Parameter-Efficient Fine-Tuning for Multilingual Natural Language Understanding

Isaac J. Carr
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
carr1985@unh.edu

Geurav Sainha
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
gauravsinha266@buffalo.edu

Hego Lawrence
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
contacthugo@unr.edu

Abstract

Parameter-efficient fine-tuning has become a decisive architectural and operational strategy for adapting large pretrained language models to multilingual natural language understanding tasks without retraining or storing full model replicas. This paper presents a systems-oriented analysis of parameter-efficient adaptation, focusing on structural trade-offs among adapter modules, low-rank decompositions, prefix-based approaches, and selective parameter updates. The discussion emphasizes cross-lingual transfer dynamics, in which task-specific and language-specific knowledge must be balanced under limited parameter budgets. Rather than treating parameter efficiency solely as a compression problem, the paper examines multilingual adaptation as a socio-technical infrastructure challenge involving deployment constraints, evaluation fairness, energy consumption, and governance. The analysis shows how small parameter updates can preserve strong zero-shot cross-lingual capacity while reducing memory and communication overhead in distributed training and serving environments. The paper also addresses risks of linguistic performance asymmetry, tokenizer-induced cost disparities, and evaluation benchmarks that obscure underperformance in low-resource languages. Through a synthesis of architectural strategies, deployment considerations, and policy implications, the paper argues that parameter-efficient multilingual fine-tuning should be governed by explicit fairness, sustainability, and auditability principles. The analysis contributes a system-level framework for designing multilingual natural language understanding systems that are not only computationally tractable but also institutionally accountable and linguistically inclusive.

Keywords

Parameter-efficient fine-tuning; multilingual natural language understanding; cross-lingual transfer; model governance; low-rank adaptation; linguistic fairness.

1. Introduction

The adaptation of large pretrained language models to downstream natural language understanding tasks has shifted from task-specific architectures toward unified multilingual encoders and decoders. Full fine-tuning of these models is increasingly difficult to justify in

operational settings because each downstream language-task pair can require storing and serving a full model replica. Parameter-efficient fine-tuning has therefore emerged as a structural answer to the growing gap between model scale and deployment budgets [1]. In contrast to full fine-tuning, parameter-efficient methods freeze the majority of pretrained weights and modify a small set of task-specific or language-specific parameters. Low-rank adaptation further reduces the trainable footprint by decomposing weight updates into low-dimensional components [2].

Multilingual natural language understanding makes this problem more acute because linguistic diversity introduces many axes of variation. A single multilingual model may need to handle typologically distant languages, code-switched input, nonstandard orthographies, and varied levels of digital resource availability. Multilingual pretraining objectives such as those used in XLM-R have produced strong cross-lingual representations [3], but downstream adaptation still demands careful allocation of new parameters. Adapter-based frameworks such as MAD-X separate language and task adapters, allowing some parameters to be shared across languages while others remain specialized [4]. This separation is structurally significant because it decouples the cost of adding a new language from the cost of adding a new task.

This article examines parameter-efficient fine-tuning for multilingual natural language understanding as a systems problem rather than as an isolated model compression technique. The analysis covers architectural strategies, cross-lingual transfer behavior, inference and serving infrastructure, evaluation fairness, robustness, and governance. The central argument is that parameter efficiency cannot be evaluated solely by task accuracy; it must be assessed alongside linguistic equity, operational sustainability, auditability, and the institutional capacity to maintain multilingual systems over time.

2. Motivations and Structural Pressures for Parameter-Efficient Multilingual Adaptation

Full parameter fine-tuning remains a powerful adaptation strategy when compute budgets are large and the number of target tasks is small. However, in multilingual settings the number of language-task combinations grows combinatorially, making full model replication infeasible for many institutions. Parameter-efficient transfer learning reduces this cost by inserting compact adapter modules into transformer layers while keeping the main model frozen [1]. The core motivation is not merely to reduce training time, but to create a stable shared backbone that can support many downstream languages and tasks without multiplying infrastructure demands.

