Preprints.org

Article

Not peer-reviewed version

Optimizing AI Language Models: A Study of ChatGPT-4 vs. ChatGPT-4o

Md Nurul Absar Siddiky *, Muhammad Enayetur Rahman *, MD Fayaz Bin Hossen *, Muhammad Rezaur Rahman *, Md. Shahadat Jaman

Posted Date: 3 February 2025

doi: 10.20944/preprints202502.0066.v1

Keywords: ChatGPT-4; ChatGPT-4o; Natural Language Processing; Optimization; Artificial Intelligence; Deep Learning; Transformer Architecture

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Optimizing AI Language Models: A Study of ChatGPT-4 vs. ChatGPT-4o

Md Nurul Absar Siddiky [1,*], Muhammad Enayetur Rahman [2,*], MD Fayaz Bin Hossen [3,*], Muhammad Rezaur Rahman [4,*] and Md. Shahadat Jaman [5]

[1] Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC, 28223, USA
[2] Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, 23509, USA
[3] Computer Science, Old Dominion University, Norfolk, VA, 23509, USA
[4] Research and Innovation Department, Agile Crafts, Khilgaon, Dhaka, 1219, Bangladesh
[5] Electrical Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh; shahadat30eee@gmail.com
* Correspondence: msiddiky@uncc.edu (M.N.A.S.); mrahm011@odu.edu (M.E.R.); mhoss006@odu.edu (M.F.B.H.); refayetbd@gmail.com (M.R.R.)

Abstract: This paper presents a comparative analysis of OpenAI's GPT-4 and its optimized variant, GPT-4o, focusing on their architectural differences, performance, and real-world applications. GPT-4, built upon the Transformer architecture, has set new standards in natural language processing (NLP) with its capacity to generate coherent and contextually relevant text across a wide range of tasks. However, its computational demands, requiring substantial hardware resources, make it less accessible for smaller organizations and real-time applications. In contrast, GPT-4o addresses these challenges by incorporating optimizations such as model compression, parameter pruning, and memory-efficient computation, allowing it to deliver similar performance with significantly lower computational requirements. This paper examines the trade-offs between raw performance and computational efficiency, evaluating both models on standard NLP benchmarks and across diverse sectors such as healthcare, education, and customer service. Our analysis aims to provide insights into the practical deployment of these models, particularly in resource-constrained environments.

Keywords: ChatGPT-4; ChatGPT-4o; natural language processing; optimization; artificial intelligence; deep learning; transformer architecture

1. Introduction

OpenAI's GPT-4 has revolutionized natural language processing (NLP) with its ability to generate coherent, contextually relevant text across various applications, such as language translation, conversational agents, summarization, and content generation. By leveraging vast amounts of textual data and the transformer architecture's self-attention mechanisms, GPT-4 can understand and generate human-like responses with remarkable accuracy. This has had profound implications across multiple industries, including healthcare, customer service, and education, where natural language processing is integral to automation and enhanced user experience [1]. However, despite its success, GPT-4 has also faced challenges, particularly concerning its large model size, high computational costs, and energy consumption. Running such a model often requires substantial hardware resources, which can be prohibitive for smaller organizations or real-time applications.

In response to these challenges, OpenAI introduced GPT-4o, an optimized variant designed to reduce resource consumption and increase operational efficiency. GPT-4o retains much of GPT-4's advanced capabilities while integrating optimization techniques such as model compression, parameter pruning, and memory-efficient computation, making it more suitable for deployment in environments where computational resources are limited [2]. These optimizations are particularly relevant in scenarios that require real-time processing, such as interactive chatbots, virtual assistants, and other AI-powered systems where latency and response time are critical.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

This research investigates the technical and functional differences between these models and highlights their respective strengths, shortcomings, and potential use cases. By analyzing their architectures, we aim to understand the trade-offs made between raw performance and computational efficiency. Specifically, while GPT-4 excels in tasks requiring high accuracy and deep contextual understanding, GPT-4o offers a more balanced approach, ensuring a lower computational footprint without sacrificing too much performance. This balance makes GPT-4o a viable solution for industries prioritizing speed and cost-efficiency over marginal gains in performance [3,4].

In this paper, we aim to provide an in-depth comparison of GPT-4 and GPT-4o, both from a technical and application-oriented perspective. We begin by reviewing the architectural foundations of both models, focusing on the core mechanisms that drive their performance. Then, we delve into the specific optimizations implemented in GPT-4o, including techniques like quantization and adaptive batching, and evaluate their impact on both inference speed and overall system efficiency. Finally, we present performance evaluations based on standard NLP benchmarks and real-world applications to illustrate the practical implications of using these models across different sectors, such as education, healthcare, and customer service [5–8]. Through this analysis, we seek to provide actionable insights for practitioners and researchers in AI, helping them make informed decisions about model selection based on their specific requirements.

The paper is structured as follows: Section II provides an overview of beam management standards and summarizes related survey papers; Section III reviews state-of-the-art 6G beam management approaches; Section IV discusses possible avenues for future work; and Section V concludes the paper.

2. Overview of ChatGPT-4

2.1. Architecture

ChatGPT-4 is built upon the Transformer architecture, introduced by Vaswani et al. (2017), which has since become a cornerstone for most state-of-the-art NLP models due to its efficiency in handling long-range dependencies in text. The Transformer model, unlike traditional recurrent neural networks (RNNs) or long short-term memory (LSTM) models, leverages self-attention mechanisms, allowing it to focus on different parts of the input sequence dynamically, regardless of their distance from the current token being processed. This ability to capture complex relationships between words and phrases in a parallelizable manner is one of the key factors that enable GPT-4 to generate coherent, contextually relevant responses across diverse tasks [9].

GPT-4 employs a multi-layer architecture, where each layer consists of two primary sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feedforward network. The self-attention mechanism enables the model to weigh the importance of different tokens in the input sequence, which is crucial for tasks that require understanding context, such as translation, text generation, and summarization [10]. Additionally, GPT-4 includes positional encodings to account for the sequence order of words, addressing one of the main challenges in the original Transformer model, which was its inability to inherently capture sequential data.

One of the distinguishing features of GPT-4, compared to its predecessors like GPT-3, is its significantly larger scale. GPT-4 boasts billions of parameters, vastly expanding its capacity to learn and generalize from large and diverse datasets. This scale enables GPT-4 to not only handle a wide array of language tasks but also excel in zero-shot and few-shot learning scenarios, where the model can generalize to new tasks with little to no additional training data [1]. The ability to perform well with minimal training data highlights the model's deep understanding of language structure and context.

Moreover, GPT-4's training data includes an immense variety of text sources, ranging from books and research papers to websites and social media, allowing it to develop a nuanced understanding of different domains, genres, and styles. This diverse training corpus is essential for the model's generalization across multiple applications, such as conversational AI, content generation, and summarization, making it highly versatile for industry use [11]. For example, in conversational AI, GPT-4

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

can understand user inputs and generate contextually relevant responses with a level of coherence and fluency that closely mimics human conversation.

The deep and wide architecture of GPT-4 also benefits from enhanced regularization techniques, such as layer normalization and dropout, which prevent overfitting and improve generalization across tasks [12]. These mechanisms ensure that despite its massive scale, GPT-4 maintains robustness and stability during both training and inference. Additionally, the model employs advanced optimization techniques, such as Adam [13], to manage the complex gradients and vast parameter space during training, ensuring that the model converges efficiently.

As the Transformer architecture forms the backbone of GPT-4, its scalability and parallelism make it ideal for training on modern, distributed hardware systems such as GPUs and TPUs. This ability to scale efficiently is critical for training models like GPT-4, which require immense computational resources. The architecture's design allows for parallel processing of data, which significantly reduces the time and cost of training large-scale models [14]. These computational efficiencies are crucial for practical deployment in real-world applications, where response times and scalability are often key performance indicators.

In summary, GPT-4's architecture is a refined and massively scaled version of the original Transformer model, incorporating self-attention mechanisms, multi-layer structures, and advanced optimization techniques. Its ability to learn from vast datasets and generalize across tasks makes it a powerful tool for a wide array of natural language processing applications, ranging from conversational AI to complex text generation. However, the sheer size and computational demands of GPT-4 also highlight the ongoing challenges in model efficiency and scalability for practical deployment.

2.2. Training and Fine-Tuning

GPT-4 has undergone rigorous training on a massive corpus of diverse data sources, including books, academic papers, websites, articles, and user-generated content, allowing it to develop a broad understanding of language and context across different domains. This training data is not limited to any particular subject or genre, which gives GPT-4 the versatility to perform well across various tasks such as question answering, language translation, summarization, and conversational AI [1]. The data is carefully curated to include information from different languages, cultures, and disciplines, allowing the model to handle both common and specialized queries with a high degree of accuracy and fluency.

During the training phase, GPT-4 uses an autoregressive approach, predicting the next token in a sequence based on the previous tokens. This method helps the model generate coherent text that follows logical progression, which is particularly important for tasks like text completion and story generation [15]. The training process is computationally intensive, typically requiring vast amounts of GPU/TPU resources to process billions of parameters over extended periods. However, this large-scale pre-training equips GPT-4 with a deep understanding of linguistic structures, syntax, semantics, and even some level of reasoning, which can be leveraged across multiple use cases.

