Feasibility and Best Practices for Distillation Training of Large Language Models on Microsoft Azure

Introduction

Distillation training, also known as knowledge distillation, represents a significant paradigm in the optimization of large language models (LLMs). This technique involves the transfer of knowledge from a large, complex model, often termed the "teacher," to a smaller, more efficient model, known as the "student" [1]. The benefits of successful model distillation are manifold, including a reduction in the model's size, leading to faster inference speeds, lower computational costs, and enhanced accessibility across various deployment environments [5]. By effectively transferring the learned representations and predictive capabilities of a larger model, a smaller model can achieve performance levels comparable to its teacher, but with significantly fewer computational resources [1]. Furthermore, some studies suggest that the distillation process can even lead to improvements in the student model's ability to generalize to unseen data [9]. The core motivation behind employing distillation is to enable the deployment of sophisticated AI models in resource-constrained settings, thereby increasing the scalability and cost-effectiveness of AI applications. This trade-off between model size, inference speed, and accuracy forms a central consideration in the application of distillation techniques.

This report aims to investigate the feasibility and establish the best practices for conducting distillation training on three specific large language models: Gemma 3 27B, PH4 14B, and QWQ32B. The intended platform for this training is Microsoft Azure, with a focus on utilizing either Tensor Processing Units (TPUs) or Graphics Processing Units (GPUs) available within the Azure infrastructure. A critical aspect of this investigation is to determine strategies for achieving two distinct training time targets: under 3 hours and approximately 27 hours. It is important to note that this distillation training will occur after a phase of significant prior development, implying that the foundation models or related resources have already undergone substantial refinement or adaptation. This prior work suggests that the distillation efforts are likely targeted towards a specific application or domain, and the existing knowledge base can inform the selection of the teacher model and the overall distillation strategy. The context of prior development also implies that the distillation process is not starting from a completely untrained state, which could influence the time required for knowledge transfer.

Understanding the Target Models

To effectively address the feasibility and best practices for distillation, a thorough understanding of the characteristics of the target models is essential. Each model possesses unique architectural features, training datasets, and inherent properties that will influence the distillation process and its outcomes.

Gemma 3 27B

Gemma 3 27B is a prominent member of the Gemma family of open-source language models developed by Google [10]. This family includes models of varying sizes, with the 27 billion parameter variant representing a highly capable option [11]. The Gemma 3 models, including the 27B version, were trained on a substantial dataset comprising 14 trillion tokens [12]. This extensive training corpus incorporates a diverse range of text and code, including web documents, code repositories, and mathematical texts [12]. Notably, the original training of Gemma 3 models leveraged Google's TPUs [12]. The training methodology for Gemma models involves a combination of techniques, including distillation, reinforcement learning, and model merging, aimed at enhancing performance across various tasks [12]. Gemma 3 27B also supports a significant context window of 128,000 tokens, allowing it to process and generate longer sequences of text [13]. Furthermore, instruction-tuned versions of Gemma 3 are available, indicating their suitability for tasks requiring specific response formats and adherence to instructions [14]. Given that the pre-training of Gemma 3 itself incorporates distillation, it suggests that this model is a viable candidate for further distillation, either as a teacher or a student. The multimodal capabilities of Gemma 3, which include vision and language understanding, might also be relevant depending on the nature of the prior development and the intended application [12].

PH4 14B

PH4 14B belongs to Microsoft's Phi family of small language models (SLMs) [18]. This particular model is a dense decoder-only Transformer architecture with 14 billion parameters [18]. The training of phi-4 involved a large dataset of 9.8 trillion tokens [18]. A key characteristic of the Phi family is its primary focus on English language text [18]. The training process for PH4 14B was conducted on Microsoft Azure, utilizing NVIDIA GPUs, specifically H100s [18]. The model underwent a rigorous enhancement and alignment process, incorporating both supervised fine-tuning and direct preference optimization (DPO) to ensure precise instruction adherence and robust safety measures [18]. PH4 14B has a context length of 16,000 tokens [18]. It is noteworthy that a multimodal version, Phi-4-multimodal-instruct, exists, although with a smaller parameter count of 5.6 billion [23]. The design philosophy behind the Phi family emphasizes efficiency and strong performance in environments with limited computational resources, making PH4 14B a suitable candidate for distillation into

even smaller and faster models. Its training history on Azure GPUs is directly pertinent to the user's query.

