efficient Fine-Tuning (PEFT), a paradigm shift in the adaptation of LLMs for specialized tasks. Techniques such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) are explored in depth, offering insights into how these methods optimize the use of computational resources while maintaining or enhancing task performance. The discussion includes an analysis of additional PEFT methods like T-Few, AdaMix, and MEFT, highlighting the delicate balance between efficiency and effectiveness in LLM adaptation. Generative artificial intelligence (Generative AI) represents a significant evolution in the field, with the ability to produce text, images, and multimodal outputs that have broad applications across industries [4–6]. This research delves into the transformative impact of transformer-based models such as GPT-3, Copilot, Bard, and LLaMA, alongside text-to-image generation systems like Stable Diffusion, Midjourney, and DALL-E. As these technologies gain traction in areas ranging from art and creative writing to healthcare and finance, the research critically examines both the opportunities and ethical challenges associated with their widespread use [7–9]. The investigations also provide a historical context, tracing the evolution of AI from its conceptual origins in the mid-20th century through to its current state as a driver of innovation in the 21st century. By reflecting on the contributions of pioneers like Alan Turing and the development of early automated systems, the research underscores the philosophical and ethical debates that have shaped AI's trajectory [10–12]. The rise of Generative AI in recent years is presented as the latest chapter in this ongoing narrative, with a focus on its applications in robotics, planning, and business intelligence [13–15]. Furthermore, the research examines the software and hardware ecosystems that support Generative AI, comparing the benefits and limitations of local deployments versus cloud-based services.

This analysis highlights the importance of accessibility, scalability, and the protection of privacy and intellectual property in the deployment of AI technologies. Through a detailed exploration of these themes, the research aims to provide a comprehensive understanding of the current state of Generative AI, its potential future directions, and the ethical considerations that must guide its development and implementation.

2. Methods and experimental analysis

This research adopts a multi-faceted approach to evaluate the performance and adaptability of the PaLM-E model in robotic environments, as well as to explore the broader implications of Parameter-Efficient Fine-Tuning (PEFT) and Generative AI technologies.

Phase 1: Evaluation of PaLM-E in Robotic Environments

The initial phase focuses on assessing PaLM-E's capabilities in a variety of robotic scenarios. A diverse set of tasks will be formulated, ranging from simple actions to complex, long-horizon maneuvers, simulating real-world conditions to evaluate PaLM-E's decision-making processes. A robust testing framework will be developed, incorporating benchmarks that simulate dynamic and unpredictable environments. The integration of PaLM-E with low-level language-to-action policies will be central to this phase, enabling the translation of textual prompts into precise robotic actions. The model's adaptability will be further tested by introducing adversarial disturbances to evaluate its robustness and generalization to tasks not encountered during the training phase. This aspect of the research is crucial for understanding PaLM-E's potential in transfer learning and its effectiveness in unforeseen scenarios, drawing on methodologies established in existing literature on robotic AI integration.

Phase 2: Exploration of Parameter-Efficient Fine-Tuning (PEFT)

The research then delves into PEFT, with detailed background research and available knowledge focusing on techniques like Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA). The historical development of PEFT will be traced, highlighting key milestones and the challenges that have shaped its evolution. The implementation of LoRA in large language models (LLMs) will be analyzed, with particular attention to the starting point preservation hypothesis, which plays a critical role in reducing the number of trainable parameters without sacrificing performance. Following this, the introduction of QLoRA will be explored, demonstrating how quantization enhances parameter efficiency. This slice will provide a comprehensive understanding of PEFT's role in optimizing LLMs, supported by previous studies that have documented its impact on computational resource management.

Phase 3: Analysis of Generative AI Technologies

In the third phase, the research shifts to an in-depth examination of Generative AI, specifically transformer-based models like GPT-3, GPT-4, Copilot, Bard, LLaMA, Stable Diffusion, Midjourney, and DALL-E. The study will assess these models' capabilities in processing natural language prompts and generating diverse outputs across multiple modalities, including text, code, and images. This analysis will be contextualized within the broader historical development of AI, tracing its evolution from speculative fiction to its current status as a transformative force in various industries. The research will explore both unimodal and multimodal systems, emphasizing their real-world applications in fields such as robotics, planning, and business intelligence, and will investigate the role of Generative AI in synthetic data generation, with a particular focus on its implications for business analytics.

Phase 4: Examination of Software, Hardware, and Ethical Considerations

The research will then inspect the integration of Generative AI features into commercial products, assessing the accessibility and usability of these technologies on consumer devices. The scalability of Generative AI models will be evaluated, comparing local deployment on consumer-grade hardware with cloud-based services. Special attention will be given to the advantages of local model deployment, including privacy protection, intellectual property safeguards, and the avoidance of rate limiting and censorship. This phase will also incorporate an ethical dimension, critically examining privacy and security concerns associated with Generative AI, particularly in relation to deepfakes and synthetic media. The exploration will propose ethical guidelines for the responsible development and deployment of these technologies, addressing societal impacts, biases, misinformation, and manipulation concerns. This ethical exploration will be grounded in existing frameworks and will contribute to the ongoing discourse on AI ethics.

Synthesis and Future Directions

Finally, the research will synthesize the findings from each phase, highlighting the interconnectedness of vision, language, and robotics in pushing the boundaries of AI capabilities. The implications of this research for the future development of AI models and technologies will be discussed, with recommendations for future research aimed at addressing emerging challenges and opportunities in the rapidly evolving field of Generative AI. To provide a better understanding, **Figure 1** illustrates the visualization concerning the matters.

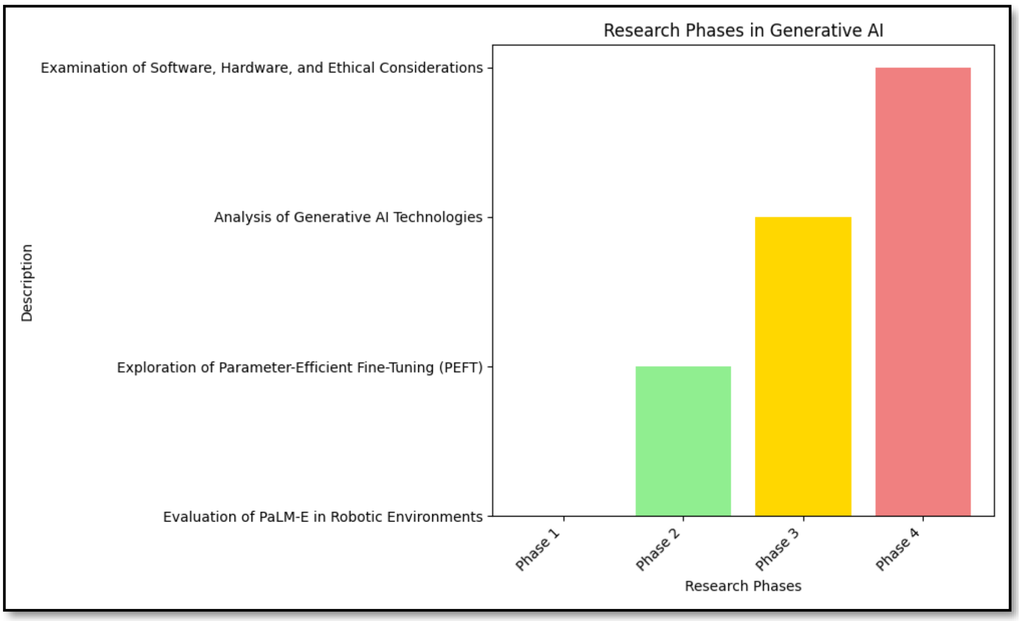


Figure 1. An overall visualization of the research exploration experimentations.

