Advancing Language Understanding: A Review of Challenges and Solutions in Training Large Language Models for Low-Resource Languages

Ege Acar

Received December 31, 2024
Accepted June 12, 2025
Electronic access July 15, 2025

This paper investigates the adaptation of state-of-the-art training methods for Large Language Models (LLMs) to address the specific challenges posed by low-resource languages, languages that lack sufficient digital corpora and linguistic tools. Through a structured literature review, the study examines how current training approaches—primarily designed for high-resource languages—can be modified to improve performance in underrepresented linguistic contexts. Key techniques for adaptations evaluated include data augmentation methods such as back-translation and paraphrasing, cross-lingual transfer learning, and community-driven dataset development. The paper also explores fine-tuning strategies like Low-Rank Adaptation (LoRA), which reduce the number of trainable parameters and make training feasible in constrained environments. In addition to technical methods, the study highlights the significance of addressing data bias, computational accessibility, and ethical considerations in LLM deployment. The findings synthesize both theoretical advancements and practical approaches to provide a structured analysis of current solutions for developing more inclusive, efficient, and culturally aware language technologies. By identifying gaps in current research and consolidating viable solutions, this paper aims to contribute to ongoing efforts to democratize natural language processing and extend its benefits to speakers of low-resource languages worldwide.

Introduction

A Large Language Model (LLM) is defined as a type of deep learning model, typically utilizing architectures like transformers, trained on massive datasets to process and generate human-like text. LLMs possess emergent abilities, such as understanding context and generating coherent responses, which become evident as the scale of model parameters and data increases [1].

The field of Natural Language Processing (NLP) has been fundamentally reshaped by the rapid advancements in Large Language Models (LLMs). These models, representing a culmination of decades of research progressing from statistical language models through neural language models (NLMs) and pre-trained language models (PLMs), exhibit remarkable capabilities in understanding and generating human-like text [2]. Trained on massive datasets often comprising web-scale text corpora, LLMs like GPT-3 [3], LlaMA [3], BLOOM [4], have demonstrated emergent abilities not present in their smaller predecessors. These include in-context learning (learning tasks from examples provided only at inference time), instruction following (generalizing to new tasks based on natural language instructions after specific tuning), and complex reasoning. Such capabilities have positioned LLMs as foundational components for a wide array of applications and potentially general-purpose artificial intelligence agents.

LLMs enabled machines to perform complex linguistic tasks such as translation, summarization, and conversational AI with high accuracy. While these models perform well in high-resource languages such as English, their effectiveness significantly drops when applied to low-resource languages.

Low resource languages, which have limited digital presence and NLP infrastructure, present a major challenge for the global scalability and inclusivity of language technologies. There is a critical imperative to extend the benefits of LLMs beyond the handful of high-resource languages (predominantly English) where development and evaluation have historically been concentrated.

Despite recent advancements, most LLM training methodologies remain data and compute-intensive, making them poorly suited for low-resource language contexts. Prior research has explored solutions such as data augmentation, transfer learning, and multilingual models, but these efforts are often fragmented or not tailored to the specific constraints of underrepresented languages.

Objectives

This paper aims to address the research question: *What are the state-of-the-art methods for training LLMs, and how can these methods be adapted to better support low-resource languages?*

The study focuses on identifying existing training strategies and analyzing how they can be modified to address challenges such as data scarcity, computational limitations, and linguistic diversity. It examines solutions including data augmentation, cross-lingual transfer learning, and community-driven resource development. By offering a structured synthesis of current approaches, the paper seeks to support the development of more inclusive and accessible NLP technologies.

Methodology

This paper employs a structured literature review approach, analyzing peer-reviewed publications, technical reports, and applied examples from recent studies. Emphasis is placed on identifying common patterns, emerging best practices, and practical solutions for training LLMs in low-resource contexts.

Background

Neural Networks and Evolution of LLMs

LLMs are built upon artificial neural networks, which are inspired by the interconnected structure of the human brain. These networks consist of layered units (neurons) that transform input data through mathematical functions, enabling the model to learn complex patterns. Key to this learning process is gradient descent, an optimization algorithm that adjusts the network's internal weights to minimize prediction error. By calculating gradients through backpropagation, the model iteratively refines its predictions, improving performance over time. This architecture—comprising input, hidden, and output layers—combined with modern optimization techniques, underpins the ability of LLMs to perform sophisticated natural language understanding and generation tasks.

Before the application of neural networks, the journey of language models began with statistical methods, primarily n-gram models, which, while foundational, struggled to capture long-range dependencies in language[5]. The shift towards neural networks marked a significant advancement, with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, enabling the processing of sequential data and improved handling of contextual dependencies. Simultaneously, word embeddings like Word2Vec and GloVe revolutionized semantic representation, allowing models to grasp word meanings and relationships in a vector space. From a scale point of view, both RNNs and LSTMs process sequences sequentially, making parallelization during training challenging as the computation at each step depends on the previous step's hidden state.

The pivotal moment arrived with the introduction of the Transformer architecture, which leveraged self-attention mechanisms to drastically enhance language modeling performance[6]. The Transformer architecture innovation became the backbone of many state-of-the-art LLMs, such as GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), LLaMA (Large Language Model Meta AI) and their numerous variants.

In contrast to RNNs and LSTMs, the Transformer architecture enabled large-scale model training by making sequence processing parallel, efficient, and scalable. Its use of self-attention, modular design, and compatibility with GPU-based parallelism allowed researchers to train much larger models on massive datasets, which laid the foundation for LLMs. In their seminal paper, Kaplan et al.[7] demonstrated that the performance of Transformer-based language models improves predictably with increases in model size, dataset size, and compute resources. This empirical observation, now known as the "scaling law," provided strong motivation for training ever-larger models and helped solidify the Transformer as the backbone of modern LLM architectures.