Fine-tuning, which is typically done after the initial training, plays a crucial role in adapting GPT-4 to specific applications. The fine-tuning process often involves supervised learning, where human feedback is used to guide the model towards desired outputs. This feedback is typically collected through human evaluators who review the model's responses and provide corrections or rankings based on quality, coherence, and relevance [16]. By iteratively adjusting the model's parameters based on this feedback, fine-tuning helps to improve the quality of the generated text in specific use cases, such as customer service chatbots, educational tools, or medical diagnostics [17,18].

An important aspect of fine-tuning is the incorporation of Reinforcement Learning from Human Feedback (RLHF). In RLHF, the model's responses are evaluated by human reviewers who score the quality of the generated outputs. These scores are then used as rewards in a reinforcement learning framework, helping GPT-4 to learn which types of responses are most preferred by users [19]. This approach has been particularly successful in improving the conversational capabilities of GPT-4, as it allows the model to better align its outputs with human preferences, making it more engaging and useful in real-world scenarios.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

The fine-tuning process also helps address specific challenges such as bias and safety in language models. While pre-trained models like GPT-4 can inadvertently generate biased or inappropriate content due to the nature of the data they are trained on, fine-tuning provides a mechanism to mitigate these issues. Through the use of curated datasets and explicit guidelines, the model can be adjusted to minimize harmful outputs and produce safer, more responsible content [20]. For example, when fine-tuned for customer support applications, the model is optimized to handle sensitive customer interactions with care, ensuring that it provides accurate and non-offensive responses.

Moreover, fine-tuning allows for domain-specific applications. For instance, a healthcare provider may fine-tune GPT-4 using medical literature and guidelines to ensure that the model is capable of providing reliable information related to healthcare [5–8]. Similarly, legal professionals can fine-tune the model on legal texts and case law, making it suitable for assisting with legal research or drafting contracts [21]. This ability to specialize the model for particular fields further enhances its utility across various industries.

In conversational AI, GPT-4's fine-tuned models have become popular choices for creating sophisticated chatbots and virtual assistants. These models are able to maintain coherence and relevance throughout extended interactions, making them more reliable for customer service, technical support, or personal assistants like OpenAI's own ChatGPT, which is based on fine-tuned versions of GPT-4 [22]. Fine-tuning also allows these systems to handle a wide range of conversational contexts, from casual small talk to more complex queries that require a deeper understanding of the user's intent.

In conclusion, GPT-4's rigorous training on a massive, diverse dataset, combined with targeted fine-tuning techniques such as supervised learning, RLHF, and domain-specific adjustments, enables it to perform at a high level across various NLP tasks. This flexibility and adaptability are key reasons why GPT-4 is widely adopted in both general-purpose and specialized applications. By incorporating human feedback and reinforcement learning, fine-tuning helps align the model's outputs with user expectations, ensuring that it remains coherent, relevant, and useful in real-world contexts.

2.3. Performance

As one of the most advanced versions of OpenAI's GPT series, GPT-4 achieves state-of-the-art results on several NLP benchmarks, including tasks like language generation, question answering, and text classification [1]. GPT-4 has shown remarkable performance on a wide range of tasks, often outperforming previous models in both zero-shot and few-shot learning scenarios, where the model is tested on tasks with little to no specific task-related training data [15]. Its ability to generate human-like, contextually coherent text has set new standards in natural language processing, making it suitable for diverse applications, including content creation, summarization, code generation, and even creative writing.

The model's strength lies in its massive scale, both in terms of its architecture and the volume of data it has been trained on. GPT-4 features billions of parameters, allowing it to model complex linguistic patterns and capture nuanced relationships between words, phrases, and concepts. This large parameter space enables GPT-4 to exhibit impressive generalization abilities, making it highly adaptable across domains. For example, it can handle technical jargon in healthcare, finance, and legal texts as well as engage in casual conversation with users in customer service settings [11]. Additionally, GPT-4's fine-tuning capabilities allow developers to tailor the model to specific use cases, further enhancing its versatility across industries.

Despite these advancements, GPT-4's resource demands pose significant challenges. The model requires considerable computational power and storage due to its vast number of parameters and the complexity of its operations. During both the training and inference phases, GPT-4 relies on large-scale GPU clusters or specialized hardware like TPUs (Tensor Processing Units), making it difficult for smaller organizations or developers without access to such resources to deploy the model effectively [14]. Training GPT-4 from scratch can cost millions of dollars in computational resources

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

and energy, and even using pre-trained models can require expensive cloud computing infrastructure to handle real-time, large-scale deployments [23].

In terms of environmental impact, the computational intensity of training and deploying GPT-4 raises concerns about energy consumption and carbon emissions. Recent studies have shown that large language models like GPT-4 contribute significantly to carbon footprints due to the extensive energy demands of training and inference on high-performance hardware [24]. This has led to increasing discussions in the AI community about the need for more energy-efficient models and the development of green AI technologies that balance performance with sustainability.

Moreover, the sheer size of GPT-4 introduces latency issues in real-time applications, especially those requiring fast response times, such as voice assistants, real-time translation, or conversational AI in customer service. The model's inference times can be too slow for certain real-time applications, leading to delays that may negatively impact user experience [25]. To mitigate this, organizations often resort to techniques such as model distillation or the use of smaller, optimized versions like GPT-4o, which offer faster processing times with a reduced computational footprint [2].

Another limitation of GPT-4's scale is the significant memory and storage requirements, both for storing the model and during runtime. These memory demands can become prohibitive, especially when deploying the model at scale across multiple devices or systems. Running GPT-4 for enterprise-level solutions, such as in large customer support centers or interactive learning platforms, often requires specialized infrastructure capable of handling these demands, making it impractical for smaller organizations or individual developers to implement [3,4].

Despite these challenges, the impact of GPT-4 on NLP research and industry is undeniable. Its state-of-the-art results across various benchmarks, including the General Language Understanding Evaluation (GLUE) benchmark and SuperGLUE, demonstrate its cutting-edge capabilities in text understanding and generation [26]. The model's ability to achieve high scores in multiple NLP tasks without requiring task-specific architectures or extensive task-specific training is a testament to the robustness and flexibility of the Transformer-based approach that underpins it.

In summary, while GPT-4 represents a remarkable leap in NLP performance with its ability to achieve state-of-the-art results across several benchmarks, it comes with significant trade-offs in terms of computational requirements, energy consumption, and latency. These limitations underscore the importance of ongoing research into model optimization and efficient deployment strategies, as well as the development of lighter, more resource-efficient models like GPT-4o to make such technologies more accessible and sustainable.

3. Overview of ChatGPT-4o

3.1. Architectural Enhancements

ChatGPT-4o, where the "o" stands for "optimized," is a variant of GPT-4 designed to offer similar performance with reduced computational demands, making it more accessible for real-time and resource-constrained applications. It retains the core elements of the Transformer architecture, which enables it to process text efficiently by employing self-attention mechanisms and multi-layer architectures. However, several crucial optimizations have been introduced to enhance its efficiency without significantly compromising performance. These optimizations include more efficient parameter distribution, layer-wise pruning, quantization techniques, and improved memory management strategies to reduce latency and energy consumption during both training and inference.

One of the key techniques employed in GPT-4o is parameter pruning, which reduces the number of parameters in the model by identifying and removing redundant or less important connections between layers [27]. Pruning not only reduces the overall size of the model but also accelerates inference times by minimizing the number of computations needed for each forward pass. Layer-wise pruning further improves this process by selectively reducing parameters in specific layers where redundancy is high, while maintaining the integrity of layers that are more critical to the model's

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

overall performance. This approach allows GPT-4o to preserve its ability to handle complex NLP tasks, such as language generation and text classification, with minimal loss in accuracy compared to GPT-4.

In addition to pruning, quantization techniques are applied to GPT-4o to optimize memory usage and computational efficiency. Quantization involves representing the model's weights and activations with lower precision (e.g., using 8-bit integers instead of 32-bit floating-point numbers), thereby significantly reducing memory requirements and computational costs [28]. This is especially beneficial during inference, where faster response times are often required for real-time applications like chatbots and virtual assistants. Quantization has been shown to have minimal impact on the model's performance while providing substantial gains in speed and energy efficiency, making it a key optimization technique in GPT-4o.

Another critical optimization in GPT-4o is dynamic batching. Unlike GPT-4, which typically processes a fixed number of inputs per batch, GPT-4o dynamically adjusts batch sizes based on the complexity and length of the input data. This adaptive batching mechanism allows the model to process simpler tasks more quickly, thereby reducing latency in real-time applications [14]. Dynamic batching is particularly useful in environments where varying query lengths and complexity levels are common, such as customer service platforms where the model might simultaneously handle brief inquiries and more complex support requests.

GPT-4o also incorporates improved memory management techniques, which optimize how data is stored and accessed during training and inference. One such method is activation checkpointing, where intermediate activations (data produced by layers during the forward pass) are not stored for every layer but instead recomputed during the backward pass when needed. This significantly reduces the memory footprint during training, allowing larger models to be trained on the same hardware without exhausting memory resources [29]. Additionally, memory-efficient attention mechanisms have been integrated into GPT-4o, allowing it to handle longer input sequences without suffering from the quadratic scaling of traditional attention mechanisms [30]. This is crucial for tasks like document summarization and large-scale text analysis, where the ability to process longer inputs efficiently can greatly enhance performance.