QWQ32B

QWQ32B is a 32 billion parameter language model developed by Alibaba Cloud, with a strong emphasis on reasoning capabilities [24]. The model was trained on a comprehensive dataset derived from Alibaba Group's internal historical data up to December 2024, with a particular focus on diverse text and code [27]. A significant aspect of QWQ32B's training is the incorporation of reinforcement learning (RL) techniques [24]. The model supports a long context length of 32,768 tokens, enabling it to handle more extended sequences of information [27]. To enhance efficiency, QWQ32B employs grouped-query attention (GQA) and quantization techniques, which reduce its memory footprint without a significant compromise in accuracy [27]. Notably, QWQ32B is an open-source model with commercially usable weights, making it accessible for various applications and further development [24]. It is primarily a text-only model focused on advanced reasoning and problem-solving tasks [24]. Given its open nature and strong reasoning abilities, QWQ32B presents an interesting opportunity for distillation, potentially to create more specialized reasoning models or to adapt its capabilities to resource-constrained environments. The existence of distilled versions of models with similar architectures, such as DeepSeek-R1 based on Qwen, provides a relevant precedent [28].

Feasibility of Distillation on Microsoft Azure

Microsoft Azure offers a robust and comprehensive suite of AI services and infrastructure that are highly relevant to the task of large language model distillation. These resources provide the necessary tools and computational power to conduct the training effectively and efficiently.

Azure AI Services and Infrastructure

Azure provides a wide array of AI services, with Azure Machine Learning serving as a central platform for managing the entire lifecycle of machine learning projects, including model training, deployment, and tracking [7]. Specifically for model distillation, Azure AI Foundry offers dedicated capabilities, currently accessible through a notebook-based experience [7]. This service simplifies the process of knowledge transfer from a larger teacher model to a smaller student model. Furthermore, Azure enables the utilization of stored completions generated by larger models, such as GPT-4o, to create high-quality fine-tuning datasets specifically for distillation purposes [1]. This integration of tools and services indicates the platform's strong

support for and the inherent feasibility of model distillation within the Azure environment.

Availability and Suitability of TPUs and GPUs on Azure

Azure provides access to both NVIDIA GPUs, including high-performance options like H100s, and Google's TPUs, both of which are well-suited for the computational demands of training large language models [18]. TPUs are specifically designed for the matrix operations that are fundamental to machine learning workloads and can often deliver faster training times and better cost efficiency, particularly for models optimized for frameworks like TensorFlow and JAX [14]. On the other hand, GPUs offer a high degree of versatility and broad support across various deep learning frameworks, along with substantial memory capacity, which can be advantageous for very large models [34]. The availability of both types of hardware on Azure provides flexibility in choosing the optimal infrastructure for the specific requirements of distilling Gemma 3 27B, PH4 14B, and QWQ32B. The decision will likely depend on factors such as the target model's architecture, the scale of the training process, cost considerations, and the desired training duration.

Computational Resources for Distilling Each Model

The process of distillation, while aiming to produce a smaller and more efficient model, still necessitates significant computational resources for training the student model. However, it is generally less resource-intensive than training the teacher model from scratch [9]. The specific resource requirements will be influenced by the size and architecture of the student model, as well as the complexity of the chosen distillation technique. Fine-tuning a smaller model using knowledge transferred through distillation is typically more efficient than undertaking full-scale training [36]. Parameter-efficient fine-tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA), can further reduce the memory footprint during the fine-tuning stage, making the process more manageable [9]. Therefore, while careful resource allocation will be necessary, the distillation of Gemma 3 27B, PH4 14B, and QWQ32B on Azure is computationally feasible.

Initial Assessment of Achieving Target Training Times

Achieving a training time of under 3 hours for the distillation of these relatively large models on Azure will likely be a challenging endeavor. It will likely require the adoption of highly efficient distillation techniques, the utilization of aggressively scaled hardware configurations (potentially involving a large number of TPUs or high-end GPUs), and possibly targeting a student model that is significantly smaller than the teacher. A 27-hour training timeframe appears more attainable and would allow for a

more comprehensive distillation process, potentially with less extreme hardware scaling and a greater focus on maximizing the performance of the distilled model. The "significant prior development" mentioned in the query could play a crucial role in influencing the starting point of the distillation process, potentially shortening the overall time required. However, a detailed strategy tailored to each model and the specific time constraint will be necessary to determine the exact feasibility and required resources.

Best Practices for Distillation Training

To effectively perform distillation training and achieve the desired outcomes, adhering to established best practices is crucial. These practices encompass the selection of appropriate distillation techniques, model-specific considerations, optimization strategies, dataset management, and leveraging the benefits of pre-training.