Versatility and Contextual Understanding in LLMs

This diversity of tasks is a key feature of LLMs. They can translate languages, summarize documents, answer questions, and even engage in conversation. Their versatility stems from the vast amount of data they are exposed to during training, which includes examples of these various tasks. The Transformer architecture, which relies on self-attention mechanisms, enables LLMs to weigh the importance of different words in a sentence, allowing them to understand context much more effectively than earlier models such as Convolutional Neural Networks (CNNs), RNNs[8], and LSTMs.

Training Paradigms of LLMs

LLMs function by processing large volumes of text data to learn the statistical properties of language. In this context, a "token" refers to a unit of text, which can be as small as a character, a word, or a subword segment, depending on how the model is designed. They are trained to predict the next token in a sequence given the preceding words, a task known as next-token prediction. Through this process, the models learn not just word associations but also grammar, context, and even some level of reasoning. The larger the model and the more diverse the training data, the better it becomes at understanding and generating text across various contexts.

This learning process underpins the ability of LLMs to perform a surprisingly diverse range of tasks. Understanding these methods is crucial for grasping how LLMs achieve their versatility and effectiveness.

Training LLMs is a complex and resource-intensive process. It typically involves several key stages: pre-training, fine-tuning, and in some cases, continuous learning[9].

Pre-training

This is the first and most computationally intensive stage. During pre-training, the model is exposed to a vast corpus of text, learning to predict the next token in a sentence. The model is trained on massive datasets that include many terabytes of text from books, websites, articles, and other sources. The goal of this stage is to create a model that has a broad understanding of language, even if it is not yet specialized in any specific task.

Pre-training involves optimizing the model's parameters using gradient descent, where the model minimizes the difference between its predictions and actual next tokens. The self-attention mechanism in transformers allows the model to consider the entire context of a sentence, rather than just the preceding words, which marks a significant improvement over earlier architectures such as RNNs and LSTMs.

Fine-tuning

After pre-training, the model is fine-tuned on a smaller, more specific dataset that is tailored to a specific task. Fine-tuning allows the model to adapt its broad linguistic knowledge to applications, such as sentiment analysis, machine translation, or question-answering. This phase requires less data and computational power than pre-training but is crucial for achieving high performance on specific tasks.

During fine-tuning, the model's parameters are adjusted slightly, allowing it to specialize while retaining the general language understanding it acquired during pre-training. Fine-tuning can be supervised, where the model learns from labeled examples, or unsupervised, where it continues to learn from unstructured text.

Transfer Learning and Supervised Fine-tuning

LLMs are often adapted to new tasks using transfer learning, where a pretrained model is fine-tuned on data related to a different but relevant task. This method is highly efficient because it leverages the model's existing linguistic and contextual knowledge. Supervised fine-tuning adjusts the model's weights using labeled data, allowing it to excel at tasks such as sentiment classification or summarization while retaining its foundational language competence from pre-training [10].

Continuous Learning, Reinforcement Learning with Human Feedback (RLHF), and Reward Modeling

In dynamic domains such as social media, customer service, or news tracking, continuous learning is valuable for keeping the model current. This involves periodically updating the model with new data so it can adapt to evolving trends and language usage.

Reinforcement Learning from Human Feedback (RLHF) [11] further enhances alignment with human preferences by incorporating judgments from human evaluators. A core component of RLHF is reward modelling, where the model is trained to optimize responses based on a learned reward function. This function provides positive signals for preferred outputs and negative signals for undesirable ones. Together, continuous learning, RLHF, and reward modelling help ensure the model not only stays up to date but also produces responses that are accurate, safe, and aligned with human values.

Methods

Challenges in Low-Resource Languages

Defining Low-resource Languages

Hedderich et al. [12] defines a language as low-resource when there is a lack of digital data, annotated resources, or computational tools necessary for effective natural language processing. This scarcity may apply not only to endangered or minority languages, but also to widely spoken languages or mainstream domains that have little available labeled or unlabeled data, or lack language technologies such as taggers, parsers, or large corpora. In short, "low-resource" characterizes situations where the limited availability of essential resources poses challenges for developing and evaluating NLP models.

Limited Availability of Digital Texts

One of the primary challenges in training LLMs for low-resource languages is the scarcity of digital text data. High-resource languages like English, Mandarin, or Spanish have vast amounts of text data (and speech) available online, ranging from books and articles to social media posts and technical documentation. In contrast, low-resource languages are often spoken in communities with limited internet infrastructure or in regions where digital engagement and content creation are minimal. Consequently, there is a lack of online content such as websites, social media posts, or digitalized books that could be used to build extensive language corpora. Even when text data exists, it may often be limited to religious scriptures, government documents, or a handful of literary works and it may be inaccessible due to copyright restrictions, insufficient digitization, or language preservation efforts that prioritize oral traditions over written records.

Annotated datasets, where text is labeled with information such as part-of-speech tags, syntactic structures, or translations, are critical for training supervised models. High-resource languages benefit from extensive annotated datasets that have been curated over decades. Low-resource languages, however, often lack such datasets, as the process of creating them is time-

consuming and requires linguistic expertise that may not be readily available.

This scarcity of text data makes it difficult to compile large, diverse corpora necessary for training LLMs in low-resource languages. The absence of comprehensive and diverse datasets means that LLMs have insufficient exposure to the linguistic features, syntactic structures, and semantic subtleties of these languages. LLMs trained on limited data often struggle to achieve high performance, leading to issues such as reduced accuracy, poor robustness, and a lack of understanding of the intricate aspects of low-resource languages. This data deficiency impedes the models' ability to generate contextually relevant responses and perform complex language tasks, thereby exacerbating the technological gap between high-resource and low-resource languages. The lack of substantial training data not only affects the model's learning capabilities but also limits its ability to handle the rich diversity of linguistic expressions and variations present in low-resource languages, ultimately undermining the effectiveness and applicability of NLP technologies in these linguistic contexts.