Selective parameter updating offers another route to lower operational overhead. Methods such as BitFit demonstrate that updating only bias terms can retain substantial task performance while modifying an extremely small fraction of the total parameters [5]. In multilingual contexts, selective update strategies are attractive because they reduce the risk of catastrophic interference with pretrained cross-lingual representations. At the same time, they place greater pressure on the pretrained model to already encode linguistic features that are relevant to the target languages. The resulting trade-off is between preserving general multilingual competence and injecting task-specific signal without destabilizing shared parameters.

Beyond computational cost, institutional and environmental pressures shape multilingual adaptation. Training large models from scratch for every language is carbon-intensive and

inequitable, particularly for underserved language communities. Parameter-efficient methods can reduce energy consumption by reusing a shared pretrained backbone and training only small adapters. However, energy savings are not automatic; they depend on the number of training steps, hyperparameter search overhead, and the efficiency of data pipelines. The structural promise of parameter-efficient multilingual adaptation is therefore tied to careful experimental governance and a clear accounting of resource boundaries.

3. Architectural Strategies for Parameter-Efficient Multilingual Adaptation

Adapter-based methods insert small bottleneck modules inside transformer layers while keeping the pretrained weights frozen. These modules typically contain down-projection and up-projection layers with a nonlinearity, but the conceptual core is that the base model remains a shared multilingual encoder and only the adapters are optimized for a given task or language [1], [4]. This design allows a single pretrained model to be reused across many language-task pairs, while each adapter remains small enough to be stored, versioned, and transmitted independently. The modularity of adapter-based systems also creates opportunities for sharing adapters across related languages or tasks, although it introduces questions about compatibility and composition.

Low-rank adaptation decomposes weight updates into products of low-rank matrices, thereby reducing memory overhead and enabling efficient task switching by swapping small update matrices [2]. This approach is particularly compatible with multilingual serving because many language-specific update matrices can be stored alongside one shared base model. Prefix-based and prompt-based methods instead inject additional trainable vectors into the input or hidden state sequence. These methods avoid modifying the core computational graph, but their effectiveness depends on the length budget of the prefix and the degree to which the pretrained model can interpret newly injected context across languages. BitFit demonstrates that even bias-only updates can be effective for many natural language understanding tasks, highlighting that the transformer contains numerous small parameter groups with distinct adaptation capacity [5].

Architectural choices have direct consequences for multilingual transfer. Adapters that are inserted at all layers can learn language-specific transformations but may require careful initialization and regularization to avoid overfitting in low-resource languages. Low-rank updates applied uniformly across weight matrices can preserve general linguistic structure but may underfit when target languages require substantial syntactic or lexical adaptation. Prefix-based methods can be lightweight but are often sensitive to sequence length and may introduce positional interference. The selection of a parameter-efficient architecture should therefore be informed by the expected number of languages, the similarity across languages, the amount of labeled data, and the serving environment. A system designed for high-resource typologically similar languages may succeed with shared low-rank matrices, whereas a system serving many low-resource languages may require language-specific adapters with stronger capacity.

4. Cross-Lingual Transfer and Structural Trade-offs

Although this paper focuses on multilingual natural language understanding, adjacent research on small parameter language models has shown that scalable reasoning path filtering and alignment can improve efficiency without full retraining [6]. The underlying insight, that targeted update policies can align model behavior under constrained parameter budgets, extends to multilingual adaptation. Small parameter updates can be interpreted as policy

interventions that redirect a pretrained multilingual model toward task-relevant behavior while preserving prior linguistic knowledge.

Multilingual transformer models exhibit surprising cross-lingual transfer, but their multilingual capacity is not uniform. Studies of multilingual BERT have shown that the model encodes syntactic information across languages, yet performance varies by resource availability and typological distance [7]. Subsequent work demonstrates that zero-shot cross-lingual transfer often declines for typologically distant languages or those with limited representation in pretraining [8]. Even when cross-lingual effectiveness is strong, the distribution of benefits is uneven, with high-resource languages generally receiving greater representational capacity [9]. Parameter-efficient fine-tuning can mitigate these asymmetries by allowing language-specific adapters to capture linguistic features that are not adequately represented in the shared backbone. However, adding too much language-specific capacity can reduce cross-lingual generalization by fragmenting the shared representation space.