Moreover, GPT-4o benefits from model parallelism, which distributes the model's computations across multiple devices or processors. By breaking down the model into smaller, manageable segments, GPT-4o can leverage multiple GPUs or TPUs more effectively, reducing the overall time and energy required for both training and inference [14]. This approach not only enhances speed but also reduces power consumption, making it an ideal choice for industries seeking to deploy AI models in a more sustainable and cost-effective manner.

These optimizations in GPT-4o result in a model that offers a balance between performance and efficiency, making it suitable for use in various industries where real-time interactions and resource constraints are key considerations. For instance, while GPT-4o retains the capacity to generate high-quality, contextually relevant text for customer service chatbots, virtual assistants, and educational platforms, it does so with lower latency and reduced hardware requirements. This makes it a more scalable solution for businesses that require the advanced capabilities of GPT-4 but cannot afford the extensive computational resources needed to run the original model [3,4].

Despite these improvements, there are trade-offs in terms of model accuracy and performance in highly complex tasks. While GPT-4o can match GPT-4 in most general NLP tasks, such as text classification and conversation generation, its performance might slightly lag behind when handling tasks that require a deep understanding of nuanced context or highly specialized knowledge [11]. This is primarily due to the reduced number of parameters and lower precision in some computations, which can lead to minor losses in the model's ability to capture intricate linguistic patterns. However, for many practical applications, these performance differences are negligible, and the benefits in speed and efficiency make GPT-4o a compelling choice for organizations looking to implement cutting-edge NLP solutions at scale.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

3.2. Optimization Techniques

The key difference between GPT-4 and GPT-4o lies in the optimizations that have been implemented in the latter, aimed at enhancing computational efficiency while maintaining competitive performance. These optimizations are critical for reducing the resource consumption of large-scale models, making GPT-4o more accessible and practical for real-time applications. The improvements include techniques such as model compression, memory-efficient layers, and dynamic batching, each of which contributes to the model's ability to handle language tasks with reduced latency and lower energy consumption.

3.2.1. Model Compression

One of the most significant optimizations in GPT-4o is model compression, which involves reducing the size of the model without significantly affecting its performance. Two of the primary techniques used for this purpose are quantization and pruning.

Quantization reduces the precision of the model's weights and activations, typically from 32-bit floating-point numbers to 8-bit integers or other lower-precision formats. This allows for a substantial decrease in memory usage and computational requirements. Quantization is especially effective during inference, where it reduces the load on processors, allowing for faster computations and decreased power consumption [28]. Research has shown that with proper tuning, quantized models can maintain nearly the same performance levels as their full-precision counterparts, making it an attractive solution for deploying large-scale language models like GPT-4o in resource-constrained environments [31].

Pruning, on the other hand, removes redundant or less significant weights from the model. By systematically identifying and eliminating parameters that contribute minimally to the model's performance, pruning helps reduce the overall size of the model. This leads to fewer computations during both training and inference, speeding up the model while maintaining most of its accuracy. Structured pruning, which focuses on pruning entire layers or neurons, ensures that the model remains operationally efficient and that the pruning process does not introduce irregularities that might complicate parallel processing on GPUs or TPUs [27]. In GPT-4o, this technique is applied layer-wise to ensure that model integrity is preserved, especially in layers critical for understanding and generating language.

Together, these compression techniques reduce the size of GPT-4o, making it more feasible to deploy on devices with limited memory or in situations where computational resources are at a premium. This makes GPT-4o particularly suited for mobile applications, edge computing, and other environments where real-time processing with minimal hardware is crucial [32].

3.2.2. Memory-Efficient Layers

Another key optimization in GPT-4o is the implementation of memory-efficient transformer layers. In the standard Transformer architecture, the self-attention mechanism, which computes relationships between every token in a sequence, scales quadratically with the input length. This scaling issue becomes a bottleneck for longer input sequences, leading to significant memory and time costs during both training and inference.

GPT-4o addresses this issue by incorporating more memory-efficient variations of the self-attention mechanism, such as linear attention and sparse attention, which reduce the computational complexity from quadratic to linear or sub-quadratic [30]. For example, Linformer and Reformer are architectures that approximate the full self-attention computation by compressing the attention matrix or using locality-sensitive hashing to focus on more relevant tokens. These approaches drastically reduce the memory footprint of the model while still capturing essential contextual relationships in the input data [33].

Additionally, GPT-4o uses activation checkpointing, a memory-saving technique that stores fewer intermediate activations during the forward pass and recomputes them during the backward pass. This allows the model to train on larger batches or longer sequences without exceeding memory

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

limits, improving overall efficiency during training [34]. These memory-efficient layers make GPT-4o faster and more scalable for real-world applications, especially when dealing with large-scale data or continuous interactions like customer service or conversational AI.

3.2.3. Dynamic Batching

Another notable optimization in GPT-4o is the use of dynamic batching or adaptive batching techniques, which improve the model's throughput by adjusting batch sizes based on the complexity of the input data. In traditional batch processing, models process a fixed number of inputs at a time, which can be inefficient when tasks vary significantly in complexity or length. GPT-4o adapts to these variations by adjusting the batch size in real time, allowing it to handle simpler tasks more quickly while allocating more resources to complex ones. This not only enhances the model's throughput but also ensures that latency remains low, even when dealing with a diverse set of language tasks [14].

Dynamic batching is particularly beneficial in environments where GPT-4o is deployed for real-time processing of multiple concurrent requests, such as in chatbot systems or interactive virtual assistants. For instance, short queries can be processed in smaller, more frequent batches, ensuring that users receive quick responses, while longer or more complex requests are grouped into larger batches to optimize resource utilization. This flexibility allows GPT-4o to achieve higher throughput while maintaining the responsiveness needed for time-sensitive applications.

Moreover, dynamic batching also improves the overall energy efficiency of GPT-4o, as it reduces the time spent waiting for input data and maximizes the utilization of available hardware resources. In large-scale deployments, such as in cloud-based AI services, this can result in substantial cost savings and improved environmental sustainability, which are increasingly important considerations in the development and deployment of AI technologies [24].

In summary, the optimizations implemented in GPT-4o, including model compression, memory-efficient layers, and dynamic batching, collectively make it a more efficient and practical alternative to GPT-4. These improvements reduce the computational demands of the model, making it suitable for deployment in a variety of environments, from mobile devices to large-scale cloud infrastructures. By maintaining a balance between performance and resource efficiency, GPT-4o opens up new possibilities for real-time applications in industries such as customer service, healthcare, and education, where quick and reliable responses are crucial. These advancements reflect the growing trend toward creating AI models that are not only powerful but also more accessible and sustainable.

3.3. Performance

GPT-4o performs similarly to GPT-4 on several benchmarks but does so with a significantly lower computational footprint, making it a valuable alternative for many real-time applications and resource-constrained environments. GPT-4o maintains much of the language understanding, text generation, and contextual reasoning capabilities of GPT-4, allowing it to deliver state-of-the-art performance across a wide range of tasks, including question answering, summarization, and translation [14]. However, the optimizations that make GPT-4o more efficient also introduce trade-offs, especially when it comes to tasks that require complex, nuanced text generation or deep contextual understanding, where GPT-4 has a slight edge.

The reduced computational footprint of GPT-4o stems from optimizations like model compression, layer pruning, and quantization, which allow the model to process data more efficiently while retaining most of its performance capabilities. These optimizations make GPT-4o a strong candidate for applications that require real-time interactions, such as conversational AI systems, customer support chatbots, and virtual assistants, where low latency and fast response times are critical [2]. In these scenarios, GPT-4o's ability to process input quickly without requiring massive computational resources ensures that users receive timely responses, making the model more scalable and cost-effective for companies that operate in high-traffic environments.

Additionally, GPT-4o's lower computational requirements make it ideal for deployment in resource-constrained environments, such as edge computing devices or mobile platforms. Many

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

industries, including automotive and healthcare, are increasingly leveraging AI models for real-time decision-making at the edge, where cloud infrastructure may not be readily available or where low-latency, on-device processing is preferred. For example, in autonomous vehicles, GPT-4o can be used to power natural language interfaces or assist with real-time voice-activated controls, all while operating within the limited processing power of the vehicle's onboard systems [35]. Similarly, in telemedicine, GPT-4o can be integrated into mobile apps or diagnostic tools to provide real-time responses to patient queries or assist clinicians with decision support, even in remote or bandwidth-limited environments [5–8].

While GPT-4o excels in efficiency, its performance in generating very complex or highly specialized text may not always match GPT-4. Tasks that require deep, intricate reasoning or high levels of creativity—such as writing long-form content, generating detailed technical documents, or engaging in highly specialized fields like legal analysis or scientific research—tend to benefit from the larger parameter space and richer representations provided by GPT-4 [11]. This makes GPT-4 the better option for tasks where maximum fidelity and precision are required, especially in academic or professional contexts that demand extensive contextual understanding or where minor errors in text generation could lead to significant consequences.