Uyghur, a Turkic language spoken by over 13 million people, exemplifies the multifaceted challenges of supporting low-resource languages with LLMs. Despite its sizable speaker base, Uyghur suffers from limited digital resources, sparse annotated corpora, and geopolitical constraints that hinder open data collection. In a recent effort, Lu et al.[13] tackled these obstacles by adopting a four-stage strategy: extensive monolingual data collection, continual pre-training on a Chinese-optimized LLaMA2 model, translation-specific instruction fine-tuning, and Direct Preference Optimization based on translation self-evolution (DPOSE). Their approach led to significant improvements in translation quality, even outperforming GPT-4 in specific Uyghur–Chinese tasks.

Despite being spoken by nearly 70 million people, Thai remains under-resourced in key areas of NLP, especially in tasks like semantic parsing and syntactic analysis. While significant advances have been made in tokenization and part-of-speech tagging—bolstered by tools like PyThaiNLP, DeepCut, and AttaCut—critical gaps persist in the development of high-quality syntactic parsers and semantic understanding tools. For instance, most Thai corpora are sourced from formal documents, which fail to capture informal expressions, regional dialects, and social media language. Semantic tools such as named entity recognizers and sentiment analyzers exist but often lack coverage and accuracy across diverse domains. The absence of robust, large-scale, and domain-diverse resources continues to hinder the effective deployment of LLMs for Thai, highlighting the ongoing need for comprehensive tool development and resource expansion[14].

Insufficient Linguistic Resources and Tools

High-resource languages typically have a wealth of linguistic tools, such as parsers, lemmatizers, bilingual dictionaries and named entity recognizers, which facilitate the processing of text data. Low-resource languages may lack these tools, making it challenging to preprocess and analyze text effectively. Additionally, the linguistic diversity within low-resource languages, including dialectal variations and different writing systems, adds another layer of complexity.

Linguistic Complexity and Diversity

Many low-resource languages are linguistically complex, with rich morphologies, multiple dialects, or unique syntactic structures that are challenging for LLMs to learn. For instance, where words are formed by stringing together morphemes, can generate a vast number of word forms from a small set of roots[15]. This complexity requires more data to model effectively, exacerbating the data scarcity problem. Additionally, the diversity within low-resource languages, including dialectal variations, poses further challenges for creating generalized models that can handle all variations effectively[16].

Model Performance and Limitations

LLMs trained on high-resource languages typically exhibit superior performance due to the availability of extensive training data. These models can generate fluent, contextually appropriate text, understand nuanced language structures, and perform well across a variety of NLP tasks. For instance, in the BLOOM LLM training corpus, the top nine languages constitute 95.75% of the data; the remaining 38 languages—including almost all low-resource ones—collectively get only 4.25% of training data, sometimes less than a megabyte per language. This results in very sparse vocabularies and poor understanding/generation for these languages.

Low-resource languages, such as Quechua, and Wolof, present a unique set of challenges in the context of training and deploying LLMs. These languages, often spoken by smaller communities or in regions with less technological infrastructure, lack the extensive datasets and linguistic resources that high-resource languages enjoy[17]. Mid-resource languages, such as Turkish and Indonesian, fall somewhere in between, with moderate amounts of data available but still facing challenges in achieving the same performance levels as high-resource languages.

According to Cahyawijaya et al.[17], the disparity between low-resource and high-resource languages is evident in the performance of LLMs. High-resource languages can leverage few-shot learning to achieve impressive results with minimal additional data, while low-resource languages struggle to reach comparable accuracy due to their limited datasets. The lack of sufficient

training data in low-resource languages often results in models that underperform in capturing the linguistic nuances and complexities unique to these languages.

Generalization and Robustness: High-resource LLMs benefit from training on diverse datasets that cover a wide range of topics, genres, and linguistic variations. This diversity enables the models to generalize well to new, unseen data. On the other hand, LLMs for low-resource languages may overfit the small, specific datasets they are trained on, resulting in poor generalization [18]. These models may also be less robust to variations in input, such as different dialects or informal language usage, which are common in natural text.

Cultural and Contextual Understanding: High-resource LLMs are often trained on datasets that include rich cultural and contextual information, enabling them to understand and generate culturally appropriate content. Low-resource LLMs, due to the lack of such data, may miss critical cultural nuances, leading to outputs that are not only linguistically flawed, but also culturally insensitive or irrelevant.

Economic and Technological Barriers

Developing NLP resources for a language requires significant financial and technological investment. High-resource languages, spoken in economically developed regions, benefit from strong institutional support, funding for research, and technological infrastructure. In contrast, low-resource languages are often spoken in economically disadvantaged areas where funding for linguistic research and technology development is limited. This economic disparity results in fewer resources being allocated to the development of LLMs for these languages.

Limited Research and Technological Support

Research and development in NLP have historically concentrated on a few dominant languages, such as English, Chinese, and Spanish, leading to a disparity in the availability of models and tools for other languages. For instance, high-performance models like BERT and GPT-3 have been extensively developed and fine-tuned primarily for English, with limited resources allocated to languages such as Swahili or Navajo. This focus is reflected in the availability of pre-trained models and benchmarks, which predominantly cover high-resource languages, leaving low-resource languages underrepresented. The uneven distribution of research efforts and technological support underscores the need for a more inclusive approach to NLP, which would involve expanding resources, developing multilingual models, and creating tools that address the needs of speakers of less researched languages.

