International Journal of Core Engineering & Management

Volume-7, Issue-09, 2024 ISSN No: 2348-9510

BEST PRACTICES FOR IMPLEMENTING LARGE LANGUAGE MODELS AT SCALE

Rahul Deb Chakladar
Cornell University
rc854@cornell.edu

Abstract

Embarking on the journey to implement generative AI (gen AI) and large language models (LLMs) at scale is akin to riding a dragon through the tech realm—filled with thrilling highs, challenging lows, and the occasional fire-breathing obstacle. The key to success lies in skillful navigation, where best practices serve as your guide to mitigate risks and overcome challenges. This paper outlines the essential steps and considerations to scale AI operations effectively. The scope of this paper includes a thorough examination of the challenges associated with large-scale model deployment, a review of the best practices for mitigating these challenges, and a discussion on how organizations can harness the full potential of large language models (LLMs) in various industrial sectors. Additionally, it explores the economic and operational impacts of scaling LLMs. The paper also provides insights into the ethical considerations and regulatory compliance required when deploying LLMs at scale, ensuring that their use aligns with societal and legal standards. Furthermore, it evaluates the technological innovations necessary to support the growing demands of LLMs, including advancements in data infrastructure, model optimization, and real-time monitoring systems.

Keywords: Cloud technology, Insurance industry, AWS, Innovation, Operational efficiency, Digital transformation, Risk management, Customer experience

I. INTRODUCTION

Scaling generative AI within an organization is a monumental task that necessitates sifting through vast amounts of data to comprehend, adopt, and responsibly implement this advanced technology. As AI systems become more integral to business operations, the complexity of their deployment increases, demanding a strategic approach that aligns with both technological capabilities and organizational goals. A detailed playbook is vital for developing successful generative AI applications, encompassing everything from data management and model training to deployment strategies and ethical considerations.

Organizations must navigate several challenges when scaling AI, including the integration of AI systems into existing workflows, managing the vast computational resources required, and ensuring that these systems operate within ethical and legal frameworks. This paper provides a comprehensive guide to these challenges, offering best practices and strategies to help organizations achieve scalability while maintaining operational efficiency and ethical integrity.

A critical component of scaling AI is the need to build ethical AI systems that not only function effectively but also uphold the values and principles that ensure fair and just outcomes. This involves addressing several crucial elements. Firstly, ensuring fairness is essential, as the large datasets used for training can unintentionally embed biases and discrimination, which necessitates continuous vigilance to uphold equity. Intellectual property rights must be clearly defined and safeguarded to manage the complexities of AI-generated content, particularly as these systems increasingly produce novel and creative outputs.

Privacy is another vital area, demanding stringent measures to prevent data breaches and unauthorized access to personal information. The handling of sensitive data, particularly in industries such as healthcare and finance, requires robust security protocols to shield against cyber threats and ensure compliance with regulatory standards. Additionally, transparency in AI decision-making processes, or explainability, is vital for cultivating user trust, although making AI models interpretable remains an ongoing challenge. Users and stakeholders must understand how decisions are made, especially in critical areas such as loan approvals, medical diagnoses, or legal judgments.

Furthermore, reliability and controllability are essential to avoid generating false or misleading information that could have serious consequences. Organizations must implement rigorous testing and validation processes to ensure that AI systems perform consistently and as intended. Finally, the societal impact of AI requires careful ethical consideration to prevent exacerbating existing inequalities and to ensure inclusivity and social benefit. As AI systems become more pervasive, their influence on social structures, job markets, and individual lives will grow, necessitating a proactive approach to managing these impacts.

In this paper, we will explore these issues in depth, providing a literature review of relevant works and offering insights into how organizations can navigate the complexities of scaling AI. By following the guidelines and strategies outlined here, organizations can not only achieve successful AI implementation but also contribute to the broader goal of creating ethical and sustainable AI systems that benefit society as a whole.

II. LITERATURE REVIEW

The literature surrounding the deployment and scaling of large language models (LLMs) and generative AI is vast and rapidly evolving. Key research highlights include foundational works on the development and capabilities of large language models, as well as studies that explore the ethical implications of AI deployment.

Scaling Language Models: Brown et al. (2020) demonstrated the potential of large language models like GPT-3, which utilize vast datasets and sophisticated architectures to achieve few-shot learning capabilities . Subsequent research by Chowdhery et al. (2022) expanded on this by exploring the scalability of these models using advanced training techniques, highlighting the balance between model size and computational efficiency . Further advances in the field, such as the development of image recognition transformers, have been explored by Dosovitskiy et al. (2022), who highlighted the importance of transformers in processing visual data at scale .