General Distillation Techniques

Several general distillation techniques are applicable to large language models. Knowledge distillation is a widely used approach that involves training the student model not only on the ground truth labels but also on the probability distributions (soft targets) produced by the teacher model [1]. This provides the student with richer information about the teacher's reasoning and uncertainty. Response-based distillation focuses primarily on the student model learning to mimic the final outputs of the teacher [3]. Feature-based distillation involves transferring knowledge from the intermediate layers of the teacher model to the student, allowing the student to learn more complex representations [5]. Beyond these core techniques, other methods such as data augmentation (creating more diverse training examples), intermediate layer distillation (focusing on knowledge transfer from specific layers), and multi-teacher distillation (learning from multiple teacher models) can further enhance the distillation process [5]. The selection of the most appropriate distillation technique will depend on the specific characteristics of the teacher and student models, the nature of the task, and the desired balance between training time and student model performance.

Model-Specific Best Practices

For each of the target models, there are specific best practices derived from their architecture, training history, and existing research. Gemma 3 models inherently utilize knowledge distillation during their pre-training [10]. Therefore, a logical approach might involve distilling from a larger Gemma model (if available or feasible) to the Gemma 3 27B variant, or further distilling the 27B model to an even smaller student [41]. For PH4 14B, given its strong pre-training on a large dataset and its design for efficiency, a

suitable strategy could be to fine-tune it using knowledge distilled from a larger, potentially more specialized model. Research has also explored distilling PH4 based on reasoning data generated by a larger reasoning model like DeepSeek-R1, suggesting a potential avenue for enhancing its reasoning capabilities [21]. For QWQ32B, which is focused on reasoning, distillation efforts could concentrate on training a smaller model to replicate its step-by-step reasoning process, possibly employing techniques like step-by-step distillation [42]. Given its open-source nature, there is also greater flexibility in experimenting with different student architectures and distillation methodologies.

Strategies for Optimizing Training Time

To meet the specific training time constraints of under 3 hours and approximately 27 hours, several optimization strategies can be employed. Utilizing a smaller student model will generally lead to faster training times due to the reduced number of parameters and computations. Employing efficient PEFT techniques, such as LoRA, can significantly speed up the fine-tuning process by only training a small subset of the model's parameters. Optimizing hyperparameters like batch size and learning rate is also critical for achieving faster convergence. For the more aggressive target of under 3 hours, leveraging distributed training across a large number of high-performance TPUs or GPUs on Azure will likely be essential [16]. Techniques like progressive distillation, which iteratively reduces the number of sampling steps required, can also contribute to faster training [47]. It is important to consider that achieving very short training times might necessitate accepting a greater trade-off in the performance of the distilled model compared to the teacher.

Considerations for Dataset Preparation and Utilization

The preparation and utilization of the dataset play a pivotal role in the success of distillation training [1]. The quality and relevance of the dataset used to train the student model are paramount for effective knowledge transfer. Generating synthetic data using the teacher model to produce outputs for a wide range of inputs can be a highly effective way to create a rich distillation dataset. Data augmentation techniques can be applied to increase the diversity of the training data and improve the student model's ability to generalize [5]. The size of the dataset will directly impact the training time; smaller datasets will result in faster training but might limit the amount of knowledge that can be transferred. Therefore, careful consideration of the dataset size and its characteristics is necessary to balance training time and model performance.

The Role of Pre-training

Pre-training plays a significant role in the efficiency of distillation training. Distillation often leverages pre-trained models as a starting point for both the teacher and the student [9]. The initial pre-training phase equips the models with a strong foundation of general language understanding, which allows the student model to learn task-specific knowledge more rapidly and efficiently during the distillation process [36]. An alternative approach is pre-training distillation, where the knowledge transfer occurs during the initial pre-training phase of the student model [52]. Regardless of the specific approach, leveraging the knowledge acquired during pre-training is a key factor in reducing the overall training time required for successful distillation. Selecting a suitable pre-trained student model with an architecture that aligns well with the teacher model and the target task is a crucial best practice.

Hardware Considerations: TPUs vs. GPUs on Azure

The choice of hardware accelerator, specifically TPUs versus GPUs on Microsoft Azure, is a critical decision that will significantly impact the performance, cost-effectiveness, and overall feasibility of achieving the target training times for distillation.