Enhancing Resources for Low-Resource Languages in LLMs

One of the primary methods to combat the lack of data is through data augmentation and synthesis. This involves creating additional training data from existing resources to expand the dataset for low-resource languages. Techniques such as back-translation, where text is translated to another language and then back to the original, can generate new variations of the same content. This process helps in diversifying the training data, making the model more robust to different linguistic patterns and usages. Another approach is paraphrasing, where the same information is expressed in multiple ways. Synthetic data generation, including using language models to generate plausible text based on limited examples, also contributes to building a larger dataset. These methods can alleviate the data scarcity issue by providing more training examples, which is crucial for improving model performance in low-resource languages [19].

Another effective method involves leveraging pre-trained transformer models for conditional data augmentation. This article by Kumar et al. [20] highlights that, models like GPT-2, BERT, and BART can significantly enhance data diversity and class-label preservation through techniques such as class-label prepending. Their study demonstrates that these models can generate additional training data from limited resources, thereby improving performance in low-resource settings.

These methods not only address the data scarcity issue but also helps in building more effective models for low-resource languages.

Cross-lingual transfer learning

Cross-lingual transfer learning is another effective strategy that leverages the wealth of data available for high-resource languages to benefit low-resource languages. By training multilingual models on a combination of high-resource and low-resource languages, the model can transfer knowledge from the more data-rich languages to those with fewer resources. This approach enables the low-resource languages to benefit from the general linguistic patterns learned from high-resource languages. Transfer learning techniques, such as fine-tuning pre-trained models on low-resource languages, allow these languages to adapt the learned representations to their specific linguistic characteristics. This method not only enhances the model's performance but also helps in overcoming the data imbalance between high-resource and low-resource languages.

Crowdsourcing

Engaging with community involvement and crowdsourcing offers a practical solution to the data limitation problem. By involving native speakers of low-resource languages in the data collection and annotation processes, researchers can obtain more authentic and representative data. Crowdsourcing platforms can

facilitate the collection of text data and annotations from a diverse set of contributors, including those from different dialects and regions. This approach ensures that the collected data reflects the true linguistic diversity and usage of the language. Additionally, community-driven initiatives can lead to the development of language resources that are more culturally relevant and aligned with the needs of the speakers.

One example annotation platform for crowdsourcing is CroAno[21], a web-based crowd annotation platform for the Chinese named entity recognition. The platform provides 3 roles; the crowd annotator, the annotation expert and the algorithm expert for annotating. Respective user with these roles, execute the annotation and conflict resolution thru a streamlined user interface.

All Voices[22], developed by African Languages Lab is a mobile application that enables users work with others to help gather data for low resource languages. Users interact by rating each other's submissions in a gamified environment that encourages participation.

Another example is Malmon[23], an open-source, language-agnostic platform designed to crowdsource parallel corpora for Automatic Text Simplification (ATS), particularly in low-resource languages. It enables contributors to simplify sentences through a user-friendly interface and incorporates validation features to ensure quality.

Collaborative research and open resources

Collaborative research and open resources play a vital role in enhancing the availability of linguistic tools and datasets for low-resource languages. Collaborative efforts between academic institutions, technology companies, and governmental organizations can lead to the creation of shared resources such as linguistic databases, annotated corpora, and open-source language models. These resources can be made publicly available, allowing researchers and developers to build upon existing work and contribute to collective knowledge. Initiatives such as digitizing historical texts and developing open-source tools for text processing can provide a solid foundation for further research and development in low-resource languages.

Educational Investment

Investing in education and training for researchers and developers working with low-resource languages can contribute to long-term improvements. By providing resources, funding, and training opportunities, stakeholders can build expertise in the field of NLP for low-resource languages. This investment can lead to the development of innovative methods and technologies tailored to the specific needs of these languages, further advancing the state of language modeling.

Improving Training Efficiency for Low-resource Languages

Addressing the insufficiencies faced by low-resource languages in the realm of LLMs requires a multifaceted approach that tackles the root causes of data scarcity and technological limitations. To improve the performance and accessibility of LLMs for these languages, several key strategies can be employed, each contributing to a more equitable advancement in natural language processing[24].

Parameter-efficient Fine-Tuning (PEFT)

Fine-tuning LLMs requires extensive computational resources, memory, and time. In response, parameter-efficient fine-tuning (PEFT) methods are developed, which aim to adapt pre-trained models to new tasks or languages by updating only a small subset of their parameters. PEFT techniques reduce the cost and carbon footprint of model adaptation while enabling practitioners to fine-tune large models on limited hardware. These methods are particularly valuable in low-resource settings, where data, compute, and storage are constrained. LoRA, described in detail in the next section, is a prime example of PEFT.

Low-Rank Adaptation (LoRA)

LoRA[25] (Low-Rank Adaptation) is a parameter-efficient fine-tuning method introduced to address the high computational and memory demands of adapting large pre-trained language models. Instead of updating all model weights during fine-tuning, LoRA inserts small, trainable low-rank matrices into the architecture while keeping the original weights frozen. This approach significantly reduces the number of trainable parameters and lowers the memory footprint, enabling efficient adaptation even on limited hardware. Although not originally designed with low-resource languages in mind, LoRA's lightweight design makes it particularly well-suited for such contexts. Its ability to fine-tune large models with minimal compute has made it a popular choice in multilingual and domain-specific NLP settings, including biomedical and legal applications. In the context of low-resource languages, where both data and computational resources are often constrained, LoRA offers a practical and scalable solution for adapting state-of-the-art language models with minimal overhead.

As an example, with LoRA fine-tuning using rank set to 64, the number of trainable parameters for a 70 billion parameter model is reduced down to 131 million parameters ($\sim$0.19% of the original model size)[26].