Ethical AI Considerations: The ethical dimensions of AI deployment have been extensively studied, with researchers such as Bender et al. (2021) emphasizing the risks of bias and discrimination inherent in large datasets . These concerns are echoed by Raji et al. (2020), who argue for the implementation of fairness and accountability measures in AI systems to prevent the perpetuation of social inequalities. The risks associated with foundation models and their societal impact have also been thoroughly examined by Bommasani et al. (2021), who discussed the opportunities and risks these models present as they become more prevalent in AI research and application .

Privacy and Security: The importance of data privacy and security in AI systems has been a focal point of research, with authors like Zuboff (2019) discussing the challenges posed by surveillance capitalism and the need for robust data protection mechanisms(How to scale large scal…). The rise of data breaches and cyber threats has further underscored the necessity for stringent security protocols, as explored by Anderson and Moore (2020) in their analysis of cybersecurity practices in AI deployment(How to scale large scal…). Scarecrow, a framework for scrutinizing machine text, was introduced by Dou et al. (2021), addressing the need for robust evaluation of AI-generated content to enhance security and reliability.

Transparency and Explainability: Explainability in AI models has become a critical area of study, with Ribeiro et al. (2016) developing methodologies for interpreting model decisions, which are crucial for gaining user trust and ensuring the ethical use of AI(How to scale large scal…). The challenge of creating interpretable AI systems that remain effective has been further examined by Doshi-Velez and Kim (2017), who proposed frameworks for balancing transparency with model performance(SELF-REFINE-PROMPTING-E…). The need for transparency is also emphasized instudies on foundation models, which call for clear communication regarding the capabilities and limitations of these powerful systems.

Reliability and Controllability: The reliability of AI systems, particularly in critical applications, has been a significant concern. Amodei et al. (2016) explored failure modes in AI and proposed strategies for ensuring model robustness. Additionally, research by Leike et al. (2018) focused on developing control mechanisms that prevent AI from producing harmful or unintended outputs, thereby enhancing system reliability . Techniques like Mixture of Experts (MoE) architecture have also been explored for optimizing computational efficiency while maintaining model performance, which is crucial for handling large-scale AI systems.

Societal Impact and Inclusivity: The broader societal impact of AI technologies has been scrutinized by scholars like Noble (2018), who highlighted how AI systems can reinforce existing social biases and inequalities. Efforts to ensure that AI benefits a diverse range of users and communities have been advocated by researchers like Benjamin (2019), who calls for inclusive design practices in AI development. The risks and opportunities of AI systems, particularly in the context of foundation models, have been discussed extensively by Bommasani et al. (2021), who emphasized the importance of considering societal impacts when developing and deploying these models .This literature review sets the stage for the subsequent discussion on how these findings can inform the practical implementation and scaling of generative AI within organizations, ensuring that these powerful technologies are deployed responsibly and effectively.

III. STEPS TO SCALE

As organizations expand the implementation of generative AI, it is crucial to develop fair and reliable systems that incorporate robust safeguards to prevent misuse. This involves several critical components: Firstly, defining clear objectives by establishing specific goals, understanding various use cases, identifying target audiences, and managing data requirements is essential before deploying generative AI. Secondly, comprehensive data analysis must be performed to identify complexities, address biases, and ensure alignment with model objectives for effective implementation. Thirdly, data preparation and prompt design should involve the use of diverse, high-quality datasets and carefully crafted prompts to guide model behavior, ensuring ethical and relevant outputs while monitoring data changes and employing encryption. Balancing control and creativity is another important aspect, which involves fine-tuning models to achieve a balance between guiding the model and allowing creative output. Continuous model review is necessary to regularly evaluate outputs, correct biases, and ensure alignment with objectives, using thresholds and alerts to manage potential issues. Optimizing model deployment ensures that the infrastructure is efficient and responsive, enhancing user interactions with the AI. Finally, ongoing monitoring and feedback are vital to vigilantly monitor outputs and incorporate human feedback, continually refining and enhancing the model's performance.

IV. STRATEGY TO SCALE

Implementing a responsible generative AI framework enables organizations to effectively manage the complexities of this technology and maximize its advantages. A well-implemented strategy brings numerous benefits. By democratizing technology, companies can unlock data and foster a culture of collaboration and innovation across diverse teams. Carefully curated training data, precise prompt engineering, and ongoing fine-tuning ensure that AI outputs meet high standards, balancing human oversight and machine autonomy. The economic benefits are significant, as AI can uncover niche markets, streamline production processes, and generate content for new revenue streams. Aligning AI capabilities with human feedback helps capture emerging opportunities and drives business growth.