However, for many commercial and industrial applications, the marginal difference in performance between GPT-4 and GPT-4o is outweighed by GPT-4o's improved efficiency. Industries like e-commerce, finance, and logistics, which rely on large-scale customer interactions, data analysis, and automation, can greatly benefit from GPT-4o's balance of performance and scalability [21]. For instance, in customer service chatbots, GPT-4o can deliver fast, contextually relevant responses to frequently asked questions while running on more affordable hardware, thereby reducing operational costs for companies that need to handle high volumes of customer interactions. In financial services, GPT-4o's ability to perform sentiment analysis on customer feedback or analyze large datasets for trends and insights, all while minimizing infrastructure costs, makes it a practical solution for many real-world applications [36].

Another key advantage of GPT-4o is its reduced energy consumption, which is becoming an increasingly important factor in AI deployment, especially as companies and researchers focus on making AI more sustainable. Large language models like GPT-4 have been criticized for their significant carbon footprints due to the vast amounts of energy required for training and inference [24]. GPT-4o's optimizations help mitigate this issue by lowering the computational power and energy required for each inference, making it a more environmentally friendly option. This reduced energy consumption can also translate into lower operational costs for businesses running AI models at scale, particularly in industries that depend on cloud infrastructure where energy use is directly tied to cost [23].

Moreover, GPT-4o is particularly well-suited for personalization at scale, where models must process large amounts of user data in real time to provide tailored experiences. In marketing and advertising, GPT-4o can be deployed to deliver personalized product recommendations or generate targeted content based on user preferences, without the need for extensive backend infrastructure. Its ability to handle real-time data streams efficiently makes it a versatile tool for creating dynamic, personalized user experiences across digital platforms [3,4].

In conclusion, while GPT-4o may not match GPT-4 in every domain—particularly in generating complex, highly nuanced text—its significantly lower computational footprint makes it a viable option for a wide range of commercial and industrial applications. By maintaining strong performance across several NLP benchmarks and reducing latency, energy consumption, and operational costs, GPT-4o has the potential to revolutionize real-time applications in fields ranging from customer service to healthcare, education, and beyond.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

4. Comparative Analysis

4.1. Accuracy and Efficiency

GPT-4 maintains a slight edge in terms of pure accuracy, particularly in tasks that require a deeper understanding of nuanced context, complex reasoning, or the generation of highly sophisticated language. This is largely due to the sheer size of GPT-4's model architecture, which includes billions of parameters, enabling it to capture more intricate relationships between words and phrases across longer contexts [1]. For example, in tasks such as long-form content generation, abstract reasoning, or creative writing, GPT-4's ability to leverage its larger parameter space allows it to better handle subtle nuances, idiomatic expressions, and complex sentence structures. This makes it especially valuable in highly specialized domains like legal research, academic writing, or scientific text generation, where precision and a deeper understanding of the subject matter are critical [11].

The bar chart in Figure 1 compares the performance of multiple large language models across six benchmark datasets: MMLU, GPQA, MATH, HumanEval, MGSM, and DROP. The models evaluated include GPT-4o, GPT-4T, GPT-4 (initial release on 23-03-14), Claude3 Opus, Gemini Pro 1.5, Gemini Ultra 1.0, and Llama3 400b. GPT-4o consistently demonstrates the highest performance across most benchmarks, with standout accuracy on MMLU (88.7%) and MGSM (90.5%). GPT-4T and GPT-4 perform closely in many datasets, with Gemini Pro 1.5 and Gemini Ultra 1.0 also achieving competitive results, particularly in MATH and HumanEval. Notably, Claude3 Opus exhibits relatively weaker performance, especially in GPQA and MATH. This chart highlights the comparative strengths and weaknesses of the evaluated models in a variety of accuracy-based tasks.

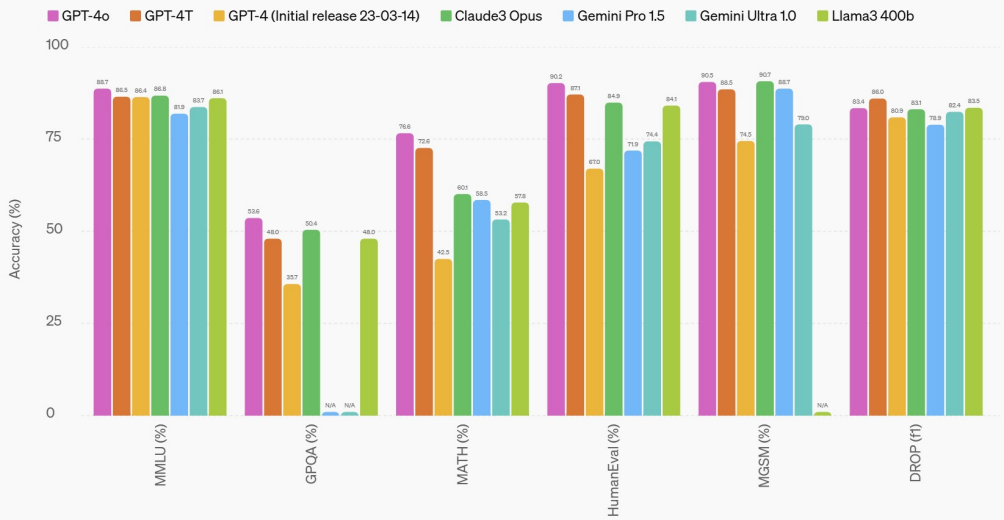

Figure 1. Text Evaluation performance evaluation (By OpenAI) [37].

On the other hand, GPT-4o offers a more balanced trade-off between performance and efficiency, making it more suitable for a wide array of applications where real-time response times and lower hardware requirements are essential. While GPT-4o sacrifices some of the nuanced accuracy found in GPT-4, it achieves nearly equivalent results on many standard benchmarks, particularly in tasks that do not require extensive reasoning or highly specific domain knowledge [14]. For example, GPT-4o performs exceptionally well in tasks like question answering, sentiment analysis, summarization, and conversational AI, where the primary requirement is delivering contextually relevant and coherent responses at high speed. This makes it ideal for applications such as customer service chatbots, interactive voice response (IVR) systems, and real-time text translation, where responsiveness and scalability are often more important than achieving perfect linguistic precision [3,4].

The bar chart in Figure 2 illustrates the M3Exam zero-shot results comparing the performance of GPT-4 and GPT-4o across various languages and question categories. Languages evaluated include Afrikaans, Chinese, English, Italian, Javanese, Portuguese, Swahili, Thai, and Vietnamese, with

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

questions categorized into "All Questions" and "Vision Questions." GPT-4o consistently outperforms GPT-4 across most languages and question types, demonstrating significant improvements, particularly in Afrikaans, Chinese, and English. The chart highlights the robust enhancements of GPT-4o in multilingual and vision-related tasks, showcasing its superior adaptability and accuracy compared to GPT-4.

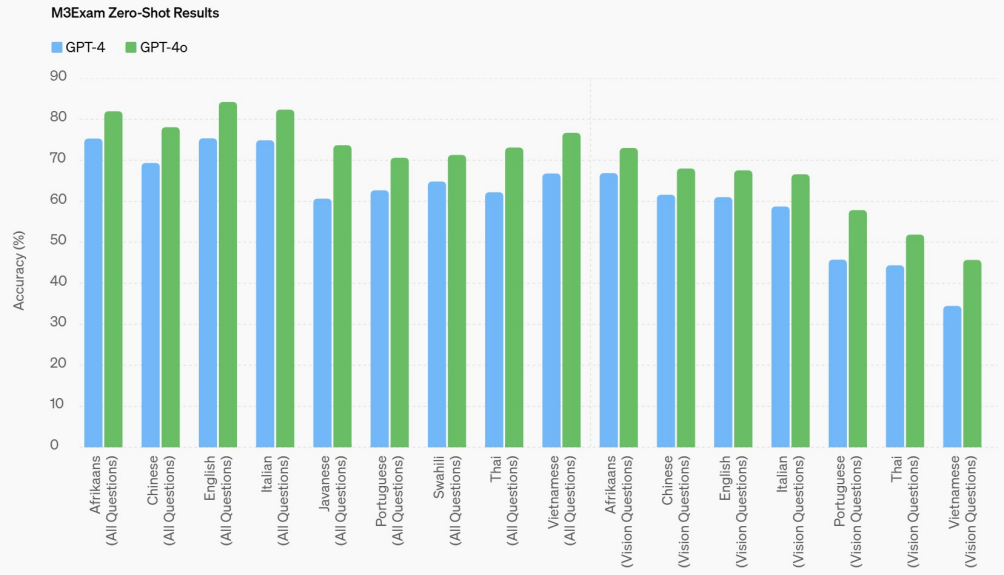

Figure 2. M3Exam Zero-Shot Performance Evaluation [By OpenAI] [37].

The trade-off between performance and efficiency in GPT-4o is largely achieved through optimizations such as parameter pruning, quantization, and memory-efficient layers, which reduce the model's overall size and computational complexity without significantly degrading its ability to understand and generate text [28]. These optimizations allow GPT-4o to run on a wider range of hardware, including mobile devices, edge computing platforms, and lower-powered cloud infrastructure, making it more accessible for businesses and developers with limited computational resources. For instance, in mobile applications or edge AI environments, where processing power and memory are often constrained, GPT-4o's lower resource requirements enable faster inference times and lower energy consumption, making it feasible to deploy AI-driven functionalities like voice assistants, real-time translation tools, and text-based interaction systems directly on-device [35].