2.1. Background research and available knowledge explorations

Generative artificial intelligence (Generative AI, or GenAI) represents a significant advancement in the capabilities of AI systems, particularly in their ability to produce text, images, and other forms of media through generative models. These models, which learn patterns and structures from vast datasets, are capable of generating new data that mirrors the characteristics of the training data. The early 2020s witnessed remarkable progress in transformer-based deep neural networks, leading to the emergence of Generative AI systems [1–15]. Notable examples include large language model (LLM) chatbots and text-to-image AI art systems, which have garnered widespread attention for their ability to accept and process natural language prompts. The application of Generative AI spans a broad spectrum of industries, highlighting its versatility and transformative potential. In the fields of art, writing, and scriptwriting, Generative AI has been used to create content that pushes the boundaries of creativity. In software development, tools like GitHub Copilot assist programmers by generating code snippets, streamlining the development process.

The healthcare and finance sectors have also embraced Generative AI for tasks such as predictive analytics and automated reporting. Additionally, the gaming, marketing, and fashion industries are leveraging these technologies to enhance user experiences and design processes. The early 2020s saw a significant surge in investment from major technology companies, including Microsoft, Google, and Baidu, as well as numerous smaller firms, reflecting the growing interest in the potential of Generative AI [16–20]. Despite its promising applications, the development of Generative AI has raised concerns regarding its potential misuse. The ability of these models to generate realistic content has led to fears of cybercrime, the creation of fake news, and the production of deepfakes—manipulated media that can deceive viewers [11–22]. These concerns underscore the need for ethical guidelines and regulatory frameworks to ensure the responsible development and deployment of Generative AI technologies.

The historical evolution of artificial intelligence provides essential context for understanding the development of Generative AI [21–33]. The field of AI was formally established as an academic discipline in 1956, but the roots of automated creativity can be traced back much further. Ancient Greek civilization explored the concept of automated art, and over the centuries, the development of creative automatons laid the groundwork for the sophisticated Generative AI systems of today.

One of the seminal contributions to the conceptual foundation of AI was Alan Turing's 1950 paper, which posed fundamental questions about machine reasoning and laid the groundwork for future advancements in the field. Over the decades, AI has experienced several waves of progress and periods of optimism, leading to the development of Generative AI planning systems and, more recently, advanced generative models capable of performing complex tasks [34–49].

Generative AI now operates across various modalities, including text, code, images, audio, video, molecules, robotics, planning, and business intelligence. Large language models, such as GPT-4 and PaLM, typically run on powerful data center computers, but there has been significant progress in developing smaller models that can operate on more accessible devices, such as smartphones, embedded systems, and personal computers.

This accessibility has facilitated the integration of Generative AI into a wide range of products, from conversational agents like ChatGPT to programming tools like GitHub Copilot. Furthermore, many of these models are available as open-source software, enabling broader experimentation and application [34–49].

One of the key advantages of running Generative AI models locally, as opposed to relying solely on cloud-based services, is the enhanced protection of privacy. Local deployment mitigates the risks associated with data exposure and provides users with greater control over their intellectual property. Additionally, running models locally can help avoid issues related to rate limiting and censorship that may arise with cloud services.

However, the largest models, which often contain hundreds of billions of parameters, still require the computational power of data center computers and are typically accessed through cloud services.

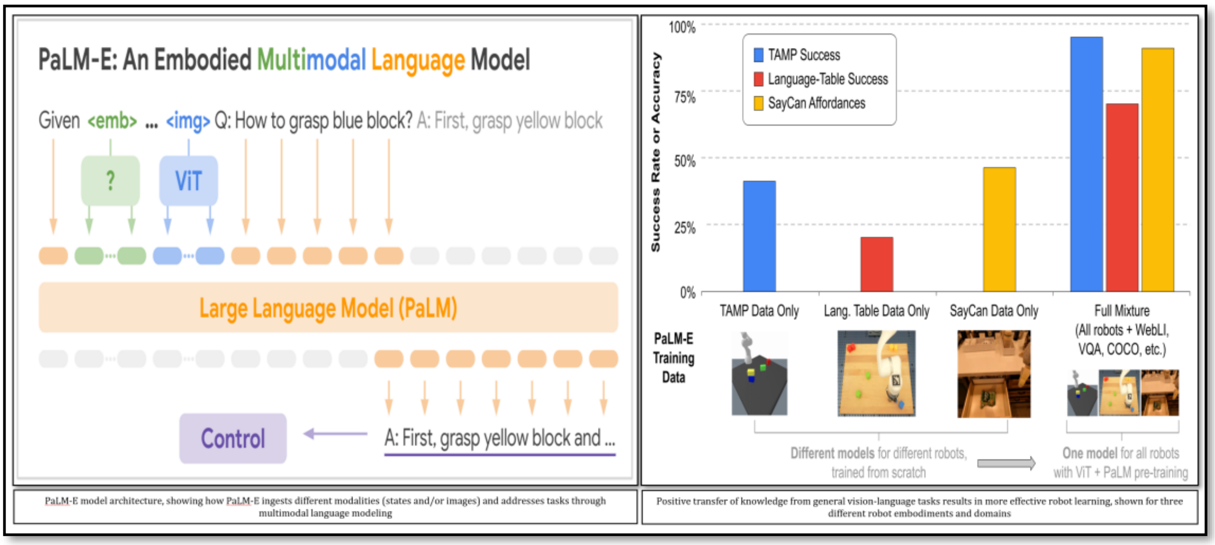
This research highlights the profound impact of Generative AI on various industries while also acknowledging the potential challenges and ethical considerations that accompany its rapid development. As Generative AI continues to evolve, it is crucial to address these concerns to ensure that the technology is harnessed for the benefit of society [34–49].

2.2. Experimental designs and simulation investigations

This research undertakes a detailed exploration of PaLM-E, a generalist robotics model developed by Google, which addresses the significant challenges posed by the lack of large-scale datasets in robotics. PaLM-E's innovative architecture integrates sensor data from robotic agents directly with a powerful language model, PaLM, to create a comprehensive visual-language model. The experimental design is structured to evaluate PaLM-E's effectiveness in performing a range of tasks across multiple robots and modalities, such as processing images, robot states, and neural scene representations. These experiments aim to demonstrate PaLM-E's ability to transfer knowledge from large-scale training data to various robotic applications, thereby improving the model's performance in both vision-language tasks and robotic decision-making [32–34]. The first phase of the experimental design involves setting up robotic environments where PaLM-E is tested on diverse tasks, ranging from basic operations to complex, long-horizon maneuvers. A simulation framework is employed to create realistic robotic scenarios, enabling the assessment of PaLM-E's adaptability and decision-making capabilities in dynamic environments.

The integration of PaLM-E with low-level language-to-action policies is crucial in this phase, allowing the model to translate textual prompts into executable robot actions. The experiments further introduce adversarial disturbances to test the model's robustness and its ability to generalize to tasks that it was not explicitly trained for. This phase also examines how visual-language data enhances the model's performance in robotic tasks, thereby underscoring the potential of PaLM-E to function as an efficient and effective generalist model for robotics. The second phase of the research delves into Parameter-Efficient Fine-Tuning (PEFT) for Large Language Models (LLMs), focusing on methodologies such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA). The experiments are designed to evaluate how PEFT techniques optimize LLMs for specific tasks while managing computational and memory requirements effectively. A step-by-step simulation process is employed, where pretrained LLMs undergo fine-tuning to adapt to specialized tasks. This phase highlights the benefits of PEFT, including reduced memory usage and lower storage costs, while also addressing potential challenges like increased training time. In the context of LoRA, the experiments introduce trainable low-rank matrices into each layer of the Transformer architecture during the fine-tuning process. This technique is designed to minimize the number of trainable parameters, thus reducing the computational burden without compromising task performance. The experimental design elucidates LoRA's working principles, particularly its focus on starting point preservation and the use of low-rank matrices as adapters.