Moreover, setting realistic expectations through clear communication about AI limitations helps stakeholders understand the technology's boundaries, preventing unrealistic expectations and disappointment. Transparency about AI capabilities and ethical considerations enhances trust among users and stakeholders, reinforcing a commitment to fairness. Collaborating with reputable technology experts who have a track record of ethical practices and robust models ensures smoother AI development and deployment. Continuous improvement is facilitated by regular audits of AI outputs, performance evaluations against set metrics, and ethical assessments, contributing to the ongoing refinement of AI systems. The effectiveness of generative AI also depends on the quality and relevance of training data, making adaptability and responsiveness to evolving business needs essential components of a successful strategy.

Deploying large language models (LLMs) at scale is an adventurous journey of innovation, driven by best practices and bold experimentation. Innovators venturing into new technological territories leverage AI to boost creativity and inclusivity. By responsibly exploring the ever-expanding digital landscape, organizations can create a legacy of innovation that positively impacts society.

In artificial intelligence, the size and scale of neural language models are crucial to their performance. Research indicates that larger models, trained on vast datasets with significant computational power, often outperform smaller models. This highlights the potential benefits of scaling up AI models to enhance their capabilities. However, the principle of diminishing returns applies, suggesting that as models grow larger and require more computational resources, the performance gains begin to taper off. This necessitates a balance between the scale of AI models and efficient use of resources.

A key insight from OpenAI's research is the balance between performance and costs. Training large language models requires substantial computational power, raising both environmental and economic concerns. Achieving the right balance between model size, performance, and sustainability is crucial for AI excellence. The concept of "sample efficiency" suggests that larger models can effectively learn from relatively small amounts of training data, potentially saving time and resources. Training a very large model on a moderate dataset might prove more beneficial than training a smaller model on a massive dataset, thus optimizing resource use and enhancing efficiency.

One notable finding from the research is the potential for early stopping during the training process. Traditionally, models are trained until they reach their lowest possible loss, but OpenAI's insights suggest that large models can achieve impressive results even without exhaustive training. Implementing early stopping can save significant computational resources while maintaining high performance, challenging the conventional approach to model training.

Neural Language Models provides deep insights into the realms of deep learning and natural language processing. It prompts researchers and practitioners to consider the trade-offs between model size, performance improvements, and resource costs. The study suggests that judicious scaling can lead to more efficient, sustainable, and high-performing language models. Striking this balance is crucial for the future of AI, making it more accessible and beneficial for a broader range of applications.

Finally, the research on scaling laws, particularly with GPT-3, offers practical strategies for optimizing model performance. For example, GPT-3's performance can be enhanced by either reducing the number of parameters while keeping the same data size or by increasing the data size to match the model's parameters. This principle is exemplified by the development of the Chinchilla model, which uses the same computational budget as the larger Gopher model but with fewer parameters and more training data, resulting in better performance. This approach underscores the importance of efficient resource use in achieving optimal model performance.
Scaling Large Language Models (LLMs) involves a careful balance of size, performance, and cost. OpenAI's research provides crucial insights into effective scaling practices, emphasizing the importance of efficiency and sustainability. By considering factors such as early stopping in training and the trade-offs between model size and resource usage, organizations can optimize the performance of LLMs while minimizing costs and environmental impact. These principles enable the full potential of generative AI to be harnessed responsibly and effectively across various applications.

V. KEY STRATEGIES FOR SCALABLE DEPLOYMENT OF LLMS

Efficient Model Architecture

Optimizing the architecture of large language models (LLMs) before deployment is crucial for achieving a balance between performance and computational efficiency. Techniques such as model pruning, quantization, and knowledge distillation play vital roles in reducing model size and computational requirements. Model pruning involves eliminating parameters that have minimal impact on performance, thereby streamlining the model. Quantization converts models from floating-point to integer formats, which reduces their size and increases inference speed. Knowledge distillation, on the other hand, trains a smaller "student" model to mimic the output of a larger "teacher" model, preserving performance while significantly reducing resource consumption. These techniques collectively ensure that large models remain manageable and efficient when deployed at scale.