Comparative Analysis of TPUs and GPUs

TPUs and GPUs offer distinct advantages for large language model distillation. TPUs are custom-designed by Google for machine learning workloads and are highly optimized for TensorFlow and JAX, often exhibiting superior performance in terms of throughput and power efficiency for these frameworks [14]. They typically offer higher memory bandwidth, which is particularly beneficial for the large tensor operations prevalent in LLMs [34]. However, their availability might be more restricted to the Google Cloud ecosystem, although Azure also provides access to them. GPUs, on the other hand, are more versatile and enjoy extensive support across a wider range of deep learning frameworks, including PyTorch, which is widely used in the LLM community [34]. High-end GPUs often possess greater total memory capacity compared to TPUs, which can be a crucial factor when dealing with very large models or complex training setups [34]. The cost-effectiveness of TPUs versus GPUs on Azure can vary depending on the specific instance types, the duration of the training run, and the utilization rates achieved [34]. GPUs are generally more widely available across different cloud providers and regions, offering greater flexibility in terms of accessibility.

Recommendations on Optimal Hardware Choice

Considering the original training hardware used for each target model provides a useful starting point for recommendations. Given that Gemma 3 models were trained

on Google TPUs, utilizing TPUs on Azure for their distillation might offer better performance and efficiency due to the optimized architecture and software stack [12]. For PH4 14B, which was trained on NVIDIA H100 GPUs on Azure, continuing with a similar GPU infrastructure for distillation would likely be the most straightforward and potentially performant approach [18]. For QWQ32B, the choice might be more flexible due to its open-source nature and potential compatibility with various frameworks. The decision would then hinge on the availability and cost of high-performance TPUs versus comparable GPUs on Azure, as well as the preferred deep learning framework for the distillation process. For the aggressive target training time of under 3 hours, it will almost certainly necessitate leveraging a large cluster of either the most powerful TPUs or the highest-end GPUs available on Azure in a well-optimized distributed training setup. The longer 27-hour target might be achievable with a less extreme hardware configuration, potentially using a smaller number of very powerful accelerators or a moderate-sized distributed setup.

Discussion on Scaling Strategies

Azure provides robust infrastructure for scaling training across multiple hardware accelerators. For GPU-based training, Azure supports distributed training using libraries like PyTorch Distributed or DeepSpeed, allowing for efficient parallelization across multiple GPUs [39]. For TPU-based training, Azure offers TPU Virtual Machines (VMs) and larger-scale TPU Pods, which consist of interconnected TPU devices, enabling significant acceleration for large model training [16]. Common strategies for distributed training include data parallelism, where the training data is split across devices, and model parallelism, where the model itself is partitioned. Achieving efficient scaling requires careful attention to factors such as the batch size used, the learning rate schedule, and minimizing communication overhead between the different hardware units. Azure's networking infrastructure is designed to support high-bandwidth, low-latency communication, which is crucial for effective distributed training of large language models.

Achieving Target Training Times

The feasibility of achieving the target training times of under 3 hours and approximately 27 hours will depend on a combination of strategic choices regarding the distillation technique, the size of the student model, the efficiency of the implementation, and the scale of the computational resources utilized on Azure.

Strategies for Achieving Training Times Under 3 Hours

Reaching a distillation training time of under 3 hours for models of this scale will likely

require a highly aggressive approach. This might involve focusing on distilling the knowledge into a significantly smaller student model with fewer parameters, which inherently reduces the computational workload. Employing highly optimized parameter-efficient fine-tuning (PEFT) techniques, such as quantization and low-rank adaptation, can further accelerate the training process. A large-scale distributed training setup utilizing a substantial number of the most powerful TPUs or GPUs available on Azure will almost certainly be necessary to achieve this timeframe. Exploring techniques like progressive distillation, which aims to reduce the number of training steps required for knowledge transfer, could also be beneficial [47]. It is important to acknowledge that achieving such a rapid training time might necessitate accepting a greater trade-off in the final performance of the distilled model compared to its teacher. The priority in this scenario would be speed, potentially at the cost of some accuracy or generalization capability.

Strategies for Achieving Training Times Around 27 Hours

Achieving a training time of approximately 27 hours offers more flexibility in the distillation strategy and the potential to create a higher-performing student model. This timeframe would likely allow for the use of a student model with a larger capacity, potentially closer in size to the teacher model. More comprehensive distillation techniques that focus on transferring a broader range of knowledge, rather than just the final outputs, could be employed. A moderate-scale distributed training setup, or even a smaller number of very powerful GPUs or TPUs, might suffice to meet this target. The goal in this scenario would be to strike a better balance between the training time and the performance of the resulting distilled model, aiming for a student that closely matches the teacher's capabilities in the target domain.