These experiments aim to showcase LoRA's efficiency in task-switching and its applicability to real-time applications, making it an optimal choice for fine-tuning large language models in resource-constrained environments. Building on the findings from LoRA, the research introduces QLoRA, an extension of LoRA that incorporates quantization techniques to achieve further parameter efficiency. The experiments simulate NF4 quantization and Double Quantization processes, demonstrating how QLoRA reduces memory requirements while maintaining, or even enhancing, model performance. The results from this phase are critical in understanding how QLoRA can be employed across various LLMs, offering a versatile and highly efficient approach to parameter-efficient fine-tuning. The experimental design also includes detailed simulations to compare the performance of PaLM-E and various PEFT techniques against existing models. These simulations provide insights into the strengths and limitations of each method, allowing for a comprehensive evaluation of their potential impact on language processing and robotics tasks. The discussions and findings from these simulations are intended to equip researchers and practitioners with a nuanced understanding of PEFT, LoRA, and QLoRA, guiding their application in real-world scenarios where efficient fine-tuning of large language models is essential. By systematically evaluating PaLM-E and PEFT techniques through carefully designed experiments and simulations, this research contributes valuable insights into the practical implementation of advanced AI models in robotics and language processing domains. The findings emphasize the potential of these techniques to significantly enhance the efficiency and effectiveness of AI-driven systems, paving the way for future developments in Generative AI and robotics. To provide a better understanding, **Figures S1** (see supplementary materials file) and **2** provide an overall visualization concerning the matters.


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 ...
100%
75%
50%
25%
0%
Success Rate or Accuracy
TAMP Success
Language-Table Success
SayCan Affordances
TAMP Data Only
Lang. Table Data Only
SayCan Data Only
Full Mixture (All robots + WebLI, VQA, COCO, etc.)
PaLM-E Training Data
Different models for different robots, trained from scratch
One model for all robots with ViT + PaLM pre-training

PaLM-E model architecture, showing how PaLM-E ingests different modalities (states and/or images) and addresses tasks through multimodal language modeling

Positive transfer of knowledge from general vision-language tasks results in more effective robot learning, shown for three different robot embodiments and domains

Figure 2. PaLM-E an embodied multimodal language model in action 2.

2.3. A deep dive into general artificial intelligence (GAI)

Generative artificial intelligence (AI) represents a significant advancement in the field of artificial intelligence, with algorithms like ChatGPT at the forefront, capable of producing a wide array of content types, including text, code, images, audio, simulations, and videos.

These transformative capabilities have garnered widespread attention, particularly following the release of ChatGPT by OpenAI in November 2022. As a highly capable chatbot, ChatGPT quickly rose to prominence, attracting over a million users within just five days of its launch. Its ability to generate diverse forms of content, from computer code and essays to creative works like poems, highlights its versatility and potential to revolutionize content creation across various industries [23–33]. The impact of Generative AI, exemplified by ChatGPT, extends beyond mere content generation, raising both opportunities and challenges. Tools like DALL-E, another AI system developed by OpenAI for generating art, further illustrate the potential of Generative AI to disrupt traditional workflows and job markets [32–34].

However, this disruption is accompanied by uncertainties and risks, particularly regarding the implications for employment, content quality, and ethical considerations. As Generative AI continues to evolve, it becomes increasingly important to balance its transformative potential with a careful understanding of its limitations and the risks it may pose. To fully grasp the significance of Generative AI, it is essential to distinguish it from broader concepts within artificial intelligence and machine learning.

While AI broadly refers to machines that mimic human intelligence, machine learning is a subset of AI focused on models that learn from data patterns without explicit programming. Generative AI, as a breakthrough in machine learning, enables models to create new content on demand, going beyond traditional tasks of pattern recognition and classification.

Text-based models like ChatGPT rely on self-supervised learning, where they are trained on vast amounts of text data to generate predictions and responses that closely resemble human language. Although ChatGPT has captured public attention, it is part of a lineage of text-based models that includes predecessors like GPT-3 and BERT, which have also made significant contributions to the field [34–36].

The development of Generative AI models requires substantial resources, typically available only to well-funded tech companies. Training these models involves large-scale data processing and significant computational power. For instance, GPT-3, one of the models underlying ChatGPT, was trained on approximately 45 terabytes of text data, reflecting the massive investment in both infrastructure and talent necessary to achieve high levels of performance [32–34].

These investments enable companies like OpenAI, DeepMind, and Meta to push the boundaries of what Generative AI can achieve, creating models capable of producing outputs that rival human-generated content in terms of quality and diversity.

Generative AI models are particularly adept at generating content that is not only lifelike but also creative, introducing random elements that add diversity to the outputs. This capability has broad practical applications across industries such as IT, software development, marketing, and healthcare [34–49]. For example, Generative AI can rapidly produce written content, optimize code, or create marketing copy, saving time and resources for organizations. However, the effectiveness of these outputs depends largely on the quality and relevance of the training data used to develop the models, as well as the specific use cases for which they are applied.

Despite the impressive capabilities of Generative AI, there are significant limitations and risks that must be acknowledged. While the outputs of Generative AI models can be convincing, they are not immune to errors, biases, or inappropriate content generation. These risks can lead to reputational and legal challenges, particularly if biased or offensive content is inadvertently published [34–36]. To mitigate these risks, it is crucial to carefully curate training data, consider the use of smaller, more specialized models, and maintain human oversight to review and approve AI-generated content before it is disseminated. As a rapidly evolving field, the long-term effects and risks associated with Generative AI are still being understood. Organizations adopting these technologies must remain vigilant, staying informed about regulatory developments and emerging risks. While Generative AI holds immense promise, its responsible implementation, coupled with continuous monitoring, is essential to maximize its benefits while minimizing unintended consequences. By approaching Generative AI with a balanced perspective, we can harness its transformative potential while addressing the ethical and practical challenges it presents [34–49].

3. GAI: From a techspertive point of view

Generative AI, often synonymous with large language models (LLMs), represents a significant subset of machine learning, recognized for its ability to generate natural-sounding language. The emergence of tools like Bard, an experimental platform designed for collaboration with Generative AI powered by a large language model, underscores the growing influence of this technology. To fully understand Generative AI, it is essential to first explore the broader context of artificial intelligence (AI). Most modern AI is rooted in machine learning, a process where neural networks—complex computer systems—learn from vast amounts of data. These networks are trained to perform specific tasks, such as image classification or predicting the next word in a sentence, by identifying patterns within the data [34–36]. Language models, a particular type of neural network, are central to Generative AI. They are trained on extensive datasets and are capable of predicting the next word in a sequence, becoming increasingly sophisticated as the amount of training data grows. These models are already in use in everyday applications, such as Gmail’s Smart Compose and Smart Reply features, which assist users by suggesting contextually relevant responses.

Bard, powered by such models, leverages this predictive capability to generate coherent and contextually appropriate language. Generative AI, by its nature, goes beyond simply predicting text. It is designed to create entirely new content based on the patterns and structures it has learned from its training data. This ability to generate novel combinations of text in natural-sounding language is what distinguishes Generative AI as a powerful tool in content creation. However, its capabilities extend beyond text; Generative AI can also produce images, audio, and even video, offering vast potential across multiple creative domains. The potential impact of Generative AI on creative fields is profound. It has the ability to transform the way we approach creativity, much like the drum machine did for music production.