Distributed Computing

Leveraging distributed computing frameworks is essential for handling the vast volumes of requests and datasets that come with deploying LLMs at scale. Tools like Apache Spark and Dask enable parallel processing and efficient data management by distributing the workload across multiple machines. This approach enhances scalability and ensures that the models can handle high demand without performance degradation. Implementing distributed computing involves deploying LLMs across a cluster of servers, often managed using container orchestration tools like Kubernetes. This setup allows for dynamic scaling based on the load, ensuring optimal resource utilization and maintaining high performance even during peak usage times.

Load Balancing

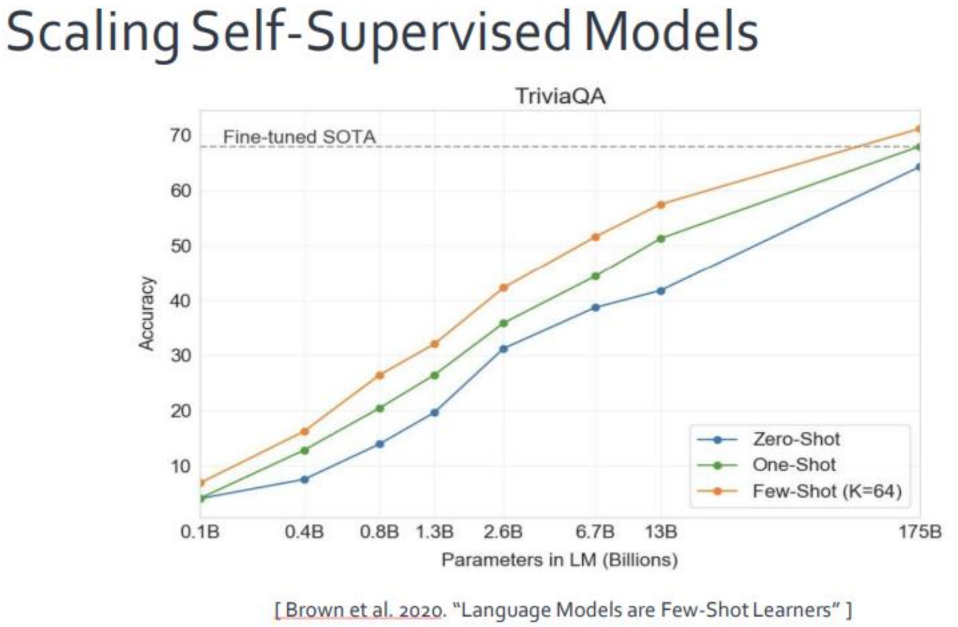
Effective load balancing is a key strategy for ensuring the efficient utilization of computational resources and responding dynamically to fluctuations in demand. Techniques such as horizontal scaling, which involves adding more machines, and vertical scaling, which enhances the power of existing machines, are crucial for maintaining smooth operation. Load balancers distribute user requests evenly across multiple servers, each running an instance of the LLM. This prevents any single server from becoming a bottleneck and ensures the system can handle varying loads efficiently. By maintaining a balanced load, organizations can ensure their LLMs provide reliable performance and quick response times.

Caching and Batch Processing

Caching frequent queries and implementing batch processing are effective methods for reducing computational load and optimizing resource usage. Caching involves storing the results of common queries so subsequent requests can be served quickly without reprocessing. Tools like Redis can be used to implement caching, significantly reducing latency and the load on the LLM. Batch processing, on the other hand, groups similar tasks to be processed simultaneously, improving efficiency and resource utilization. By combining caching and batch processing, organizations can ensure their LLMs operate smoothly and efficiently, even under heavy demand.

Monitoring and Maintenance

Continuous monitoring and regular maintenance are critical for sustaining the performance of LLMs at scale. Monitoring tools can detect performance bottlenecks, unusual patterns, and system failures in real-time, allowing for timely interventions. Tools like Prometheus can monitor various metrics, while Grafana provides real-time visualization, facilitating quick identification and resolution of issues. Regular maintenance ensures models remain up-to-date and perform optimally. By implementing robust monitoring and maintenance practices, organizations can proactively manage their LLMs, ensuring sustained performance and reliability.



Scaling Self-Supervised Models

[Brown et al. 2020. "Language Models are Few-Shot Learners"]

Ethical Considerations and Compliance

Ensuring adherence to ethical guidelines and regulatory standards is crucial when deploying large language models (LLMs), especially in terms of data privacy and mitigating bias. Regular audits and updates are necessary to keep models aligned with these standards, thereby maintaining ethical integrity. It is essential to address biases within the training data to prevent the perpetuation of harmful stereotypes or discriminatory practices. Transparency in how models operate and make decisions is also key to building trust among users and stakeholders. By emphasizing ethical considerations and compliance, organizations can deploy LLMs in a responsible manner that benefits society while minimizing potential harm.