Another significant advantage of GPT-4o is its ability to handle high-volume, real-time workloads more efficiently than GPT-4. In environments such as e-commerce platforms, financial services, and telecommunication networks, where AI models must process hundreds or thousands of user interactions simultaneously, GPT-4o's optimizations lead to faster throughput and lower latency, ensuring that customers receive timely responses without overwhelming the system's resources. For example, a customer support chatbot powered by GPT-4o can efficiently handle a large number of queries in parallel, delivering accurate, contextually appropriate responses in a fraction of the time it would take GPT-4 to process the same number of requests, all while using fewer computational resources.

Although GPT-4o excels in efficiency, its performance can fall short of GPT-4 in tasks that require long-term coherence or deeper context retention across lengthy documents. In domains such as medical diagnostics or legal document review, where a small error or misinterpretation of context can have significant consequences, GPT-4's more comprehensive understanding and ability to maintain context over longer sequences give it an edge [5–8]. However, for most practical applications, particularly those that involve shorter or less complex interactions, the difference in performance between GPT-4 and GPT-4o is often negligible, making GPT-4o the more cost-effective and operationally efficient choice.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

Moreover, GPT-4o's optimizations are particularly beneficial in terms of environmental sustainability and cost-efficiency. Large-scale AI models like GPT-4 are notorious for their high energy consumption and significant carbon footprints due to the extensive computational resources required for both training and inference [24]. By contrast, GPT-4o's reduced size and more efficient architecture lead to lower energy use, which translates into lower operational costs and a smaller environmental impact. For companies looking to scale AI solutions while minimizing both financial and ecological costs, GPT-4o provides a compelling alternative to the more resource-intensive GPT-4 [23].

The bar chart in Figure 3 illustrates the performance of various AI agents and humans under different time constraints, measured by the fraction of tasks completed (weighted) with a 95% confidence interval. The AI agents evaluated include GPT-4o mini, GPT-4t, GPT-4o, Claude 3 Sonnet, Claude 3 Opus, and Claude 3.5 Sonnet. Human performance serves as a benchmark under four time conditions: no time limit, 8 hours, 2 hours, and 30 minutes, with an additional reference for human performance within 10 minutes.

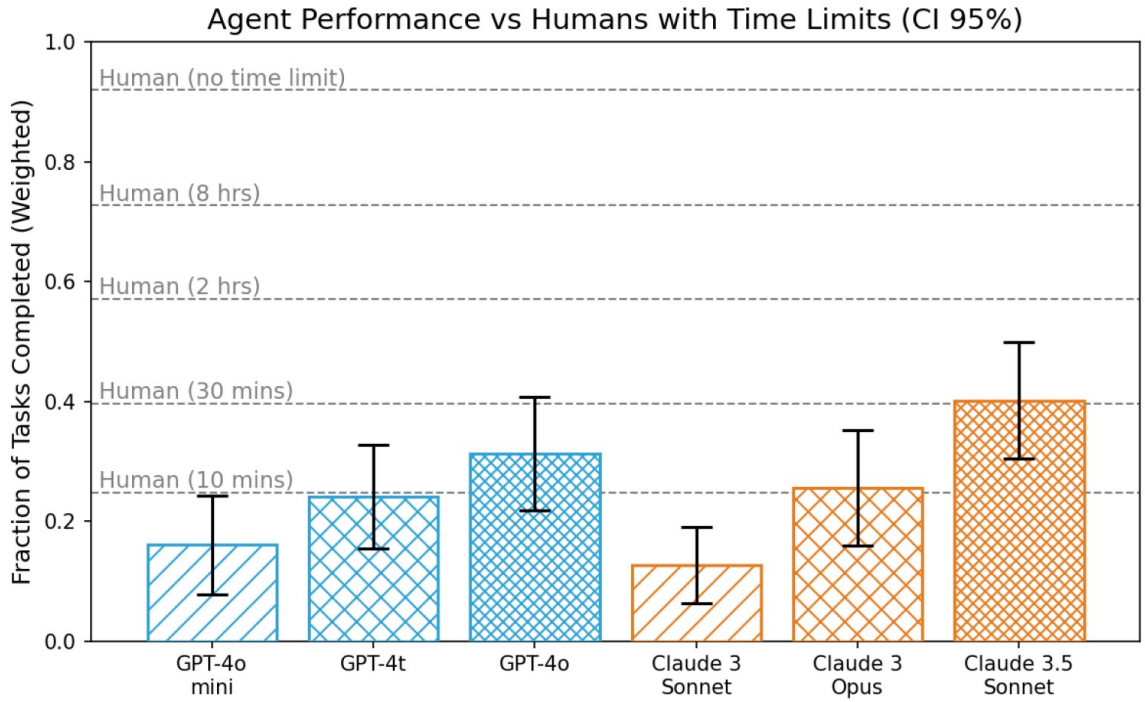

Figure 3. Agent Performance compared with Human [38].

Among the AI agents, Claude 3.5 Sonnet achieves the highest task completion rate, approaching the performance of humans under a 2-hour limit. GPT-4o demonstrates notable performance as well, outperforming Claude 3 Sonnet and Claude 3 Opus but falling short of Claude 3.5 Sonnet. GPT-4t and GPT-4o mini show lower completion rates, with substantial gaps compared to their higher-performing counterparts. The chart highlights the varying levels of capability among AI agents and underscores the gap between AI performance and human benchmarks under constrained conditions.

The Figure 4 summarizes the performance of five AI models—GPT-4o, GPT-4T (2024-04-09), Gemini 1.0 Ultra, Gemini 1.5 Pro, and Claude Opus—across various evaluation datasets: MMMU, MathVista, AI2D, ChartQA, DocVQA, ActivityNet, and EgoSchema. GPT-4o consistently achieves the highest or near-highest scores across the majority of datasets, with standout results such as 94.2% on AI2D, 92.8% on DocVQA, and 72.2% on EgoSchema. GPT-4T follows closely, demonstrating strong performance on datasets like AI2D (89.4%) and DocVQA (87.2%), though it trails GPT-4o by a noticeable margin.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.5	63.2	

Figure 4. Vision understanding evaluation [By OpenAI] [37].

The Gemini models show competitive results, with Gemini 1.5 Pro outperforming Gemini 1.0 Ultra in most datasets. For example, Gemini 1.5 Pro achieves 81.3% on ChartQA and 90.9% on DocVQA, slightly edging ahead of Gemini 1.0 Ultra in these tasks. Claude Opus performs well in specific datasets, notably AI2D (88.1%) and DocVQA (89.3%), but generally lags behind GPT-4o. This comparative analysis highlights GPT-4o's superior versatility and accuracy across a range of challenging evaluation tasks.

The Table 1 presents a comparison of the performance between GPT-4T (May 2024) and GPT-4o across multiple medical and clinical datasets, evaluated in both 0-shot and 5-shot settings. The datasets include MedQA (USMLE, Taiwan, and Mainland China) and MMLU categories such as Clinical Knowledge, Medical Genetics, Anatomy, Professional Medicine, College Biology, and College Medicine, as well as the MedMCQA Dev dataset.

GPT-4o consistently outperforms GPT-4T across all datasets and settings. For instance, in the MedQA USMLE 4 Options dataset, GPT-4o achieves 0.89 in both 0-shot and 5-shot settings compared to GPT-4T's 0.78 and 0.81, respectively. Similar improvements are observed in the MMLU Medical Genetics dataset, where GPT-4o scores 0.96 and 0.95 in 0-shot and 5-shot settings, surpassing GPT-4T's scores of 0.93 and 0.95. Notably, GPT-4o shows substantial gains in challenging datasets like MedQA Taiwan (0.91 in both 0-shot and 5-shot) and MMLU Anatomy (0.89 in both settings). These results highlight GPT-4o's superior ability to handle complex medical and clinical tasks compared to its predecessor, GPT-4T.

Table 2 compares the accuracy of four models—GPT-3.5 Turbo, GPT-4o mini, GPT-4, and GPT-4o—on the Translated ARC-Easy dataset across six languages: English, Amharic, Hausa, Northern Sotho (Sepedi), Swahili, and Yoruba. The evaluation is conducted in a 0-shot setting, with higher percentages indicating better performance.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

Table 1. Medical Knowledge Task Performance Comparison of GPT-4T and GPT-4o [38].