The Trade-offs Between Training Time and Model Performance

There is an inherent trade-off between the time invested in distillation training and the performance that can be expected from the resulting student model. Faster training times often imply a less thorough transfer of knowledge, which can lead to lower accuracy or reduced capabilities in the student model compared to the teacher. Conversely, investing more time in the distillation process allows for a more comprehensive transfer of knowledge, potentially resulting in a student model that more closely mirrors the teacher's performance. However, this comes at the cost of increased computational resources and longer training durations. The optimal point in this trade-off depends heavily on the specific requirements of the application for which the distilled model is intended, as well as the available computational budget and time constraints. Understanding and carefully considering this balance is crucial for making informed decisions about the distillation strategy and the allocation of

resources.

Azure Implementation Details and Best Practices

To effectively conduct distillation training on Microsoft Azure, it is important to leverage the platform's specific services and adhere to best practices for managing the training process, data, and models.

Leveraging Azure AI Foundry

Azure AI Foundry provides a streamlined environment for model distillation, offering a guided notebook experience that simplifies the process [7]. This service supports various task types relevant to language models and provides pre-configured environments that can expedite the setup process. Users can select both the teacher and student models from a catalog of available options. For users seeking a more guided and integrated approach to distillation on Azure, AI Foundry presents a valuable resource.

Utilizing Azure Machine Learning

For a more comprehensive and customizable approach, Azure Machine Learning offers a robust platform for managing the entire distillation workflow [7]. This service allows users to create and manage compute clusters consisting of GPUs or TPUs, providing the necessary computational power for training. Azure ML also offers tools for tracking experiments, monitoring performance metrics during training, and effectively managing the datasets used in the distillation process. Furthermore, it supports a wide range of deep learning frameworks, including PyTorch and TensorFlow, and provides various options for deploying the final distilled models into production environments.

Best Practices for Data Management

Effective data storage, access, and security are paramount when conducting distillation training on Azure. Large datasets should be stored and managed using scalable and cost-effective services such as Azure Blob Storage or Azure Data Lake Storage. Secure access control can be implemented using Azure Active Directory to manage permissions and ensure that only authorized users and services can access the data. For sensitive information such as API keys and credentials, Azure Key Vault provides a secure and centralized management solution [29]. Implementing data encryption both at rest and during transit is a critical security measure to protect the integrity and confidentiality of the training data.

Monitoring and Evaluation Strategies

Thorough monitoring and evaluation are essential to ensure the success of the distillation process. Azure Monitor can be used to track the utilization of computational resources and monitor the progress of the training jobs. Azure Machine Learning's experiment tracking capabilities allow for the logging and visualization of key performance metrics, such as loss and accuracy, throughout the training. For a more in-depth assessment of the distilled model's quality, Azure AI Evaluation provides tools for evaluating model performance on specific tasks, allowing for comparisons against the teacher model and other baseline models [2]. These comprehensive monitoring and evaluation strategies are crucial for identifying potential issues, optimizing the training process, and verifying that the distilled model meets the desired performance and efficiency targets.

Conclusion and Recommendations

The analysis indicates that distillation training of Gemma 3 27B, PH4 14B, and QWQ32B models on Microsoft Azure using TPUs or GPUs is feasible. Azure provides the necessary infrastructure and services to support this process. Achieving the target training times, however, will require careful planning and the application of specific strategies tailored to each model and time constraint.

For achieving training times **under 3 hours**, it is recommended to:

- Target a student model significantly smaller than the teacher.
- Utilize highly optimized PEFT techniques.
- Employ a large-scale distributed training setup with a substantial number of high-end TPUs (for Gemma 3) or GPUs (for PH4 and potentially QWQ32B).
- Consider progressive distillation techniques.
- Be prepared to potentially accept a trade-off in model performance for the sake of speed.

For achieving training times **around 27 hours**, it is recommended to:

- Allow for a larger student model with more capacity.
- Employ more comprehensive distillation techniques for better knowledge transfer.
- Utilize a moderate-scale distributed training setup or a smaller number of very powerful GPUs/TPUs.
- Aim for a better balance between training time and model performance.

The optimal hardware choice will likely align with the original training hardware: TPUs for Gemma 3 and GPUs for PH4. For QWQ32B, the decision will depend on cost and

availability on Azure. Leveraging Azure AI Foundry can simplify the distillation process, while Azure Machine Learning offers a more comprehensive platform for managing the entire workflow. Adhering to best practices for data management, security, monitoring, and evaluation will be crucial for success.

Key considerations include the inherent trade-off between training time and model performance, the importance of high-quality distillation datasets, and the benefits of leveraging pre-trained models. Potential challenges might involve optimizing distributed training for the aggressive 3-hour target and ensuring that the distilled models retain the desired capabilities of their larger counterparts. Further experimentation and fine-tuning will likely be necessary to achieve the optimal balance of training time and model performance for the specific application requirements.