A performance comparison between LoRA and full-parameter fine-tuning in terms of GPU-hours has been performed[27] on Falcon 7B model and the results are as follows:

These results align with findings reported by Singh et al.[26] who observed approximately a sixfold reduction in training time when comparing full-parameter fine-tuning to LoRA. LoRA

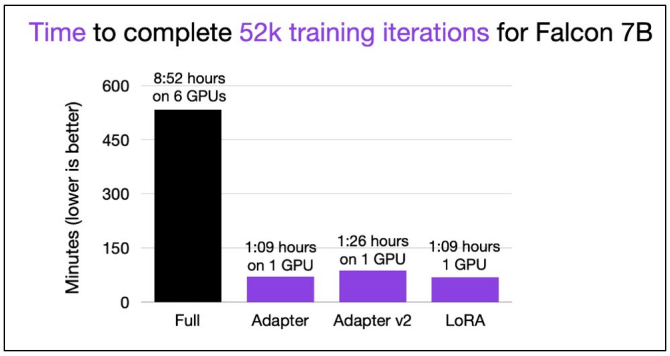

Fig. 1 LoRa vs full-parameter fine-tuning (Reproduced from Raschka[27])

represented a breakthrough in the field of fine-tuning large language models, dramatically reducing memory requirements and improving resource efficiency. Building on its success, numerous researchers have proposed optimizations and extensions (QLoRA, DyLoRA, QDyLoRA etc.) to further enhance its performance and adaptability across diverse tasks and computational environments.

Quantized Low-Rank Adaptation (QLoRA)

An important, emerging strategy for reconciling the trade-off between model performance and computational efficiency in low-resource language settings is the use of quantized parameter-efficient fine-tuning methods, such as QLoRA (Quantized Low-Rank Adaptation). QLoRA[28] demonstrates that it is possible to fine-tune large language models—up to 65 billion parameters—on a single GPU by quantizing the model to 4-bit precision while using LoRA adapters for task adaptation.

Reducing the model's weights from 16-bit (BFloat16/FP16) to 4-bit means each parameter takes only a quarter of the space. For a 65B parameter model, this brings the memory requirement for fine-tuning from over 780GB down to less than 48GB. This innovation drastically reduces the hardware and memory requirements for training, enabling researchers and practitioners with limited resources to fine-tune state-of-the-art models without significant loss in performance. QLoRA achieves parity with standard 16-bit fine-tuning in terms of accuracy on both general and instruction-following tasks, as evidenced by benchmarks where QLoRA-finetuned models rivaled ChatGPT and outperformed other open-source alternatives[28].

DyLoRA

DyLoRA[29] (Dynamic Low-Rank Adaptation) is a recent extension of LoRA that enables models to support a range of low-rank configurations after a single round of training. Traditional LoRA fixes the rank of the low-rank adapters during fine-tuning, which means any change in desired rank (for deployment on hardware with different capabilities, for example) would require retraining. DyLoRA solves this by training the model to support multiple ranks through a dynamic ranking mechanism. This approach allows flexible adaptation during inference and better balances the trade-off between model size, memory, and performance across tasks and hardware platforms.

QDyLoRA

QDyLoRA[30] (Quantized DyLoRA) advances this line of research by integrating quantization strategies (notably 4-bit quantization as in QLoRA) with the dynamic low-rank architecture of DyLoRA. The result is a fine-tuning approach that is both highly memory-efficient (enabling fine-tuning of very large models, e.g., Falcon-40B, on modest GPUs) and dynamically configurable. QDyLoRA allows a single model to be fine-tuned for a range of LoRA ranks in a single training run, letting users select the optimal trade-off between efficiency and performance post hoc, without additional retraining. Experiments demonstrate that QDyLoRA can outperform QLoRA when the rank is chosen optimally for the deployment setting. This makes it particularly appealing for low-resource and edge-device applications, where memory and computational budgets are highly variable.

Drawbacks of LoRA

Even LoRA enabled researchers to optimize computer resources for fine-tuning LLMs, it is not without its drawbacks. Shuttleworth et al.[31] lists some considerations regarding to use LoRA. Lower-rank LoRA models may trade off generalization for adaptation, potentially resulting in the loss of knowledge learned during pre-training, especially if the adaptation task is very different from the pre-training domain. Another consideration is LoRA-adapted models may "forget" more of the original pre-trained distribution compared to models that undergo full fine-tuning, resulting in lower robustness when exposed to new or sequential tasks. Finally, for certain challenging domains, such as code generation, long-form text generation, or tasks requiring deep compositional reasoning, LoRA can fall short of full fine-tuning performance.

Han et al.[32] also observes that LoRA may not fully capture the needed adaptation for domains that are highly divergent from the pre-trained distribution, due to the limited parameter space it modifies.

adaptMLLM

adaptMLLM[33] is a multilingual fine-tuning framework for fine-tuning pre-built MLLMs to enhance machine translation (MT) with a particular focus on low-resource language pairs. DeepSpeed[34], a free software library developed by Microsoft, is a critical component of adaptMLLM. It enables efficient distribution of models across both GPU and system memory, allowing

large models to be fine-tuned even on limited hardware. By integrating DeepSpeed with a streamlined workflow and user-friendly interface, adaptMLLM enables model fine-tuning at a fraction of the time and cost required for training a full model from scratch.

Computational Resources

The training of LLMs requires enormous computational power, often involving hundreds or thousands of GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units). While GPUs are widely used for a range of parallel processing tasks, TPUs are specialized hardware accelerators designed specifically by Google for efficiently training large neural networks. Their architecture is optimized for the types of operations commonly performed during machine learning, such as matrix multiplications, making them particularly well-suited for the demands of LLM training. However, this level of resource consumption raises concerns about the environmental impact and the accessibility of such technology. The high costs associated with these powerful processing units mean that only well-funded institutions and companies can afford to train and deploy these models at scale, particularly during the initial pre-training stage. This concentration of resources worsens the disparity between institutions with different levels of funding, limiting the ability of smaller organizations to contribute to advancements in the field. The deployment of a larger number of GPUs significantly accelerates training times and enhances model stability, giving organizations with greater GPU resources a substantial advantage in advancing AI research and development. For example, the xAI Colossus AI supercomputer developed by X company has more than 300000 H100 GPUs[35], while one of the biggest clusters in Turkey has a maximum of only 100 H100 GPUs[36].