By automating repetitive tasks and eliminating drudgery, Generative AI can enhance creative workflows, allowing human creators to focus on more innovative and imaginative aspects of their work. However, it is crucial to recognize that Generative AI is not a replacement for human creativity but rather a tool that augments and facilitates it [34–49]. Despite its potential, Generative AI also presents challenges, particularly in educational contexts. The ease with which AI-generated content can be produced raises important questions about how we measure success and originality in education. These concerns highlight the need for a thoughtful and responsible approach to the development and deployment of machine learning technologies.

Companies like Google have taken steps to address these challenges by establishing AI principles and creating internal governance structures aimed at ensuring the ethical development of AI. These guidelines are designed to prevent harm and mitigate issues related to bias and toxicity in AI-generated content. By adhering to these principles, organizations can help ensure that Generative AI technologies are developed and used in ways that benefit society while minimizing potential risks. While Generative AI holds immense potential to revolutionize creative processes and workflows, it also demands careful consideration of its societal impacts. By fostering responsible development and use, we can harness the power of Generative AI to tackle new challenges and open up fresh perspectives in various fields, ultimately contributing to a more innovative and creative future.

4. Machine learning (ML) mystery: A case study investigation analysis

A recent study, conducted by researchers from MIT in collaboration with Google Research and Stanford University, delves into the intriguing phenomenon of in-context learning observed in large language models, such as OpenAI’s GPT-3 and GPT-4. In-context learning refers to the ability of these models to perform new tasks after being exposed to just a few examples, without the need for retraining on new data. This capability has piqued the interest of researchers, who have sought to understand the underlying mechanisms that enable such models to learn without the traditional process of parameter updates [34–36]. The study centers on the hypothesis that these massive neural network models, particularly transformers like GPT-3, might encapsulate smaller, simpler linear models within their vast architecture. These smaller models, the researchers suggest, could be trained to execute new tasks using straightforward learning algorithms, all while leaving the parameters of the overarching model unchanged. This notion challenges the conventional understanding of how learning occurs within large language models, opening the door to new theories about their inner workings. To explore this hypothesis, the researchers conducted a theoretical investigation into transformer models specifically engineered for in-context learning. Transformers, which form the backbone of models like GPT-3, are neural networks known for their ability to process sequences of data, such as text, in a highly efficient manner. The findings of the study reveal that within these transformers, a linear model can be “written” into the hidden states of the network.

This process involves embedding the linear model within the earliest layers of the transformer, allowing the larger model to simulate and train this smaller model using pre-existing information.

As a result, the model can effectively perform in-context learning, adapting to new tasks without the need for parameter modification. The implications of these results are profound. By demonstrating that a transformer can house and train a linear model within its hidden states, the study suggests a novel method by which large language models might achieve in-context learning. This insight could lead to the development of more efficient learning algorithms capable of performing new tasks without the extensive retraining that is typically required.

The lead authors of the study emphasize the practical advantages of in-context learning, particularly its potential to streamline the learning process. By eliminating the need for complex engineering and the collection of domain-specific data, in-context learning presents a more efficient approach to training models. The researchers propose that these in-context learners do not merely mimic patterns observed in their training data; instead, they might actually acquire the ability to perform new tasks.

This challenges the prevailing notion that large language models simply memorize tasks, suggesting that they possess a more sophisticated capacity for learning. Ultimately, this study represents a significant step toward unraveling the capabilities of modern large language models. By shedding light on the mechanisms behind in-context learning, the research contributes to a deeper understanding of how these models can be leveraged for complex learning tasks. As the field of machine learning continues to evolve, insights such as these will be crucial in shaping the future of AI development.

5. GAI: The creative work perspectives

Generative AI, particularly through the use of large language and image models, is revolutionizing the landscape of creative work and business functions. These models, often referred to as foundation models, present a wide array of opportunities, including the automation of content generation, enhancement of content quality, diversification of content types, and the ability to personalize outputs across various domains. Models like OpenAI’s GPT-3 and GPT-4 are prime examples of Generative AI’s capacity to produce diverse types of content, such as text, images, and videos. These models are trained on vast datasets and require significant computational resources to develop. However, once trained, they can be fine-tuned for specific content domains using relatively smaller datasets. This process underscores the continued necessity for human involvement, both in generating prompts for the AI and in evaluating or editing the content produced by these models.

One of the key applications of Generative AI lies in the marketing sector. For instance, specialized versions of GPT-3, such as Jasper, are being utilized to create blogs, social media posts, and other customer-facing content. These tools are invaluable for maximizing search engine optimization (SEO) and tailoring personalized pitches in public relations. Additionally, image generation tools like DALL-E 2 are already making an impact in advertising, with brands such as Heinz and Nestle adopting these technologies to enhance their campaigns.

Generative AI also shows significant potential in code generation. GPT-3’s Codex, for example, can create code snippets based on textual descriptions, dramatically improving the efficiency of software development. Experiments by companies like Deloitte have demonstrated up to a 20% increase in code development speed using Codex, highlighting the practical benefits of these models in the tech industry [34–36]. Furthermore, Generative AI is increasingly being integrated into conversational AI and chatbots, leading to more sophisticated conversation understanding and context awareness. However, challenges persist, particularly regarding the replication of biased language. These issues necessitate ongoing efforts to refine and filter AI outputs, especially in sensitive contexts. Another area of interest is knowledge management. Large language models, when fine-tuned on specific content, have the potential to manage and streamline an organization's knowledge base. For example, Morgan Stanley is working with OpenAI's GPT-3 and GPT-4 to fine-tune models for wealth management training, aiming to leverage Generative AI for more effective information dissemination within the company. However, the ethical and legal implications of Generative AI cannot be overlooked. The rise of deepfakes and concerns over content ownership are central to discussions about the future of AI in creative work. As AI systems become more capable of generating a wide range of content—including emails, articles, computer programs, and more—the questions of intellectual property and content ownership become increasingly complex and pressing. The perspective presented acknowledges that while Generative AI models offer unprecedented opportunities, they also bring about challenges and risks that must be carefully managed.

As these technologies continue to evolve, they are likely to create unforeseen opportunities and implications for creative work and knowledge management. The ongoing discourse on these issues will play a critical role in shaping the future of Generative AI and its integration into various aspects of work and life.

6. GAI: Ethics, accountability, trust, public interest and proactive managements

The rise of Generative AI marks a pivotal moment in the evolution of artificial intelligence, offering transformative potential across various sectors. Unlike traditional AI, Generative AI, driven by large language models, responds to user prompts with outputs that closely mimic human language, making this technology accessible to a broader audience. However, as Generative AI gains prominence, it also brings forth significant ethical, accountability, and trust-related challenges that require careful consideration [36]. In the business world, there is a growing interest in harnessing Generative AI to enhance enterprise operations. Yet, as this technology rapidly evolves, the importance of trust and ethical management becomes paramount. One of the most pressing concerns is whether business users can trust the outputs generated by these AI models. The potential risks associated with Generative AI, such as producing inaccurate or hallucinated outputs, pose serious challenges for end-users, who may struggle to assess the factual accuracy of content that appears convincingly eloquent.

Generative AI models are also prone to biases as they are trained on vast datasets that may contain inherent prejudices. This raises the risk of users placing undue confidence in biased or erroneous outputs, which could have significant consequences in decision-making processes. Moreover, the issue of attribution is critical. Since Generative AI outputs are closely aligned with the original training data, there is a heightened risk of plagiarism and copyright violations. Balancing trust in these outputs with human oversight becomes a complex challenge, particularly when considering the legal and brand implications for enterprises. Transparency and explaining ability are equally crucial. End-users, who may not have a deep technical understanding of AI, need accessible explanations of how Generative AI works. This underscores the necessity for enterprise-wide AI literacy and risk awareness, which can help mitigate potential negative consequences. Without transparency, users may misinterpret AI-generated content, leading to misguided decisions that could have far-reaching effects.