Key Tables:

1. Comparison of Gemma 3 27B, PH4 14B, and QWQ32B

Feature	Gemma 3 27B	PH4 14B	QWQ32B
Number of Parameters	27 Billion	14 Billion	32 Billion
Original Training Hardware	Google TPUs	NVIDIA H100 GPUs on Azure	Alibaba Cloud Infrastructure (Likely GPUs/TPUs)
Context Length	128K Tokens	16K Tokens	32,768 Tokens
Key Training Techniques	Distillation, RL, Model Merging	Supervised Fine-tuning, DPO	Reinforcement Learning, Quantization, GQA
Relevant for Distillation	Multimodal, Instruction-tuned versions exist	Designed for efficiency, Azure GPU trained	Reasoning focused, Open-source

2. Azure Hardware Options for LLM Distillation

Hardware Type	Key Specifications	Availability on Azure	Cost Considerations	Recommended Use Cases
NVIDIA H100 GPU	High Memory Bandwidth, High Compute	Widely Available	Relatively High	PH4 14B, High-performance distillation
Google TPU v5e	Optimized for TensorFlow/JAX, High Bandwidth	Available	Potentially Cost-Effective	Gemma 3 27B, Large-scale parallel training
Google TPU v4/v5p	Very High Performance, Scalable in Pods	Available	Higher Cost	Sub-3 hour training target for all models

3. Distillation Strategies and Expected Training Times

Target Model	Target Training Time	Recommended Distillation Techniques	Suggested Azure Hardware Configuration	Expected Performance Trade-offs
Gemma 3 27B	Under 3 Hours	Aggressive PEFT, Smaller Student	Large TPU v4/v5p Cluster	Potential reduction in accuracy, focus on core capabilities
Gemma 3 27B	Around 27 Hours	Comprehensive KD, Moderate Student	Moderate TPU v5e Cluster	Better accuracy retention, more comprehensive

				knowledge transfer
PH4 14B	Under 3 Hours	Aggressive PEFT, Smaller Student	Large NVIDIA H100 GPU Cluster	Potential reduction in accuracy, focus on core capabilities
PH4 14B	Around 27 Hours	Comprehensive KD, Moderate Student	Moderate NVIDIA H100 GPU Cluster	Better accuracy retention, more comprehensive knowledge transfer
QWQ32B	Under 3 Hours	Step-by-Step Distillation, Smaller Student	Large NVIDIA H100/TPU v4/v5p Cluster	Potential reduction in reasoning complexity
QWQ32B	Around 27 Hours	Step-by-Step Distillation, Moderate Student	Moderate NVIDIA H100/TPU v5e Cluster	Better retention of reasoning capabilities

Works cited

1. Knowledge Distillation: Empowering Efficient AI Models | by Isaac Kargar | Medium, accessed on March 20, 2025, https://kargarisaac.medium.com/knowledge-distillation-empowering-efficient-ai-models-2505b7781045
2. Model Distillation - Humanloop, accessed on March 20, 2025, https://humanloop.com/blog/model-distillation
3. What is Knowledge distillation? | IBM, accessed on March 20, 2025, https://www.ibm.com/think/topics/knowledge-distillation
4. What Is Model Distillation? | Built In, accessed on March 20, 2025, https://builtin.com/artificial-intelligence/model-distillation
5. LLM Distillation Explained: Applications, Implementation & More - DataCamp, accessed on March 20, 2025, https://www.datacamp.com/blog/distillation-llm
6. Everything You Need to Know about Knowledge Distillation - Hugging Face, accessed on March 20, 2025, https://huggingface.co/blog/Kseniase/kd
7. Distillation in Azure AI Foundry portal (preview) - Microsoft Learn, accessed on March 20, 2025, https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/concept-model-distillation