Carbon Footprint

Carbon emissions depend on both the type of energy sources used to power the electricity grid and the total amount of energy consumed. The exact amount of CO2 emissions made during training an LLM model like OpenAI's is difficult to estimate as such figures are usually not publicly disclosed. Since most large-scale LLM training is conducted on cloud platforms, the environmental impact is influenced by the carbon intensity of the grid in the region where the data center is located. Some cloud providers publish their carbon related emissions in their websites like Google[37]. Google defines an emission metric, Grid Carbon Intensity (gCO2eq/kWh), which indicates the average operational gross carbon emissions per unit of energy from the grid. For example, Google's europe-north2 region (Stockholm) reports a grid intensity of 22 gCOeq/kWh, whereas the us-east1 region (South Carolina) reports 560 gCOeq/kWh—reflecting the greater reliance on renewable energy in Stockholm[37].

Emissions are a function of both energy consumption and the type of fuel powering the grid. Therefore, energy usage becomes a critical factor when assessing the environmental impact of different model training approaches. For example, training a machine translation model from English to Irish (ga) using AdaptMLLM takes approximately 4 hours and consumes around 1.1 kWh of energy. In contrast, training a large-scale model like GPT-3 is estimated to consume approximately 1,287,000 kWh, illustrating the substantial energy - and consequently, emissions - gap between lightweight fine-tuning and full-scale pretraining[38].

Ethical Considerations in Low-Resource NLP

Data Quality and Bias

The quality of the training data directly impacts the performance of the model. However, large datasets often contain biases that can be inadvertently learned by the model. For instance, if the training data over-represents certain demographics or viewpoints, the model may produce biased outputs, which can perpetuate stereotypes or misinformation.

Ethical and Societal Implications

The deployment of LLMs in real-world applications raises ethical questions, particularly concerning privacy, misinformation, and the potential for misuse. For example, LLMs can generate convincing fake news or impersonate individuals in text, leading to concerns about their role in spreading misinformation or being used in malicious ways.

A notable case study highlighting these concerns is the use of GPT-3, a powerful language model developed by OpenAI. In 2020, researchers demonstrated that GPT-3 could generate highly realistic and coherent text, including fake news articles and deceptive content. This capability raised alarms about the potential for misuse, as malicious actors could exploit such models to create misleading information or impersonate individuals, thereby amplifying the spread of misinformation and undermining trust in digital communications[39]. The ability of GPT-3 to generate human-like text, highlighted the pressing need for ethical guidelines and safeguards to prevent the misuse of advanced language models.

While issues like bias and misinformation are critical in evaluating the ethical dimensions of Large Language Models (LLMs), data sovereignty, community consent, and the rights of speakers of low-resource languages are equally important concerns. Many low-resource language datasets are sourced through web scraping or community-contributed content, yet such practices often overlook informed consent and cultural rights of the communities involved.

A growing body of literature warns that collecting and using linguistic data without explicit permission can constitute a form of digital colonization, where communities lose control over their cultural and linguistic heritage[40].

To ethically engage with these communities, researchers and developers must adopt community-centered approaches. This includes:

- Engaging communities as collaborators, not just data sources, during data collection and annotation.
- Ensuring data sovereignty, where communities retain control over how their data is used, shared, and represented in NLP systems.
- Using frameworks such as CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, Ethics)[41].

With these steps, the pursuit of technological inclusion becomes less exploitative. Therefore, advancing NLP for low-resource languages must not only be a technical goal but also a moral and political commitment to respecting linguistic and cultural rights.

Deployment Harms and Equity Risks

While the technical challenges of training LLMs for low-resource languages are considerable, equal attention must be paid to the potential deployment harms these models may cause. In domains such as healthcare and legal systems, underperforming models trained on sparse or biased data can propagate systemic inequalities. For instance, a language model deployed in a clinical setting that misinterprets inputs in a low-resource language may lead to incorrect diagnoses or treatment plans, disproportionately affecting already marginalized populations. Similarly, legal language models that are lacking cultural and linguistic nuance may fail to understand or translate critical legal concepts, thereby preventing fair access to justice. These risks are heightened by the tendency of LLMs to reflect and amplify biases present in training data—an issue particularly important in low-resource settings where the available data may be unbalanced or unrepresentative. Therefore, ensuring model transparency, fairness, and continual evaluation is essential to prevent the amplification of existing social inequities through automated systems.

Discussion

The investigation into training Large Language Models (LLMs) underscores a multifaceted challenge influenced by several critical factors that shape the effectiveness of these models, particularly in the context of low-resource languages. Our review of the literature reveals that the efficacy of LLMs is profoundly impacted by the availability of training data, the computational resources required, and the inherent biases in data and models.

High-resource languages, such as English and Mandarin, benefit from vast and diverse datasets, which enable LLMs to learn detailed linguistic patterns and generalize effectively across various tasks. This extensive training data allows these models to achieve high performance and robustness in tasks ranging from text generation to sentiment analysis. However, the situation for low-resource languages is markedly different. These languages often lack sufficient digital text corpora and annotated datasets, which are crucial for training effective models. The limited data availability hampers the ability of LLMs to capture the nuances and complexities of these languages, leading to models that may produce less accurate and less reliable outputs. This scarcity of data also limits the ability to create comprehensive and representative models that can generalize well to different contexts and dialects.

The computational resources required for training LLMs further compound the challenge. Training state-of-the-art models demands substantial hardware and energy investments, often only accessible to well-funded institutions. This disparity creates an additional barrier for low-resource languages, as the necessary computational power and resources are beyond the reach of many researchers and institutions working in these language areas. The environmental impact of such intensive computational processes also raises concerns about the sustainability of current training practices and highlights the need for more efficient methodologies.