Medical Knowledge Task	GPT-4T (May 2024)	GPT-4o
MedQA USMLE 4 Options (0-shot)	0.78	**0.89**
MedQA USMLE 4 Options (5-shot)	0.81	**0.89**
MedQA USMLE 5 Options (0-shot)	0.75	**0.86**
MedQA USMLE 5 Options (5-shot)	0.78	**0.87**
MedQA Taiwan (0-shot)	0.82	**0.91**
MedQA Taiwan (5-shot)	0.86	**0.91**
MedQA Mainland China (0-shot)	0.72	**0.84**
MedQA Mainland China (5-shot)	0.78	**0.86**
MMLU Clinical Knowledge (0-shot)	0.85	**0.92**
MMLU Clinical Knowledge (5-shot)	0.87	**0.92**
MMLU Medical Genetics (0-shot)	0.93	**0.96**
MMLU Medical Genetics (5-shot)	0.95	**0.95**
MMLU Anatomy (0-shot)	0.79	**0.89**
MMLU Anatomy (5-shot)	0.85	**0.89**
MMLU Professional Medicine (0-shot)	0.92	**0.94**
MMLU Professional Medicine (5-shot)	0.92	**0.94**
MMLU College Biology (0-shot)	0.93	**0.95**
MMLU College Biology (5-shot)	0.95	**0.95**
MMLU College Medicine (0-shot)	0.74	**0.84**
MMLU College Medicine (5-shot)	0.80	**0.89**
MedMCQA Dev (0-shot)	0.70	**0.77**
MedMCQA Dev (5-shot)	0.72	**0.79**

Table 2. Accuracy on Translated ARC-Easy (%, higher is better), 0-shot [38].

Model	**English** *(n=523)*	**Amharic** *(n=518)*	**Hausa** *(n=475)*	**Northern Sotho** *(Sepedi, n=520)*	**Swahili** *(n=520)*	**Yoruba** *(n=520)*
GPT 3.5 Turbo	80.3	6.1	26.1	26.9	62.1	27.3
GPT-4o mini	93.9	42.7	58.5	37.4	76.9	43.8
GPT-4	89.7	27.4	28.8	30	83.5	31.7
GPT-4o	**94.8**	**71.4**	**75.4**	**70**	**86.5**	**65.8**

GPT-4o consistently achieves the highest accuracy across all languages, demonstrating superior multilingual capability. For English, GPT-4o scores 94.8%, outperforming GPT-4 (89.7%) and GPT-4o mini (93.9%). In low-resource languages such as Amharic and Hausa, GPT-4o achieves remarkable accuracy of 71.4% and 75.4%, respectively, compared to GPT-4's 27.4% and 28.8%. Similar trends are observed for Northern Sotho and Yoruba, where GPT-4o scores 70% and 65.8%, significantly surpassing the other models. For Swahili, GPT-4o achieves 86.5%, a notable improvement over GPT-4's 83.5%. These results underscore GPT-4o's advancements in handling both high- and low-resource languages effectively, making it a robust model for multilingual tasks.

Table 3 reports the accuracy of four models—GPT-3.5 Turbo, GPT-4o mini, GPT-4, and GPT-4o—on the Translated TruthfulQA dataset across six languages: English, Amharic, Hausa, Northern Sotho (Sepedi), Swahili, and Yoruba. The evaluation is conducted in a 0-shot setting, with accuracy scores indicating the percentage of truthful responses.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

Table 3. Accuracy on Translated TruthfulQA (%, higher is better), 0-shot [38].

Model	English *(n=809)*	Amharic *(n=808)*	Hausa *(n=808)*	Northern Sotho *(Sepedi, n=809)*	Swahili *(n=808)*	Yoruba *(n=809)*
GPT 3.5 Turbo	53.6	26.1	29.1	29.3	40	28.3
GPT-4o mini	66.5	33.9	42.1	36.1	48.4	35.8
GPT-4	81.3	42.6	37.6	42.9	62	41.3
GPT-4o	**81.4**	**55.4**	**59.2**	**59.1**	**64.4**	**51.1**

GPT-4o consistently demonstrates the highest accuracy across all languages, reflecting its superior ability to generate truthful outputs in multilingual contexts. For English, GPT-4o achieves an accuracy of 81.4%, slightly higher than GPT-4's 81.3%, and significantly outperforming GPT-4o mini (66.5%) and GPT-3.5 Turbo (53.6%). In low-resource languages like Amharic and Hausa, GPT-4o achieves 55.4% and 59.2%, respectively, far surpassing GPT-4's scores of 42.6% and 37.6%. Similarly, for Northern Sotho, Swahili, and Yoruba, GPT-4o scores 59.1%, 64.4%, and 51.1%, respectively, outperforming all other models by substantial margins. These results highlight GPT-4o's advancements in truthfulness and its robust handling of both high- and low-resource languages in a zero-shot setting.

The Table 4 compares the performance of four models—GPT-3.5 Turbo, GPT-4o mini, GPT-4, and GPT-4o—on a dataset across three languages: Amharic ($n = 77$), Hausa ($n = 155$), and Yoruba ($n = 258$). GPT-4o consistently achieves the highest accuracy among all models, scoring 44.2% for Amharic, 59.4% for Hausa, and 60.5% for Yoruba. In comparison, GPT-4 performs slightly lower, with scores of 41.6%, 41.9%, and 41.9%, respectively, across the three languages. GPT-4o mini also performs competitively but lags behind GPT-4o, scoring 33.8% for Amharic, 43.2% for Hausa, and 44.2% for Yoruba. GPT-3.5 Turbo demonstrates the lowest performance, with scores of 22.1% for Amharic, 32.3% for Hausa, and 28.3% for Yoruba. These results highlight GPT-4o's superior capabilities in handling low-resource languages compared to its predecessors and smaller variants.

Table 4. Accuracy on Dataset (%, higher is better), 0-shot [38].

Model	Amharic *(n=77)*	Hausa *(n=155)*	Yoruba *(n=258)*
GPT 3.5 Turbo	22.1	32.3	28.3
GPT-4o mini	33.8	43.2	44.2
GPT-4	41.6	41.9	41.9
GPT-4o	**44.2**	**59.4**	**60.5**

In summary, while GPT-4 maintains an edge in tasks requiring intricate language generation and deep contextual understanding, GPT-4o strikes a better balance between performance and efficiency, making it more suitable for real-time applications with lower hardware requirements. Its ability to handle high-throughput, real-time processing and its lower resource demands make GPT-4o an attractive option for businesses and industries seeking to deploy AI at scale, particularly in scenarios where quick, reliable responses are critical, and computational resources are limited. This balance of efficiency and performance positions GPT-4o as a versatile tool in sectors such as customer service, finance, healthcare, and mobile technologies, where AI-driven applications need to operate efficiently without compromising too much on quality.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

4.2. Cost of Deployment

Due to its smaller computational demands, GPT-4o is generally more cost-effective to deploy in production environments, offering a significant advantage to organizations that prioritize efficiency and scalability. The reduced hardware and energy requirements of GPT-4o mean that it can be run on less expensive infrastructure compared to GPT-4, which is crucial for businesses operating on limited budgets or with restricted access to high-performance computing resources [23]. This makes GPT-4o particularly appealing to small and medium-sized enterprises (SMEs) and startups that seek to leverage advanced natural language processing (NLP) capabilities without incurring the substantial costs associated with maintaining large-scale AI models.

Organizations with limited hardware resources or those looking for cost-efficient solutions without sacrificing too much accuracy often prefer GPT-4o because it offers a good balance between performance and resource utilization. One of the key factors driving the cost savings associated with GPT-4o is its optimized model architecture, which employs techniques such as pruning, quantization, and memory-efficient attention mechanisms. These techniques reduce the number of parameters and the computational complexity of the model, enabling it to run efficiently on standard GPUs or even CPUs, rather than requiring costly high-end hardware like multi-GPU clusters or TPUs, which are typically needed to deploy larger models like GPT-4 [2].

In cloud-based environments, where businesses often pay based on the amount of computing power consumed, the cost savings of using GPT-4o can be significant. Cloud providers like AWS, Google Cloud, and Azure charge based on the computational resources used, meaning that GPT-4o's lighter architecture translates directly into lower operating costs for inference and real-time deployment [14]. For example, companies deploying large-scale conversational AI systems or customer service chatbots can use GPT-4o to handle high volumes of interactions without the need for extensive hardware scaling, thus reducing both infrastructure costs and energy consumption.

Furthermore, the ability to run GPT-4o on edge devices opens up new opportunities for cost savings in industries where cloud-based computing may not be practical or desirable due to latency, security, or privacy concerns. By deploying GPT-4o on edge devices like mobile phones, IoT devices, or local servers, organizations can perform AI inference closer to the source of data, reducing the need for continuous data transmission to and from the cloud [35]. This is particularly advantageous in sectors like manufacturing, healthcare, and automotive, where real-time processing is critical, and any delays could impact operational efficiency or safety. The ability to run complex NLP models at the edge, without requiring constant connectivity to a powerful cloud server, can result in lower operational costs and improved performance for applications like voice assistants, predictive maintenance systems, and smart medical devices.

In addition to reducing hardware and operational costs, GPT-4o's efficiency can also lower energy consumption, which has financial and environmental benefits. Large language models like GPT-4 have been criticized for their high energy use during both training and inference, contributing to the growing carbon footprint of AI systems [24]. In contrast, GPT-4o's optimizations allow it to deliver strong performance with a fraction of the energy required by larger models, making it a more sustainable option for organizations looking to reduce their environmental impact. This can also result in direct cost savings, especially for companies running AI models at scale, where energy consumption constitutes a significant portion of the overall operating expenses. By choosing GPT-4o, businesses not only save on hardware and cloud infrastructure costs but also reduce their energy bills, making it a win-win solution for both budget-conscious and environmentally conscious enterprises [23].