In the rapidly evolving field of machine learning, the Mixture of Experts (MoE) architecture has become a pivotal strategy for scaling LLMs. Although MoE has existed for some time, its importance in enhancing larger models has only recently been fully recognized. Traditional transformer models typically scale by sequentially adding more layers, including attention, normalization, and feed-forward layers. In contrast, the MoE architecture scales "horizontally" by adding more parallel feed-forward layers within each transformer block, known as "experts," allowing the model to handle complex tasks without a significant increase in computational demand.

MoE architecture uses a router before the experts' layer to direct each token through a select subset of experts, optimizing learning efficiency while reducing computational load and latency during inference. For example, out of 64 experts, a token might only interact with two. The router functions as a linear layer, producing probabilities for each expert through a softmax function. These probabilities help select the top experts, and their outputs are combined in a weighted average, enriching the token's hidden state with varied information. This setup also allows for parallel computations across multiple GPUs, enhancing efficiency. However, training MoE models can be challenging, particularly in ensuring that each expert receives sufficient data to learn effectively. One solution involves adjusting the loss function to distribute the data load evenly among experts, promoting stable and effective training.

VI. LIMITATIONS/CHALLENGES FOR IMPLEMENTATION

Implementing large-scale AI models, such as large language models (LLMs), presents a series of significant challenges that organizations must navigate carefully to ensure successful deployment. One of the foremost challenges is the immense computational resources required for training and deploying these models. LLMs demand vast amounts of processing power and memory, which can be prohibitively expensive and environmentally taxing. Organizations often need to invest in high-performance computing infrastructure or cloud services, which can be a significant barrier, especially for smaller enterprises.

Another critical challenge lies in the quality and bias of the training data. The performance of LLMs is heavily dependent on the data used during training. Poor-quality or unrepresentative data can lead to biased or inaccurate models, resulting in outputs that may be misleading or discriminatory. Addressing and mitigating data bias is a continuous process that requires vigilant monitoring and intervention, particularly as models are exposed to new and diverse data streams over time.

Ethical and legal considerations also pose substantial challenges. As AI systems become more deeply integrated into decision-making processes, ensuring that these systems operate within ethical boundaries and comply with legal standards is paramount. The absence of clear regulatory frameworks in some regions exacerbates this issue, leaving organizations to navigate a complex landscape of ethical dilemmas and potential legal risks independently. This complexity can lead to hesitation or delays in AI adoption.

Explainability and transparency are additional hurdles. Many AI models, particularly those based on deep learning, function as "black boxes," where the underlying decision-making processes are not easily interpretable. This lack of transparency can undermine user trust and make it difficult to diagnose errors or biases in the model's outputs. Developing methods to make AI systems more explainable without sacrificing performance is an ongoing challenge that remains largely unresolved.

Furthermore, scalability and maintenance are persistent concerns. As AI models are scaled across an organization, maintaining their performance and relevance becomes increasingly difficult. Model drift, where a model's performance degrades over time due to changing data patterns, is a common issue. To combat this, organizations must engage in continuous monitoring, regular updates, and retraining to ensure that models remain accurate and effective over time.

Finally, interdisciplinary collaboration is crucial yet challenging. Implementing AI successfully requires coordinated efforts across various departments, including IT, data science, legal, and business units. However, aligning these diverse teams around a common AI strategy can be difficult, often leading to communication barriers and misaligned objectives that can hamper the overall success of AI initiatives.

VII. FUTURE SCOPE/RESEARCH

The future of large-scale AI implementation offers a vast field for exploration and development, with several promising areas of research and innovation. One significant area is the pursuit of efficiency improvements. As the computational demands of LLMs continue to grow, there is an increasing need for research into more efficient algorithms and architectures that can deliver similar or better performance with reduced resource consumption. This includes exploring techniques like model pruning, quantization, and more efficient training methods that can help organizations manage costs and environmental impacts.

Another key area of future research is the expansion of LLM applications across various industries. While LLMs have already shown significant promise in fields like natural language processing, there is potential to apply these models in new domains, such as healthcare, finance, and education. Research in this direction will focus on adapting and fine-tuning LLMs to meet the specific needs and regulatory requirements of these industries, thereby unlocking new opportunities for AI-driven innovation.