Moreover, data quality and biases present significant challenges in training LLMs. High-resource language datasets, while abundant, often contain biases that can be inadvertently learned by models, leading to outputs that may reinforce existing stereotypes or propagate misinformation. For low-resource languages, the challenge is twofold: not only is there a lack of data, but the data that is available may be limited in scope and quality, further affecting the model's performance and reliability.

Our review also points to the necessity of developing innovative approaches to address these challenges. Strategies such as cross-lingual transfer learning, which leverages knowledge from high-resource languages, and data augmentation techniques, like back-translation and synthetic data generation, offer promising avenues for improving LLM performance in low-resource contexts. Engaging with community-driven efforts and collaborative research can also enhance the availability of linguistic resources and tools, contributing to more equitable advancements in NLP.

In summary, while significant progress has been made in the field of LLMs in terms of achieving advanced model architectures and performance benchmarks for high-resource languages, substantial challenges remain in adapting training methods for low-resource languages. Addressing these challenges requires a concerted effort to enhance data availability, optimize com-

putational resources, and mitigate biases, ultimately leading to more inclusive and effective language technologies that serve to a broader linguistic diversity and provide equal access to NLP advancements.

Conclusions and Future Directions

As we look to the future, several key areas hold promise for advancing Large Language Models (LLMs) and their applications in NLP. One critical focus is improving support for low-resource languages. Recent studies have highlighted the potential of few-shot learning techniques, which allow models to generalize from limited examples, and cross-lingual transfer learning, which leverages knowledge from high-resource languages to benefit low-resource ones[42]. Conneau et al.[43] demonstrated that multilingual models trained on a diverse set of languages can significantly improve performance on low-resource languages through transfer learning. Future research is likely to build on these findings by exploring more efficient methods for data augmentation and model adaptation to enhance performance with minimal annotated data.

Moreover, there is a strong emphasis on developing more efficient training methodologies. Advances in sparsity-inducing regularization, such as the work by Frankle et al.[44] have shown that pruning unnecessary weights can reduce computational costs without sacrificing model performance. The design of efficient neural architectures, like the Transformer variants proposed by Huang et al.[45] and the optimization of hardware utilization are expected to further reduce the computational and environmental costs associated with training large models. For example, the introduction of sparse transformers has led to significant improvements in training efficiency while maintaining model effectiveness. These innovations are crucial for making these technologies more accessible and sustainable.

The field will also benefit from increased efforts to improve the interpretability and explainability of LLMs. Recent research by Ribeiro et al.[46] on model interpretability techniques, such as LIME, has underscored the importance of understanding and interpreting model decisions to enhance trust and reliability. Developing methods to better understand and interpret model decisions will be particularly important in critical applications like healthcare and legal systems, where the consequences of incorrect predictions can be severe.

Ethical considerations and bias mitigation will remain critical priorities. Studies such as those by Bolukbasi et al.[47] have revealed significant biases in NLP models, emphasizing the need for ongoing work to create models that are fair and inclusive. Future work will likely focus on developing frameworks for evaluating and mitigating biases in NLP technologies to ensure that the benefits of these technologies are equitably distributed across different user groups.

In addition to advancing LLMs and NLP technologies, linguists have a crucial role to play in this evolving landscape. They should focus on collaborating with computational researchers to develop and refine models that accurately capture the nuances of diverse languages and dialects. Linguists can contribute by providing deep linguistic insights and annotations that improve the quality of training data, particularly for underrepresented languages. They should also advocate for ethical considerations and fairness in NLP applications, ensuring that linguistic diversity is respected and biases are addressed. Moreover, linguists can engage in interdisciplinary research to explore how advancements in NLP can support language preservation and revitalization efforts. By bridging the gap between theoretical linguistics and practical NLP applications, linguists can help shape more inclusive and effective language technologies.

The following topics are recommended for future research:

1. **Dynamic Parameter-Efficient Fine-Tuning:** Explore adaptive fine-tuning methods (e.g., QDyLoRA) that adjust training depth or rank based on language complexity and resource availability.

2. **Culturally-Aware Dataset Development:** Encourage the creation of multilingual datasets by creating methods (like mobile apps) that integrate local knowledge, cultural nuance, and ethical data sourcing for more representative training.

3. **Mitigating Intruder Dimensions in LoRA Fine-Tuning:** Investigate how spectral anomalies like "intruder dimensions" affect model robustness and transfer learning and develop techniques to minimize their impact.

4. **Benchmark a broader range of LLMs fine-tuned with different LoRA variants and token counts:** Conduct systematic benchmarking across various LoRA-based fine-tuning strategies (e.g., standard LoRA, QLoRA, DyLoRA) using LLMs of different sizes and training token budgets. This would help quantify trade-offs between parameter efficiency, performance, and generalization in low-resource language settings.

Overall, these future directions, grounded in current research, will drive the ongoing evolution of NLP technologies by enhancing their ability to understand and generate diverse linguistic content more accurately and efficiently. This includes improving cross-lingual capabilities, enabling nuanced sentiment analysis, and advancing real-time translation across a broader spectrum of languages. Additionally, the development of more sophisticated dialogue systems and context-aware models will contribute to more personalized and effective human-computer interactions. The positive societal impact will be profound, as these advancements can foster greater global communication, support efforts in language preservation and revitalization, and provide more

equitable access to information and digital resources for underrepresented and marginalized communities. Such progress will not only bridge linguistic divides but also contribute to a more inclusive and informed global society.

Acknowledgments

The author would like to thank Lumiere Research for their valuable support and guidance throughout the development of this study.

References

1 J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph and S. Borgeaud, *Emergent abilities of large language models*.

2 C. Mishra and D. Gupta, *Deep machine learning and neural networks: an overview*.

3 H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi and Y. Babaei, *Llama 2: open foundation and fine-tuned chat models*.