Moreover, GPT-4o is especially valuable in high-volume applications such as e-commerce, finance, and telecommunications, where AI models must handle large numbers of transactions or customer interactions in real time. For instance, in an e-commerce platform handling thousands of customer queries and transactions every second, the efficiency of GPT-4o enables the system to scale quickly and cost-effectively. The reduced computational requirements allow businesses to support more simultaneous users without having to invest heavily in expanding their infrastructure, which is partic-

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

ularly important during periods of peak demand, such as holiday shopping seasons or promotional events [3,4].

In financial services, GPT-4o's cost-efficiency is also a major advantage for tasks like fraud detection, sentiment analysis, and algorithmic trading, where real-time processing of vast datasets is required. Financial institutions can deploy GPT-4o to analyze market trends or monitor transactions at scale, ensuring fast, accurate results while keeping operational costs manageable. This makes it easier for firms to remain competitive without needing to invest in costly high-performance computing clusters or extensive cloud services [36].

For educational platforms and content creation tools, GPT-4o offers an affordable solution for generating text, tutoring students, or assisting in research, without the overhead of GPT-4. Schools, universities, and businesses offering AI-driven learning solutions can use GPT-4o to power virtual assistants, personalized learning systems, or automated content generation platforms, all while keeping their infrastructure costs within budget [5–8].

In conclusion, GPT-4o's smaller computational demands make it a highly cost-effective solution for organizations looking to deploy advanced NLP models in production environments. Its combination of lower hardware requirements, energy efficiency, and ability to scale in cloud or edge environments makes it an attractive option for businesses with limited resources or those seeking to minimize operational costs without sacrificing too much in terms of accuracy or performance. By offering a balance between efficiency and capability, GPT-4o enables a wider range of organizations, from SMEs to large enterprises, to leverage the power of advanced AI at a fraction of the cost of deploying larger models like GPT-4.

4.3. Real-World Applications

4.3.1. Education

In the realm of education, GPT-4 has demonstrated significant value in generating complex and detailed educational content, including lecture notes, exam questions, essays, and research papers. Its ability to understand and create sophisticated academic material makes it particularly useful for tasks that require a deep understanding of various subjects, including science, mathematics, history, and literature. For example, GPT-4 can assist in developing comprehensive learning modules that incorporate in-depth explanations, step-by-step problem-solving processes, and even suggest reading materials or resources for further study [11]. Educators and institutions can use GPT-4 to automate content creation, thereby streamlining curriculum development and providing personalized educational materials at scale.

However, in educational environments that require real-time, interactive feedback, GPT-4o proves to be a more practical and efficient solution. GPT-4o's lower computational requirements and faster response times make it ideal for student tutoring systems, where immediate feedback is essential for student engagement and learning outcomes. For instance, intelligent tutoring systems (ITS) powered by GPT-4o can interact with students in real time, guiding them through exercises, answering questions, and offering hints or explanations based on their progress. This interactivity is particularly beneficial for subjects like mathematics and language learning, where constant practice and immediate correction are key to mastering concepts [3,4]. In these applications, GPT-4o's ability to deliver quick responses while maintaining a high level of accuracy ensures that students receive personalized support without delays, enhancing the overall learning experience.

Furthermore, GPT-4o's efficiency allows it to be deployed in resource-constrained environments, such as rural schools or institutions in developing countries where access to advanced computing infrastructure may be limited. Educational platforms using GPT-4o can be integrated into mobile devices or low-cost computers, ensuring that students in these regions have access to high-quality tutoring and educational tools. The scalability and cost-effectiveness of GPT-4o make it a practical solution for delivering educational content to large numbers of students, democratizing access to personalized learning across different socioeconomic backgrounds [35].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

4.3.2. Healthcare

In healthcare, both GPT-4 and GPT-4o have shown promise in applications ranging from medical research and diagnostics to clinical decision support. GPT-4's capacity for generating complex, highly accurate medical content makes it particularly useful for creating research papers, analyzing large datasets, or even assisting in the diagnosis of rare or complex diseases [5–8]. Its ability to synthesize information from various medical texts and journals allows healthcare professionals to access comprehensive insights quickly, which is invaluable in fields such as oncology, cardiology, and neurology. For instance, GPT-4 can generate detailed summaries of recent research, analyze clinical trial data, or assist in generating personalized treatment plans for patients based on the latest evidence.

While GPT-4 excels in these more research-oriented tasks, GPT-4o is better suited for real-time clinical decision support systems (CDSS) due to its lower latency and faster response times. CDSS tools powered by GPT-4o can provide healthcare professionals with immediate, actionable insights during patient interactions, helping them make informed decisions on diagnoses, treatments, or prescriptions without having to wait for complex computations to complete. For example, in emergency settings, where quick decisions can save lives, GPT-4o can assist doctors by analyzing patient data in real time and suggesting potential diagnoses or treatment options based on symptoms and medical history [36]. This makes GPT-4o an invaluable tool in telemedicine platforms, where real-time interaction between patients and healthcare providers is crucial for timely and accurate medical care.

Additionally, GPT-4o's lower computational demands make it ideal for deployment in mobile health (mHealth) applications, where it can power chatbots that provide patients with healthcare advice, reminders for medication adherence, or even mental health support. In regions with limited access to healthcare facilities, GPT-4o can be used to build AI-driven diagnostic tools that operate on mobile devices, helping to bridge the gap in healthcare delivery by providing real-time support and information to patients in remote or underserved areas [39]. This ability to scale healthcare support while minimizing costs and resource requirements has the potential to transform how medical care is delivered globally.

4.3.3. Customer Service

Both GPT-4 and GPT-4o have found extensive applications in customer service, but GPT-4o's efficiency and scalability make it particularly valuable for businesses that need to provide fast, reliable customer support at scale. While GPT-4 may be better suited for generating detailed responses or handling complex customer interactions that require a higher degree of contextual understanding, GPT-4o is more appropriate for day-to-day customer service tasks, such as answering frequently asked questions (FAQs), processing simple requests, and guiding users through troubleshooting steps.

Chatbots and virtual assistants powered by GPT-4o can handle a large number of customer queries simultaneously, reducing wait times and ensuring that users receive timely responses. For example, in e-commerce, GPT-4o can assist customers in tracking their orders, finding product information, or processing returns, all while operating on cost-effective infrastructure that doesn't require extensive computational resources. This ability to operate efficiently without sacrificing too much accuracy makes GPT-4o a cost-effective solution for businesses looking to improve their customer service offerings without investing heavily in new hardware or cloud services.

Moreover, GPT-4o's real-time capabilities make it ideal for interactive voice response (IVR) systems, where customers can interact with automated systems via phone to resolve issues or get information. IVR systems powered by GPT-4o can deliver accurate, conversational responses quickly, enhancing the user experience by minimizing hold times and ensuring that customer inquiries are resolved without the need for human intervention. This is particularly beneficial for large organizations in industries such as telecommunications, banking, and retail, where customer inquiries can number in the thousands or millions per day [3,4].

Additionally, the use of GPT-4o in customer service reduces operational costs, as businesses no longer need to rely on extensive human customer support teams to handle basic queries. Instead,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

human agents can be reserved for more complex interactions that require deeper knowledge or a more personalized touch, while GPT-4o manages routine tasks with high efficiency. This hybrid approach improves overall customer satisfaction while keeping costs under control, making it an attractive option for businesses of all sizes [36].

In conclusion, while GPT-4 and GPT-4o both offer robust solutions for a variety of industries, GPT-4o's efficiency, scalability, and ability to handle real-time interactions make it particularly valuable in education, healthcare, and customer service. Its lower computational demands and faster response times allow organizations to deploy AI-powered solutions without the need for extensive hardware investments, making it a cost-effective and practical choice for businesses and institutions looking to improve their service offerings while managing operational costs. By leveraging the strengths of GPT-4o, industries ranging from education to healthcare and customer service can provide personalized, real-time support at scale, enhancing user experiences and improving overall outcomes.

5. Limitations and Future Directions

Despite the advancements brought by both GPT-4 and GPT-4o, they are not without their limitations, especially when it comes to scalability, accessibility, and certain complex tasks. GPT-4, while highly powerful and capable of achieving state-of-the-art performance across a wide range of natural language processing (NLP) tasks, is computationally expensive to train and deploy. Its massive size, consisting of billions of parameters, necessitates significant hardware infrastructure, including access to high-performance GPUs or TPUs, as well as substantial energy consumption [24]. For smaller organizations, educational institutions, and research labs with limited computational resources, deploying GPT-4 is often financially prohibitive. The cost of training and inference, coupled with the need for large-scale cloud services or in-house computational infrastructure, creates barriers to entry for businesses that could otherwise benefit from cutting-edge language models [23].

Moreover, even organizations with access to sufficient resources may face challenges when using GPT-4 for real-time applications due to latency issues and the large memory footprint required for inference. The complexity of GPT-4 also leads to longer inference times, making it less suitable for time-sensitive environments such as real-time customer service or conversational AI systems, where delays in responses can negatively impact user experience [2]. As a result, while GPT-4 excels in tasks that demand a deep understanding of context, such as legal or technical document analysis, it may not be the best choice for applications where speed and efficiency are paramount.