Accountability in the use of Generative AI is a central theme in this discussion. As these models increasingly mimic human creativity, the responsibility for their outputs must be clearly defined. It is essential to maintain human oversight and ensure that AI-generated content is subject to thorough analysis, scrutiny, and context-aware review. This human element is critical to preserving ethical standards and ensuring that AI-driven decisions align with societal values and public interest.

The need for proactive management in the deployment of Generative AI cannot be overstated. Organizations must establish robust frameworks for accountability, trust, and ethics, ensuring that the outcomes of Generative AI are transparently linked to their creators and the enterprise. As the AI landscape continues to evolve, these frameworks will be essential in navigating the complexities of this technology and safeguarding its integration into various aspects of work and life.

While Generative AI holds immense potential to revolutionize content creation and business operations, it also necessitates a careful balance between innovation and ethical responsibility. By addressing the challenges of trust, accountability, and public interest and by implementing proactive management strategies, organizations can harness the power of Generative AI while minimizing the risks and ensuring its positive impact on society.

7. The future of Generative AI (GAI) and its directions

The current excitement surrounding Generative AI, particularly models like ChatGPT, has sparked widespread interest and speculation about its potential impact. However, it's crucial to focus on the true value of Generative AI, which lies not in its capacity as a generalized solution but in its application to niche domains where it can offer significant, context-specific advantages.

While the buzz around ChatGPT is undeniable, the primary value of Generative AI will emerge in specialized contexts where it can explore and utilize highly specific information in novel ways. The development of ChatGPT plugins by various companies exemplifies this shift. These plugins are not about creating a one-size-fits-all solution but rather enhancing functionality in specific areas.

For instance, in the realm of travel planning, a Generative AI tool tailored for a company like Expedia can deliver a substantial competitive edge, particularly in a market where information discovery is critical.

This trend raises important questions about the future of Generative AI and its implications for established search giants like Google. Will these developments pose a serious threat to their dominance, or are we witnessing an "iPhone moment"—a transformative shift in user behavior and expectations? The likely outcome is a gradual change, where organizations leverage large language models (LLMs) trained on their own data to drive meaningful transformations in how they operate and interact with customers.

OpenAI's recognition of the commercial potential of Generative AI is evident in its recent move to open a waiting list for companies to access ChatGPT plugins. This decision foreshadows the emergence of numerous new products and interfaces powered by OpenAI's Generative AI systems in the coming months and years, underscoring the expanding role of Generative AI in various sectors. However, it's important to dispel the notion that OpenAI is the sole gatekeeper of Generative AI technology. While ChatGPT is a prominent tool, it is not the only one available. A broader ecosystem of tools exists, including self-hosted LLMs, which offer organizations an alternative approach to Generative AI. By deploying LLMs on their own enterprise data, organizations can address privacy concerns and maintain greater control over their AI implementations.

The future of Generative AI is also likely to see a trend toward domain-specific language models. Fine-tuning general-purpose LLMs on specific datasets could result in highly effective information retrieval tools tailored to particular industries or use cases [36]. This approach has promising applications in areas such as product information management, content creation, and internal documentation, demonstrating how Generative AI can evolve into more specialized and practical tools.

As Generative AI becomes more embedded in specific contexts, its mystique as an all-knowing entity will diminish. The future of AI will likely be less threatening and more approachable, especially as it becomes increasingly domain-specific. A comparison can be drawn to GitHub Copilot, an AI tool that supports software developers by helping them solve problems within the scope of their existing knowledge and experience. This is a key indicator of how Generative AI will be successful—not as a catch-all solution but as an integrated tool that enhances specific applications. Ultimately, the true value of Generative AI will be realized as it becomes seamlessly embedded in particular domains, leading to a more practical and grounded acceptance of its capabilities [34–49]. As users come to understand the limitations of Generative AI, its usefulness will become more apparent, fostering a balanced perspective that recognizes both its potential and its boundaries.

8. Results and findings

Google's PaLM-E, an embodied multimodal language model, demonstrates significant advancements in robotic environments and vision-language tasks, showcasing its versatility across a range of applications.

The model was rigorously evaluated in three distinct robotic scenarios, each involving real robots and a variety of tasks such as visual question answering (VQA), image captioning, and general language processing. The findings highlight PaLM-E's ability to effectively integrate language understanding with robotic control, establishing new benchmarks in the field.

Robotic scenarios and task performance:

In one scenario, PaLM-E was tasked with directing a mobile robot in a kitchen environment. The model successfully guided the robot to retrieve a bag of chips, demonstrating robustness even when the bag was placed back into a drawer, which introduced additional complexity. This capability underscores PaLM-E's resilience to environmental disturbances and its ability to execute tasks in dynamic settings. In another task, PaLM-E was instructed to grab an unseen green block. Here, the model generated a plan that extended beyond the robot's training data, effectively generalizing its actions to handle novel objects.

This ability to generalize is a critical advancement, enabling robots to perform tasks with objects or in environments not explicitly encountered during training. In a tabletop robot environment, PaLM-E tackled long-horizon tasks, such as sorting blocks by color into designated corners. The model demonstrated its capacity to process visual information and generate sequences of textually represented actions for intricate, prolonged tasks. This performance marks a significant improvement over previous models, showcasing PaLM-E's potential for handling complex, multi-step processes.

Moreover, PaLM-E exhibited impressive zero-shot generalization. For example, it successfully pushed red blocks towards a coffee cup, adapting to new tasks that were not part of its training data. This adaptability is a testament to the model's robustness in handling unforeseen challenges and tasks. In a third robotic scenario inspired by task and motion planning (TAMP), PaLM-E addressed combinatorically challenging planning tasks. The model effectively solved these tasks by leveraging knowledge transferred from visual and language models, coupled with a modest amount of training data from an expert TAMP planner. This capability highlights PaLM-E's efficiency in learning from limited data while achieving high-level task performance.

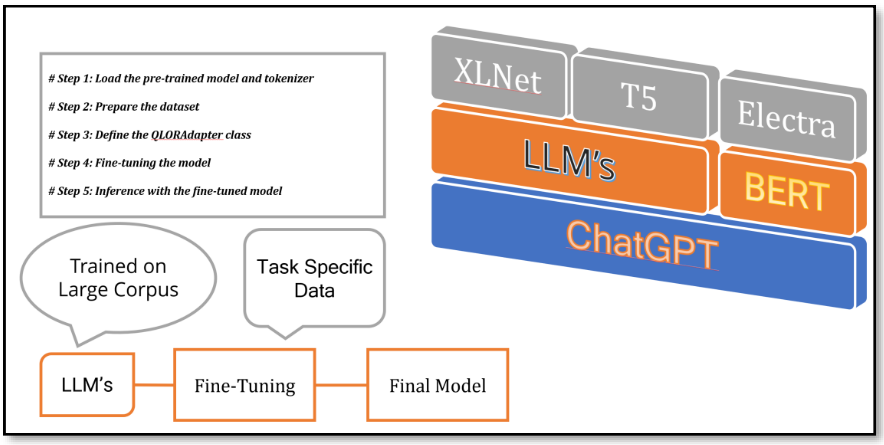
Visual-Language Generalization:

PaLM-E also proved itself as a visual-language generalist, outperforming leading vision-language-only models in several benchmarks. Notably, it achieved the highest reported score on the OK-VQA dataset, a challenging benchmark for visual question answering, without requiring task-specific fine-tuning. This performance underscores the model's strength in visual understanding and its extensive external world knowledge.