8. AzureML Model Distillation - Code Samples - Microsoft Learn, accessed on March 20, 2025, https://learn.microsoft.com/en-us/samples/azure/azureml-examples/azureml-model-distillation/
9. A Detailed Technical Comparison of Fine-Tuning and Distillation in Large Language Models, accessed on March 20, 2025, https://medium.com/@jsmith0475/a-detailed-technical-comparison-of-fine-tuning-and-distillation-in-large-language-models-cccbe629dcba
10. What Is Google Gemma? | IBM, accessed on March 20, 2025, https://www.ibm.com/think/topics/google-gemma
11. Gemma 2, knowledge distillation, llama-agents, and more ai updates | by Nabil W | Medium, accessed on March 20, 2025, https://medium.com/@nabilw/gemma-2-knowledge-distillation-llama-agents-and-more-ai-updates-2ea4a409c1ba
12. Introducing Gemma 3: The Developer Guide, accessed on March 20, 2025, https://developers.googleblog.com/en/introducing-gemma3/
13. Fine-tune Gemma 3 with Unsloth, accessed on March 20, 2025, https://unsloth.ai/blog/gemma3
14. google / gemma-3-27b-it - NVIDIA API Documentation, accessed on March 20, 2025, https://docs.api.nvidia.com/nim/reference/google-gemma-3-27b-it
15. Papers Explained 329: Gemma 3 - Ritvik Rastogi, accessed on March 20, 2025, https://ritvik19.medium.com/papers-explained-329-gemma-3-153803a2c591
16. google/gemma-3-27b-it - Hugging Face, accessed on March 20, 2025, https://huggingface.co/google/gemma-3-27b-it
17. Google claims Gemma 3 reaches 98% of DeepSeek's accuracy - using only one GPU, accessed on March 20, 2025, https://www.zdnet.com/article/google-claims-gemma-3-reaches-98-of-deepseeks-accuracy-using-only-one-gpu/
18. microsoft/phi-4 - Hugging Face, accessed on March 20, 2025, https://huggingface.co/microsoft/phi-4
19. Phi Open Models - Small Language Models | Microsoft Azure, accessed on March 20, 2025, https://azure.microsoft.com/en-us/products/phi
20. Phi-4 quantization and inference speedup - Microsoft Community Hub, accessed on March 20, 2025, https://techcommunity.microsoft.com/blog/machinelearningblog/phi-4-quantization-and-inference-speedup/4360047
21. Distillation of Phi-4 on DeepSeek R1: SFT and GRPO | Microsoft Community Hub, accessed on March 20, 2025, https://techcommunity.microsoft.com/blog/machinelearningblog/distillation-of-phi-4-on-deepseek-r1-sft-and-grpo/4381697
22. Microsoft Phi-4: The Revolutionary 14B Parameter Language Model | Galaxy.ai, accessed on March 20, 2025, https://galaxy.ai/youtube-summarizer/microsoft-phi-4-the-revolutionary-14b-parameter-language-model-w22WT1bgn5s
23. phi-4-multimodal-instruct Model by Microsoft - NVIDIA NIM APIs, accessed on

March 20, 2025, https://build.nvidia.com/microsoft/phi-4-multimodal-instruct/modelcard
24. QwQ-32B: Features, Access, DeepSeek-R1 Comparison & More | DataCamp, accessed on March 20, 2025, https://www.datacamp.com/blog/qwq-32b
25. Huaxin Securities: Alibaba Cloud QWQ-32B makes its global debut, and the open-source model has entered the phase of commercial value release. - Moomoo, accessed on March 20, 2025, https://www.moomoo.com/news/post/50290638/huaxin-securities-alibaba-cloud-qwq-32b-makes-its-global-debut
26. Alibaba Stock: China Has Low AI Revenue Compared to United States - IO Fund, accessed on March 20, 2025, https://io-fund.com/artificial-intelligence/ai-platforms/alibaba-stock-china-low-ai-revenue-vs-us
27. QwQ-32B - Cobus Greyling - Medium, accessed on March 20, 2025, https://cobusgreyling.medium.com/qwq-32b-7444e66503c7
28. What are DeepSeek-R1 distilled models? | by Mehul Gupta | Data Science in your pocket | Jan, 2025 | Medium, accessed on March 20, 2025, https://medium.com/data-science-in-your-pocket/what-are-deepseek-r1-distilled-models-329629968d5d
29. How to use Azure OpenAI Service stored completions & distillation - Microsoft Learn, accessed on March 20, 2025, https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/stored-completions
30. Introducing Model Distillation in Azure OpenAI Service | Microsoft Community Hub, accessed on March 20, 2025, https://techcommunity.microsoft.com/blog/azure-ai-services-blog/introducing-model-distillation-in-azure-openai-service/4298627
31. Model Distillation. Making AI Models Leaner and Meaner: A... | by Naveen Krishnan | Towards AI, accessed on March 20, 2025, https://pub.towardsai.net/model-distillation-41ccb09eb312
32. Large Language Models — the hardware connection | APNIC Blog, accessed on March 20, 2025, https://blog.apnic.net/2023/08/10/large-language-models-the-hardware-connection/
33. Azure sets a scale record in large language model training | Microsoft Azure Blog, accessed on March 20, 2025, https://azure.microsoft.com/en-us/blog/azure-sets-a-scale-record-in-large-language-model-training/
34. GPU vs TPU for LLM Training: A Comprehensive Analysis - Incubity by Ambilio, accessed on March 20, 2025, https://incubity.ambilio.com/gpu-vs-tpu-for-llm-training-a-comprehensive-analysis/
35. Understanding TPUs vs GPUs in AI: A Comprehensive Guide | DataCamp, accessed on March 20, 2025, https://www.datacamp.com/blog/tpu-vs-gpu-ai
36. Fine-Tuning Large Language Models: A Comprehensive Guide - Analytics Vidhya,