GPT-4o, on the other hand, addresses some of these challenges by being more computationally efficient, using techniques such as model compression, quantization, and pruning to reduce the number of parameters and memory requirements [14]. However, GPT-4o still has its limitations, particularly in handling tasks that require a very fine-tuned context awareness or long-term coherence in text generation. For instance, while GPT-4o can perform well in conversational AI settings and shorter tasks, it may struggle with more complex dialogues that require the model to maintain coherence and context over extended interactions. Similarly, tasks like creative writing, where nuance and long-term story arcs are important, might expose some of the limitations in GPT-4o's ability to fully capture the intricacies of human language [11].

Furthermore, GPT-4o's optimizations, while improving efficiency, can sometimes come at the cost of a slight reduction in model accuracy and performance. In highly specialized domains, such as medical diagnostics or scientific research, where precision is critical, even small inaccuracies or misinterpretations can lead to significant consequences. For example, in a clinical setting, a conversational agent powered by GPT-4o might provide general medical advice but could fall short when offering specific diagnoses or treatment options that require a deep understanding of medical knowledge and context [5–8]. As such, GPT-4o's efficiency makes it more suitable for general applications, but it may not always be the best option for highly specialized or mission-critical tasks.

Future research could explore even more aggressive optimization strategies that further reduce the computational demands of large language models without compromising the richness of text

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

generation. One potential avenue is the development of sparsity-based models, where only a subset of the model's parameters are active during any given task or query. This approach could lead to models that are both smaller and more efficient, while maintaining the ability to deliver high-quality text generation when needed [40]. Additionally, low-rank approximation techniques, which involve approximating large matrices with smaller, lower-dimensional ones, could further enhance model efficiency by reducing the number of computations required for each layer of the network [41].

Another promising direction for future improvements is the integration of more advanced reinforcement learning (RL) algorithms to enhance the long-term coherence and context retention in conversational agents. Current models, even sophisticated ones like GPT-4, can sometimes lose coherence in extended dialogues, where the system needs to remember context from earlier parts of the conversation. Reinforcement Learning from Human Feedback (RLHF) has already been used to fine-tune models for better alignment with user expectations, but more advanced RL algorithms could enable models to better track long-term dependencies in conversation and maintain a more natural flow [19]. For instance, hierarchical RL could be employed to manage conversational tasks at multiple levels, allowing the model to focus on both short-term responses and long-term conversational goals simultaneously, improving the user experience in multi-turn dialogues [42].

Moreover, there is growing interest in neural-symbolic integration, where neural networks like GPT-4 and GPT-4o are combined with symbolic reasoning systems to enhance their ability to handle tasks requiring logical reasoning and structured knowledge. While neural networks excel at pattern recognition and language generation, they often struggle with tasks that require reasoning over discrete entities, such as solving math problems or executing complex planning tasks. By integrating symbolic systems, future models could benefit from the strengths of both approaches, leading to more powerful AI systems capable of both generating rich, coherent text and reasoning through complex problems with logical consistency [43].

Finally, as sustainability becomes an increasingly important concern in AI research, future developments will likely focus on reducing the environmental impact of training and deploying large-scale models. By improving the energy efficiency of both training and inference processes, researchers can mitigate the significant carbon footprint associated with large language models like GPT-4 and GPT-4o [24]. Methods such as distributed training across more energy-efficient hardware, and advancements in algorithmic efficiency, will be key to making AI more sustainable in the future.

6. Conclusions

Despite their advancements, GPT-4 and GPT-4o each present distinct trade-offs in terms of performance, computational efficiency, and scalability. GPT-4 excels in tasks requiring deep contextual understanding and complex reasoning, making it a preferred choice in specialized domains such as legal analysis, scientific research, and academic content generation. However, its high computational demands pose significant challenges for deployment in real-time applications and resource-limited environments. On the other hand, GPT-4o's optimizations make it a more efficient alternative, particularly for applications that prioritize quick response times and scalability, such as customer service, healthcare, and education. While GPT-4o sacrifices some accuracy in handling more complex tasks, its reduced hardware and energy requirements, along with faster inference times, make it an attractive option for many organizations. Future research should continue to explore optimization strategies, focusing on reducing computational costs without compromising performance, and integrating advanced techniques like reinforcement learning and neural-symbolic reasoning to improve long-term coherence and logical consistency in AI models. Ultimately, the choice between GPT-4 and GPT-4o will depend on the specific needs of the application, balancing the trade-offs between accuracy, efficiency, and cost.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

References

1. Brown, T.B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* **2020**.
2. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.
3. Shahriar, S.; Lund, B.D.; Mannuru, N.R.; Arshad, M.A.; Hayawi, K.; Bevara, R.V.K.; Mannuru, A.; Batool, L. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences* **2024**, *14*, 7782.
4. Kipp, M. From GPT-3.5 to GPT-4. o: A Leap in AI's Medical Exam Performance. *Information* **2024**, *15*, 543.
5. Liu, C.L.; Ho, C.T.; Wu, T.C. Custom GPTs Enhancing Performance and Evidence Compared with GPT-3.5, GPT-4, and GPT-4o? A Study on the Emergency Medicine Specialist Examination. In Proceedings of the Healthcare. MDPI, 2024, Vol. 12, p. 1726.
6. Temsah, M.H.; Jamal, A.; Alhasan, K.; Temsah, A.A.; Malki, K.H. OpenAI o1-Preview vs. ChatGPT in Healthcare: A New Frontier in Medical AI Reasoning. *Cureus* **2024**, *16*.
7. Zhang, J.; Sun, K.; Jagadeesh, A.; Falakaflaki, P.; Kayayan, E.; Tao, G.; Haghighat Ghahfarokhi, M.; Gupta, D.; Gupta, A.; Gupta, V.; et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. *Journal of the American Medical Informatics Association* **2024**, *31*, 1884–1891.
8. Günay, S.; Öztürk, A.; Yiğit, Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists. *The American journal of emergency medicine* **2024**, *84*, 68–73.
9. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
10. Radford, A. Improving language understanding by generative pre-training **2018**.
11. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**.
12. Lei Ba, J.; Kiros, J.R.; Hinton, G.E. Layer normalization. *ArXiv e-prints* **2016**, pp. arXiv–1607.
13. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
14. Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* **2019**.
15. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
16. Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* **2019**.
17. Gite, S.; Rawat, U.; Kumar, S.; Saini, B.; Bhatt, A.; Kotecha, K.; Naik, N. Unfolding Conversational Artificial Intelligence: A Systematic Review of Datasets, Techniques and Challenges in Developments. *Engineered Science* **2024**.
18. Casheekar, A.; Lahiri, A.; Rath, K.; Prabhakar, K.S.; Srinivasan, K. A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review* **2024**, *52*, 100632.
19. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P.F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* **2020**, *33*, 3008–3021.
20. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
21. Frankenreiter, J.; Nyarko, J. Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom ed.) Forthcoming* **2022**.
22. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
23. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green ai. *Communications of the ACM* **2020**, *63*, 54–63.
24. Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.; Rothchild, D.; So, D.; Texier, M.; Dean, J. Carbon emissions and large neural network training. arXiv 2021. *arXiv preprint arXiv:2104.10350* **2014**.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2025 doi:10.20944/preprints202502.0066.v1

25. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. Springer, 2020, pp. 121–137.
26. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* **2019**, *32*.
27. LeCun, Y.; Denker, J.; Solla, S. Optimal brain damage. *Advances in neural information processing systems* **1989**, *2*.
28. Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research* **2018**, *18*, 1–30.
29. Chen, T.; Xu, B.; Zhang, C.; Guestrin, C. Training Deep Nets with Sublinear Memory Cost. *CoRR* **2016**, *abs/1604.06174*.
30. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* **2020**.
31. Han, S.; Pool, J.; Tran, J.; Dally, W. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems* **2015**, *28*.
32. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* **2015**.
33. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* **2020**.
34. Chen, T.; Xu, B.; Zhang, C.; Guestrin, C. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* **2016**.
35. Mittal, V.; Bhushan, B. Accelerated computer vision inference with AI on the edge. In Proceedings of the 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT). IEEE, 2020, pp. 55–60.
36. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
37. OpenAI. Hello GPT-4o, 2025. [Accessed 31-01-2025].
38. Hurst, A.; Lerer, A.; Goucher, A.P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* **2024**.
39. Topol, E. *Deep medicine: how artificial intelligence can make healthcare human again*; Hachette UK, 2019.
40. Gale, T.; Elsen, E.; Hooker, S. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574* **2019**.
41. Mehta, S.; Rangwala, H.; Ramakrishnan, N. Low rank factorization for compact multi-head self-attention. *arXiv preprint arXiv:1912.00835* **2019**.
42. Sutton, R.S. Reinforcement learning: An introduction. *A Bradford Book* **2018**.
43. Besold, T.R.; d'Avila Garcez, A.; Bader, S.; Bowman, H.; Domingos, P.; Hitzler, P.; Kühnberger, K.U.; Lamb, L.C.; Lima, P.M.V.; de Penning, L.; et al. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*; IOS press, 2021; pp. 1–51.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.