The largest version of the model, PaLM-E-562B, exhibited particularly advanced capabilities, including visual chain-of-thought reasoning and multi-image inference. These abilities enable the model to break down complex answering processes into smaller, manageable steps and to perform inference across multiple images, even though it was primarily trained on single-image prompts.

This sophistication in reasoning and inference represents a significant leap forward in the capabilities of embodied AI systems.

Figure S2 (see supplementary materials file) and **Figures 3–5** in the accompanying visualizations provide further insights into these findings, illustrating the model's performance across various tasks and environments and highlighting its versatility and robustness.



Step 1: Load the pre-trained model and tokenizer

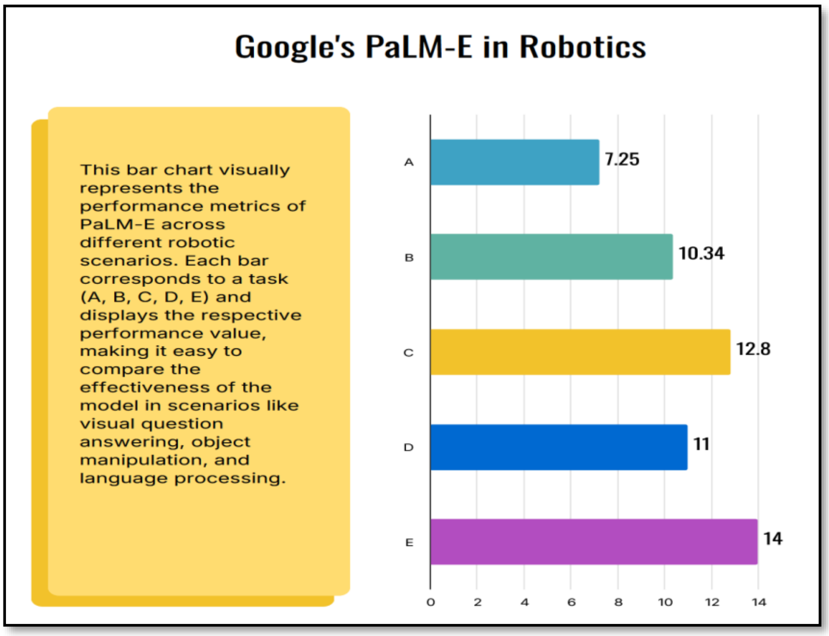
Step 2: Prepare the dataset

Step 3: Define the QLORAdapter class

Step 4: Fine-tuning the model

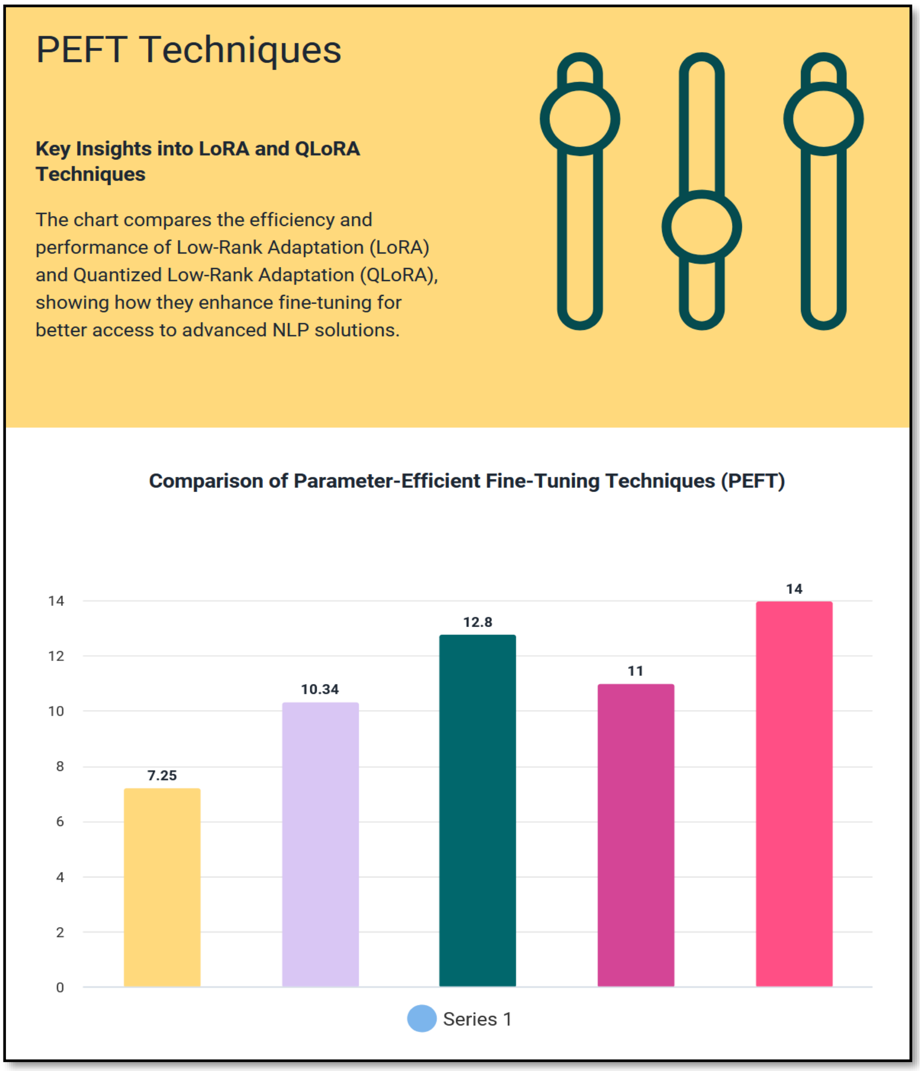
Step 5: Inference with the fine-tuned model

Figure 3. The experimental simulation processing.



This bar chart visually represents the performance metrics of PaLM-E across different robotic scenarios. Each bar corresponds to a task (A, B, C, D, E) and displays the respective performance value, making it easy to compare the effectiveness of the model in scenarios like visual question answering, object manipulation, and language processing.

Figure 4. A visualization of the research findings 1.



PEFT Techniques

Key Insights into LoRA and QLoRA Techniques

The chart compares the efficiency and performance of Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), showing how they enhance fine-tuning for better access to advanced NLP solutions.

Figure 5. A visualization of the research findings 2.

9. Discussions and future directions

The exploration of Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), brings to light several key insights and future possibilities in the field of Natural Language Processing (NLP). This discussion addresses essential queries about these techniques, providing a comprehensive overview of their goals, advantages, and potential impacts on the research community [32–49].

Parameter-Efficient Fine-Tuning Goals:

The primary goal of parameter-efficient fine-tuning is to adapt pre-trained language models to specific tasks while minimizing the computational and memory burdens traditionally associated with fine-tuning large models. This approach is particularly crucial as the scale and complexity of language models continue to grow. By reducing the number of parameters that need to be updated during the fine-tuning process, PEFT techniques enable more efficient adaptation to various downstream tasks, making it feasible to deploy sophisticated models in resource-constrained environments.

Enhancements through Quantized Low-Rank Adaptation (QLoRA):

One of the significant advancements in parameter efficiency comes from the integration of quantization into the low-rank adaptation process, as seen in QLoRA. By quantizing the weights of the adaptation layers, QLoRA reduces the memory footprint and computational load without resorting to complex quantization techniques. This method preserves the overall performance of the language model, ensuring that the efficiency gains do not come at the cost of reduced accuracy or effectiveness. The approach enhances the applicability of fine-tuning large models in practical scenarios, where hardware constraints often pose significant challenges.