accessed on March 20, 2025, https://www.analyticsvidhya.com/blog/2023/08/fine-tuning-large-language-models/
37. Transfer Learning in NLP - GeeksforGeeks, accessed on March 20, 2025, https://www.geeksforgeeks.org/transfer-learning-in-nlp/
38. Fine Tune Large Language Model (LLM) on a Custom Dataset with QLoRA | by Suman Das, accessed on March 20, 2025, https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07
39. Step-By-Step Guide to Effective LLM Distillation for Scalable AI - Lamatic Labs, accessed on March 20, 2025, https://blog.lamatic.ai/guides/llm-distillation/
40. Distillation: Turning Smaller Models into High-Performance, Cost-Effective Solutions, accessed on March 20, 2025, https://techcommunity.microsoft.com/blog/aiplatformblog/distillation-turning-smaller-models-into-high-performance-cost-effective-solutio/4355029
41. Syed-Hasan-8503/Gemma-2-2b-it-distilled - Hugging Face, accessed on March 20, 2025, https://huggingface.co/Syed-Hasan-8503/Gemma-2-2b-it-distilled
42. Towards Widening The Distillation Bottleneck for Reasoning Models - arXiv, accessed on March 20, 2025, https://arxiv.org/html/2503.01461v1
43. kz919/QwQ-0.5B-Distilled-SFT - Hugging Face, accessed on March 20, 2025, https://huggingface.co/kz919/QwQ-0.5B-Distilled-SFT
44. kz919/QwQ-0.5B-Distilled - Hugging Face, accessed on March 20, 2025, https://huggingface.co/kz919/QwQ-0.5B-Distilled
45. LLM Distillation: The Key to Efficient AI Models | by Piyush Kashyap | Feb, 2025 | Medium, accessed on March 20, 2025, https://medium.com/@piyushkashyap045/llm-distillation-the-key-to-efficient-ai-models-cb4026a655bf
46. the world's largest distributed LLM training job on TPU v5e | Google Cloud Blog, accessed on March 20, 2025, https://cloud.google.com/blog/products/compute/the-worlds-largest-distributed-llm-training-job-on-tpu-v5e
47. The paradox of diffusion distillation - Sander Dieleman, accessed on March 20, 2025, https://sander.ai/2024/02/28/paradox.html
48. Transformers and Transfer Learning: Leveraging Pre-Trained Models for Quick Wins | by Hassaan Idrees | Medium, accessed on March 20, 2025, https://medium.com/@hassaanidrees7/transformers-and-transfer-learning-leveraging-pre-trained-models-for-quick-wins-99eee633948b
49. Few-shot Learning Text Generation | Restackio, accessed on March 20, 2025, https://www.restack.io/p/few-shot-learning-answer-text-generation-cat-ai
50. Transfer Learning in Natural Language Processing (NLP): A Game ..., accessed on March 20, 2025, https://medium.com/@hassaanidrees7/transfer-learning-in-natural-language-processing-nlp-a-game-changer-for-ai-models-b8739274bb02
51. How to fine-tune a large language model (LLM) | Generative-AI – Weights & Biases - Wandb, accessed on March 20, 2025,

https://wandb.ai/byyoung3/Generative-AI/reports/How-to-fine-tune-a-large-language-model-LLM---VmlldzoxMDU2NTg4Mw
52. Pre-training Distillation for Large Language Models: A Design Space Exploration, accessed on March 20, 2025, https://www.researchgate.net/publication/385140169_Pre-training_Distillation_for_Large_Language_Models_A_Design_Space_Exploration
53. Pre-training Distillation for Large Language Models: A Design Space Exploration - arXiv, accessed on March 20, 2025, https://arxiv.org/html/2410.16215v1
54. MiniPLM: Knowledge Distillation for Pre-training Language Models - OpenReview, accessed on March 20, 2025, https://openreview.net/forum?id=tJHDw8XfeC
55. Maximize Your AI Model Performance: Evaluating Distilled Models with Azure AI Evaluation SDK, accessed on March 20, 2025, https://techcommunity.microsoft.com/blog/aiplatformblog/the-future-of-ai-maximize-your-fine-tuned-model-performance-with-the-new-azure-a/4284292