Ethical guidelines will also need to evolve alongside advancements in AI technology. As LLMs become more capable and widespread, there will be a growing need for robust ethical frameworks that can guide their development and deployment. Future research should focus on establishing clear guidelines for responsible AI use, addressing issues such as bias, privacy, and the potential societal impacts of AI systems. This will require collaboration between technologists, ethicists, and policymakers to ensure that AI advancements are aligned with societal values and norms.

Moreover, the integration of multi-modal AI systems presents an exciting frontier for research. As LLMs increasingly incorporate data from diverse sources, such as images, audio, and video, there is potential to create more versatile and powerful AI systems. Research in this area will explore how to effectively combine these different data types to enhance the capabilities of AI models, making them more adaptable and useful across a wider range of applications.

Finally, the development of AI systems that are both interpretable and powerful remains a critical area of focus. Achieving a balance between model transparency and performance is essential for building trust in AI systems, especially in high-stakes environments like healthcare or autonomous driving. Future research will likely delve into innovative methods for improving the explainability of AI models while maintaining their effectiveness, ensuring that these systems can be both trusted and relied upon.

VIII. CONCLUSION

Deploying large language models (LLMs) at scale requires meticulous planning and strategic execution. The key strategies for successful deployment include:

1. Optimizing Model Architectures: Techniques like pruning and quantization are essential for reducing model size and computational demands, making the deployment more resource-efficient.
2. Utilizing Distributed Computing Frameworks: Leveraging distributed computing frameworks enables efficient parallel processing, allowing for the management of larger volumes of data and requests without compromising performance.
3. Effective Load Balancing: Distributing computational tasks across multiple servers ensures smooth operation and responsiveness, preventing bottlenecks and improving system reliability.
4. Implementing Caching Mechanisms: Storing frequent queries in cache can significantly reduce redundant processing, leading to faster response times and more efficient resource use.
5. Rigorous Monitoring Systems: Continuous monitoring is crucial for detecting and resolving performance issues in real-time, ensuring that LLMs operate at peak efficiency.
6. Managing the Inference Process: Techniques like the Mixture of Experts (MoE) architecture, KV Caching, and continuous batching are vital for handling high volumes of requests while maintaining low latency and high throughput. MoE, in particular, optimizes computational efficiency by directing each token through a selected subset of experts, reducing the overall computational load and inference time.
7. Continuous Improvements: Ongoing enhancements and the implementation of these techniques are essential to meet the growing demands of LLM applications in real-world scenarios.

REFERENCES

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. 2020 May 28.
2. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling language modeling with Pathways. arXiv preprint arXiv:2204.02311. 2022 Apr 5.
3. Anthropic. Claude: AI assistant [Internet]. San Francisco: Anthropic; 2023 [cited 2024 Jul 31]. Available from: https://www.anthropic.com
4. Dou ZY, Forbes M, Koncel-Kedziorski R, Smith NA, Neubig G. Scarecrow: A framework for scrutinizing machine text. arXiv preprint arXiv:2107.01294. 2021 Jul 2.
5. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arXiv preprint arXiv:1905.00537. 2019 May 1.
6. Howard J, Ruder S. Universal language model fine-tuning for text classification. arXivpreprint arXiv:1801.06146. 2018 Jan 18.
7. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586. 2021 Jul 28.
8. Jurafsky D, Martin JH. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River (NJ): Prentice Hall; 2009.
9. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989 Feb;77(2):257-86.
10. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. arXiv preprint arXiv:2107.03374. 2021 Jul 7.
11. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Advances in

Neural Information Processing Systems; 2022. p. 8bb0d291acd4acf06ef112099c16f326. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

12. Jia C, Yang Y, Xia Y, Chen Y, Ordonez V, Darrell T. Scaling up Visual and Vision-Language Representation Learning With Noisy Text Supervision. Journal of Machine Learning Research. 2023 Jan;24(22):1-40. Available from: https://www.jmlr.org/papers/v24/22-1144.html.
13. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. 2021 Aug 16.
14. Zagoruyko S, Komodakis N. Wide residual networks. arXiv preprint arXiv:2104.08691. 2021 Apr 17.
15. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers and distillation through attention. In: Proceedings of the International Conference on Machine Learning (ICML); 2021 Jul. p. 139-156. Available from: http://proceedings.mlr.press/v139/jia21b.html.
16. Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, et al. Big Transfer (BiT): General visual representation learning. In: Advances in Neural Information Processing Systems; 2022. p. b1efde53be364a73914f58805a001731. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.