Advantages of Low-Rank Adaptation (LoRA):

Low-Rank Adaptation offers several advantages that make it a valuable technique for fine-tuning large language models. First, LoRA reduces the parameter overhead, which is particularly beneficial when multiple tasks require fine-tuning, as it allows for more efficient task switching without needing to retrain the entire model from scratch. Additionally, LoRA maintains inference latency, ensuring that the model’s responsiveness remains intact despite the reduced number of trainable parameters. These benefits make LoRA an effective solution for deploying adaptable models in real-world applications, where both efficiency and performance are paramount.

Implications for Researchers:

Researchers stand to gain significantly from the adoption of PEFT techniques. By leveraging methods like LoRA and QLoRA, researchers can fine-tune large language models more efficiently, optimizing their use across a range of downstream tasks without incurring excessive computational costs. This not only broadens the accessibility of powerful language models but also encourages innovation by allowing more researchers to experiment with and refine these models. The practical implications of these techniques are far-reaching, particularly in enabling the deployment of advanced NLP solutions in a wider array of contexts.

Applicability of QLoRA to Language Models:

QLoRA’s versatility is another point of discussion, as it can be applied to various types of language models, including RoBERTa, DeBERTa, GPT-2, and GPT-3. This adaptability underscores the potential of QLoRA as a standard tool for parameter-efficient fine-tuning across different architectures. The ability to fine-tune diverse models with minimal resource requirements opens up new possibilities for deploying customized NLP solutions in specific domains, further enhancing the impact of these models on industry and academia.

Future Directions:

Looking ahead, the development of more advanced PEFT techniques will likely focus on further reducing the computational demands of fine-tuning while expanding the applicability to even larger and more complex models. Future research may explore the integration of other efficiency-boosting methods, such as pruning or distillation, with PEFT techniques to create even more streamlined fine-tuning processes. Additionally, there is potential for extending the benefits of PEFT beyond NLP, applying similar principles to other domains where large models are used, such as computer vision or speech processing.

The continued evolution of these techniques will also necessitate addressing the challenges of maintaining model performance while optimizing for efficiency. As language models become increasingly embedded in real-world applications, ensuring that they remain accurate, reliable, and fair will be critical. This may involve developing more sophisticated methods for managing the trade-offs between efficiency and performance, as well as refining the mechanisms for controlling model bias and ensuring robust generalization across different tasks and datasets.

Parameter-efficient finetuning represents a promising direction in the ongoing development of NLP technologies. By enabling more efficient use of large language models, PEFT techniques like LoRA and QLoRA have the potential to significantly expand the accessibility and applicability of these models, driving innovation and enabling the deployment of advanced AI solutions across a broader range of contexts.

10. Conclusions

The development of PaLM-E marks a significant milestone in advancing the capabilities of generally-capable models by simultaneously addressing vision, language, and robotics. This research not only explores the model's versatility in unifying traditionally distinct tasks but also highlights the broader implications of PaLM-E in enhancing the integration of these domains. By leveraging knowledge transfer from vision and language to robotics, PaLM-E demonstrates the potential to create more proficient robots capable of utilizing diverse data sources. This advancement paves the way for broader applications in multimodal learning, positioning PaLM-E as a critical facilitator in the evolution of artificial intelligence across various fields.

In parallel, the rapid evolution of Parameter-Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), addresses the significant challenges posed by the computational and memory requirements of fine-tuning large language models (LLMs). These innovative strategies enhance the efficiency of the fine-tuning process while maintaining or even improving task performance. The introduction of trainable low-rank matrices in LoRA and quantization techniques in QLoRA exemplifies novel approaches to minimizing the number of trainable parameters, making the fine-tuning of LLMs more practical and accessible.

The emphasis on parameter efficiency is crucial for overcoming the resource-intensive nature of LLMs, contributing to reduced memory usage and computational costs. This focus not only addresses technical challenges but also has profound implications for the broader field of Natural Language Processing (NLP). By making the fine-tuning process more efficient, these techniques open up opportunities for deploying large language models in a wider range of real-world applications, fostering innovation and broader adoption of NLP technologies.

As these PEFT techniques continue to evolve, they are reshaping the landscape of fine-tuning processes for LLMs, making them more adaptable, resource-efficient, and applicable to a diverse array of tasks. The potential to deploy large language models with reduced resource requirements is transformative, enabling the practical application of advanced AI across various domains.

The research underscores the significant impact of PaLM-E and parameter-efficient fine-tuning techniques like LoRA and QLoRA, highlighting their role in driving the next generation of AI technologies and expanding the accessibility and applicability of NLP in real-world scenarios.

Supplementary 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 exploration findings and information of the research investigations are referenced where appropriate.

Acknowledgments: The author would like to acknowledge and thank the GOOGLE Deep Mind Research with its associated pre-prints access platforms. This research exploration was investigated under the platform provided by GOOGLE Deep Mind which is under the support of the GOOGLE Research and the GOOGLE Research Publications within the GOOGLE Gemini platform. Using their provided platform of datasets and database associated files with digital software layouts consisting of free web access to a large collection of recorded models that are found within research access and its related open-source software distributions which is the implementation for the proposed research exploration that was undergone and set in motion. There are many data sources some of which are resourced and retrieved from a wide variety of GOOGLE service domains as well. All the data sources which have been included and retrieved for this research are identified, mentioned and referenced where appropriate.

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

Conflict of interest: The author declares no conflict of interest.

References

1. McKinsey. The state of AI in 2022-and a half decade in review. Available online: https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai%20in%202022%20and%20a%20half%20decade%20in%20review/the-state-of-ai-in-2022-and-a-half-decade-in-review.pdf (accessed on 2 March 2024).
2. McKinsey & Company. McKinsey Technology Trends Outlook 2022. Available online: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20top%20trends%20in%20tech%202022/mckinsey-tech-trends-outlook-2022-full-report.pdf (accessed on 2 March 2024).
3. Quick guide to AI 2.0 Oct 2020. Mckinsey.com. Available online: https://ceros.mckinsey.com/quick-guide-to-ai-12/p/1 (accessed on 2 March 2024).
4. Chui, M., Manyika, J., & Miremadi, M. What AI can and can’t do (yet) for your business. Available online: https://www.mckinsey.com/capabilities/quantumblack/our-insights/what-ai-can-and-cant-do-yet-for-your-business (accessed on 2 March 2024).
5. Hedges R. Artificial intelligence discovery & admissibility case law and other resources. Available online: https://afbnj.org/wp-content/uploads/2023/12/AI-Written-Materials-1.25.24.pdf (accessed on 2 March 2024).
6. Griffith E, Metz C. Anthropic, an A.I. Start-Up, Is Said to Be Close to Adding $300 Million. Available online: https://www.nytimes.com/2023/01/27/technology/anthropic-ai-funding.html (accessed on 2 March 2024).
7. Fiegerman S, Lanxon N. AI Glossary: A-Z of Artificial Intelligence Terms to Know.. Available online: https://www.bloomberg.com/features/2024-artificial-intelligence-glossary/ (accessed on 2 March 2024).