4 T. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić and D. Hesslow, *BLOOM: a 176b-parameter open-access multilingual language model*.

5 C. Shannon, *A mathematical theory of communication*.

6 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones and A. Gomez, *Attention is all you need*.

7 J. Kaplan, S. McCandlish, T. Henighan, T. Brown, B. Chess and R. Child, *Scaling laws for neural language models*.

8 S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma and Z. Jiang, *A comparative study on transformer vs rnn in speech applications*.

9 K. Taghandiki and M. Mohammadi, *Large language models training, challenges, applications, and development*.

10 G. Vrbančič and V. Podgorelec, *Transfer learning with adaptive fine-tuning*.

11 S. Chaudhari, P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan and K. Narasimhan, *RLHF deciphered: a critical analysis of reinforcement learning from human feedback for llms*.

12 M. Hedderich, L. Lange, H. Adel, J. Strötgen and D. Klakow, *A survey on recent approaches for natural language processing in low-resource scenarios*.

13 K. Lu, Y. Yang, F. Yang, R. Dong, B. Ma and A. Aihemaiti, *Low-resource language expansion and translation capacity enhancement for llm*, a study on the uyghur. .

14 R. Arreerard, S. Mander and S. Piao, *Survey on thai nlp language resources and tools*.

15 S. Bessou and M. Touahria, *Morphological analysis and generation for machine translation from and to arabic*.

16 H. Park, K. Zhang, C. Haley, K. Steimel, H. Liu and L. Schwartz, *Morphology matters: a multilingual language modeling analysis*.

17 S. Cahyawijaya, H. Lovenia and P. Fung, *LLMs are few-shot in-context low-resource language learners*.

18 X. Jia, J. Wang, Z. Zhang, N. Cheng and J. Xiao, *Large-scale transfer learning for low-resource spoken language understanding*.

19 B. Li, Y. Hou and W. Che, *Data augmentation approaches in natural language processing: a survey*.

20 P. Kumar, *Large language models (llms): survey, technical frameworks, and future challenges*.

21 B. Zhang, Z. Li, Z. Gan, Y. Chen, J. Wan and K. Liu, *CroAno : a crowd annotation platform for improving label consistency of chinese ner dataset*.

22 *African Languages Lab*, https://www.africanlanguageslab.com/.

23 H. Hjartarson and S. Fririksdóttir, *Malmon: a crowd-sourcing platform for simple language*.

24 Vistatec, *How to Overcome the Need for Data for Low-Resource Languages*, https://vistatec.com/how-to-overcome-the-need-for-data-for-low-resource-languages/.

25 E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li and S. Wang, *LoRA: low-rank adaptation of large language models*.

26 A. Singh, N. Pandey, A. Shirgaonkar, P. Manoj and V. Aski, *A study of optimizations for fine-tuning large language models*.

27 S. Raschka, *Finetuning falcon llms more efficiently with lora and adapters*.

28 T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, *QLoRA: efficient finetuning of quantized llms*.

29 M. Valipour, M. Rezagholizadeh, I. Kobyzev and A. Ghodsi, *DyLoRA: parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation*.

30 H. Rajabzadeh, M. Valipour, T. Zhu, M. Tahaei, H. Kwon and A. Ghodsi, *QDyLoRA: quantized dynamic low-rank adaptation for efficient large language model tuning*.

31 R. Shuttleworth, J. Andreas, A. Torralba and P. Sharma, *LoRA vs full fine-tuning: an illusion of equivalence*.

32 Z. Han, C. Gao, J. Liu, J. Zhang and S. Zhang, *Parameter-efficient fine-tuning for large models: a comprehensive survey*.

33 S. Lankford, H. Afli and A. Way, *AdaptMLLM: fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds*.

34 J. Rasley, S. Rajbhandari, O. Ruwase and Y. He, *DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters*.

35 M. Tyson, *Elon Musk Is Doubling the World's Largest AI GPU Cluster — Expanding Colossus GPU Cluster to 200,000 "soon*, https://www.tomshardware.com/pc-components/gpus/elon-musk-is-doubling-the-worlds-largest-ai-gpu-cluster-expanding-colossus-gpu-cluster-to-200-000-soon-has-floated-300-000-in-the-past, Has Floated 300,000 in the Past. Tom's Hardware.

36 *Türkiye'nin Süper Bilgisayarı ARF ACC'den Büyük Başarı — ULUSAL AKADEMİK AĞ ve BİLGİ MERKEZİ*, https://ulakbim.tubitak.gov.tr/tr/haber/turkiyenin-super-bilgisayari-arf-accden-buyuk-basari.

37 *Carbon Free Energy for Google Cloud Regions*, https://cloud.google.com/sustainability/region-carbon., Google Cloud.

38 D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia and D. Rothchild, *Carbon emissions and large neural network training*.

39 L. Floridi and M. Chiriatti, *GPT-3: its nature, scope, limits, and consequences*.

40 S. Bird, *Decolonising speech and language technology*.

41 S. Carroll, I. Garba, O. Figueroa-Rodríguez, J. Holbrook, R. Lovett and S. Materechera, *The care principles for indigenous data governance*.

42 A. Conneau and G. Lample, *Cross-lingual language model pretraining*.

43 A. Conneau, A. Baevski, R. Collobert, A. Mohamed and M. Auli, *Unsupervised cross-lingual representation learning for speech recognition*.

44 J. Frankle and M. Carbin, *The lottery ticket hypothesis: finding sparse, trainable neural networks*.

45 X. Huang, F. Perez, J. Ba and M. Volkovs, *Improving transformer optimization through better initialization*.

46 M. Ribeiro, S. Singh and C. Guestrin, *Why should i trust you?": explaining the predictions of any classifier*.

47 T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama and A. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*.