8. Pinaya WHL, Graham MS, Kerfoot E, et al. Generative AI for Medical Imaging: extending the MONAI Framework. ArXiv. 2013. doi: 10.48550/arXiv.2307.15208
9. Pasick A. Artificial Intelligence Glossary: Neural Networks and Other Terms Explained. Available online: https://www.nytimes.com/article/ai-artificial-intelligence-glossary.html (accessed on 2 March 2024).
10. Generative models. Available online: https://openai.com/index/generative-models/ (accessed on 2 March 2024).
11. Metz C. OpenAI Plans to Up the Ante in Tech's A.I. Race. Available online: https://www.nytimes.com/2023/03/14/technology/openai-gpt4-chatgpt.html (accessed on 2 March 2024).
12. Thoppilan R, De Freitas D, Hall J, et al. (2022). LaMDA: Language Models for Dialog Applications. ArXiv. 2022; ArXiv:2201.08239.
13. Roose K. A Coming-Out Party for Generative A.I., Silicon Valley's New Craze. Available online: https://www.nytimes.com/2022/10/21/technology/generative-ai.html (accessed on 2 March 2024).
14. Don't fear an AI-induced jobs apocalypse just yet. Available online: https://www.economist.com/business/2023/03/06/dont-fear-an-ai-induced-jobs-apocalypse-just-yet (accessed on 2 March 2024).
15. Harreis H, Koullias T, Roberts R, Te K. Generative AI in Fashion | McKinsey. Available online: https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion (accessed on 2 March 2024).
16. Eapen TT, Finkenstadt DJ, Folk J, Venkataswamy L. How Generative AI Can Augment Human Creativity. Harvard Business Review. 2023.
17. The race of the AI labs heats up. Available online: https://www.economist.com/business/2023/01/30/the-race-of-the-ai-labs-heats-up (accessed on 2 March 2024).
18. Google Cloud brings generative AI to developers, businesses, and governments. Available online: https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-for-businesses-and-governments (accessed on 2 March 2024).
19. Political Machines: Understanding the Role of AI in the U.S. 2024 Elections and Beyond - Center for Media Engagement - Center for Media Engagement. Available online: https://mediaengagement.org/research/generative-ai-elections-and-beyond/ (accessed on 2 March 2024).
20. Simon FM, Altay S, Mercier H. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Available online: https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/ (accessed on 22 October 2024).
21. Anyoha R. The History of Artificial Intelligence. Science in the News. Scientific Research Publishing; 2017.
22. Benbya H, Strich F, Tamm T. Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work. Journal of the Association for Information Systems. 2024; 25(1): 23–36. doi: 10.17705/1jais.00861
23. Gagniuc PA. Markov Chains. John Wiley & Sons, Inc.; 2017. doi: 10.1002/9781119387596
24. Jebara T. Machine Learning. Springer eBooks; 2004. doi: 10.1007/978-1-4419-9011-2
25. Cao Y, Li S, Liu Y, et al. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. ArXiv. 2023; ArXiv:2303.04226.
26. openai. Available online: https://github.com/openai/finetune-transformer-lm (accessed on 2 March 2024).
27. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Blog. 2019; 1(8): 9.
28. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv. 2023; ArXiv:2303.12712.
29. Schlagwein D, Willcocks L. 'ChatGPT et al.': The ethics of using (generative) artificial intelligence in research and science. Journal of Information Technology. 2023; 38(3): 232-238. doi: 10.1177/02683962231200411
30. Islam A. A History of Generative AI: From GAN to GPT-4. Available online: https://www.marktechpost.com/2023/03/21/a-history-of-generative-ai-from-gan-to-gpt-4/ (accessed on 2 March 2024).
31. McKinsey & Company. What Is Generative AI? Available online: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai (accessed on 2 March 2024).
32. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. ArXiv. 2021; ArXiv:2108.07258.
33. Chen M, Tworek J, Jun H, et al. Evaluating Large Language Models Trained on Code. ArXiv. 2021; ArXiv:2107.03374
34. Bin Akhtar Z. From bard to Gemini: An investigative exploration journey through Google's evolution in conversational AI

and generative AI. Computing and Artificial Intelligence. 2024; 2(1): 1378. doi: 10.59400/cai.v2i1.1378

35. Bin Akhtar Z. Exploring Biomedical Engineering (BME): Advances within Accelerated Computing and Regenerative Medicine for a Computational and Medical Science Perspective Exploration Analysis. Journal of Emergency Medicine: Open Access. 2024; 2(1): 1-23. doi: 10.33140/jemoa.02.01.06
36. Akhtar ZB. 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. 2024; 11(1). doi: 10.1186/s43067-024-00145-1
37. Unraveling the Promise of Computing DNA Data Storage: An Investigative Analysis of Advancements, Challenges, Future Directions. Journal of Advances in Artificial Intelligence. 2024; 2(1). doi: 10.18178/jaai.2024.2.1.122-137
38. Akhtar ZB. The design approach of an artificial intelligent (AI) medical system based on electronical health records (EHR) and priority segmentations. The Journal of Engineering. 2024; 2024(4). doi: 10.1049/tje2.12381
39. Akhtar ZB. Securing Operating Systems (OS): A Comprehensive Approach to Security with Best Practices and Techniques. International Journal of Advanced Network, Monitoring and Controls. 2024; 9(1): 100-111. doi: 10.2478/ijanmc-2024-0010
40. Akhtar ZB, Gupta AD. Integrative Approaches for Advancing Organoid Engineering: From Mechanobiology to Personalized Therapeutics. Journal of Applied Artificial Intelligence. 2024; 5(1): 1-27. doi: 10.48185/jaai.v5i1.974
41. Akhtar ZB. Advancements within Molecular Engineering for Regenerative Medicine and Biomedical Applications an Investigation Analysis towards A Computing Retrospective. Journal of Electronics, Electromedical Engineering, and Medical Informatics. 2024; 6(1). doi: 10.35882/jeeemi.v6i1.351
42. Akhtar ZB. Accelerated Computing a Biomedical Engineering and Medical Science Perspective. Annals of the Academy of Romanian Scientists Series on Biological Sciences. 2023; 12(2): 138-164. doi: 10.56082/annalsarscibio.2023.2.138
43. Akhtar ZB. Designing an AI Healthcare System: EHR and Priority-Based Medical Segmentation Approach. Medika Teknika: Jurnal Teknik Elektromedik Indonesia. 2023; 5(1): 50-66. doi: 10.18196/mt.v5i1.19399
44. Bin Akhtar Z. A Revolutionary Gaming Style in Motion. In: Dey I (editor). Computer-Mediated Communication. IntechOpen; 2022. doi: 10.5772/intechopen.100551
45. Akhtar ZB, Stany Rozario V. The Design Approach of an Artificial Human Brain in Digitized Formulation based on Machine Learning and Neural Mapping. In: Proceedings of the 2020 International Conference for Emerging Technology (INCET); 5-7 June 2020; Belgaum, India. doi: 10.1109/incet49848.2020.9154000
46. Akhtar Z. Biomedical engineering (bme) and medical health science: an investigation perspective exploration. Quantum Journal of Medical and Health Sciences. 2024; 3(3): 1-24.
47. Akhtar ZB, Rawol AT. Uncovering Cybersecurity Vulnerabilities: A Kali Linux Investigative Exploration Perspective. International Journal of Advanced Network, Monitoring and Controls. 2024; 9(2): 11-22. doi: 10.2478/ijanmc-2024-0012
48. Akhtar ZB, Tajbiul Rawol A. Unlocking the Future for the New Data Paradigm of DNA Data Storage: An Investigative Analysis of Advancements, Challenges, Future Directions. Journal of Information Sciences. 2024. doi: 10.34874/IMIST.PRSM/JIS-V23I1.47102
49. Bin Akhtar Z. Artificial intelligence (AI) within manufacturing: An investigative exploration for opportunities, challenges, future directions. Metaverse. 2024; 5(2): 2731. doi: 10.54517/m.v5i2.2731