Multilingual adapter generation addresses the fragmentation risk by producing language adapters from a generative process rather than training each language adapter independently [10]. This approach is structurally attractive because it permits parameter sharing across languages while retaining language-specific adjustment. Typology-based adapters can further condition adaptation on linguistic features such as word order, morphological complexity, or case marking [11]. These strategies illustrate a broader principle: multilingual parameter-efficient adaptation should treat language similarity as a design variable rather than as a constant. By organizing adapters around typological structure, systems can improve transfer to low-resource languages while containing the total parameter budget.

The structural trade-offs are not limited to adapter placement. The interaction between language-specific and task-specific parameters determines whether a new language can be added without retraining the entire system. If task parameters are entangled with language parameters, adding a language may require revisiting many tasks. If language parameters and task parameters are fully separated, the system becomes more modular but may suffer from reduced cross-task generalization. The design challenge is to balance modularity and shared representation under a bounded parameter budget while ensuring that low-resource languages do not become permanent outliers in the adaptation landscape.

5. Infrastructure and Deployment Considerations

Deployment infrastructure shapes the practical value of parameter-efficient multilingual fine-tuning. AdapterHub demonstrates that transformer adaptation can be organized as a modular ecosystem in which task and language adapters are trained, shared, and composed without modifying the base model [12]. Such an ecosystem is particularly valuable for multilingual natural language understanding because it allows institutions to maintain a common backbone while distributing language-specific updates across teams. The operational model resembles a software package repository: the base model remains stable, and adaptation modules can be versioned, audited, and reused.

Scalable adaptation architectures from neural machine translation have informed multilingual fine-tuning by showing that small adapter networks can be inserted into large models without requiring full retraining for each language pair [13]. Quantized low-rank adaptation extends this logic by enabling fine-tuning on quantized base models, thereby reducing memory consumption during training and allowing larger models to be adapted on limited hardware [14]. These infrastructure gains are meaningful in multilingual settings because many research

and public service institutions lack access to large-scale accelerator clusters. Parameter-efficient methods can therefore widen participation in multilingual natural language understanding, but only if the surrounding tooling is usable and the base model remains accessible.

Tokenization and sequence processing introduce additional infrastructure pressures in multilingual systems. Multilingual models often use shared subword vocabularies that allocate fewer tokens to some languages, increasing the number of subword units and the computational cost for those languages [15]. This cost asymmetry can interact with parameter-efficient adaptation because a language that requires more subword tokens may also place greater pressure on adapter capacity. Deployment planners should therefore measure not only accuracy but also tokenization efficiency and latency across languages. A system that appears parameter-efficient in aggregate may still impose disproportionate inference costs on certain language communities.

From a serving perspective, parameter-efficient modules reduce the storage and switching costs associated with multilingual task serving. A single base model can be loaded into GPU memory while small language-task adapters are swapped dynamically. However, the system must still manage adapter routing, version compatibility, and evaluation consistency. If the base model is updated, all dependent adapters may need retraining or validation. Infrastructure governance must therefore include explicit compatibility policies and monitoring mechanisms that track accuracy drift across languages after any component update.

6. Robustness, Fairness, and Evaluation

Evaluation of multilingual parameter-efficient systems requires attention to cross-lingual transferability and representation quality. Studies of monolingual representations have shown that transferability to a target language depends on structural properties of the source and target languages, not merely on model size [16]. This finding implies that parameter-efficient fine-tuning cannot assume uniform transfer across all languages. Evaluation must therefore include targeted probing of syntactic and semantic generalization in multiple languages rather than relying on average multilingual scores.

Benchmarking frameworks such as XTREME-R emphasize more challenging reasoning and reading comprehension tasks across many languages, revealing that high aggregate performance can hide severe degradation in low-resource languages [17]. In parameter-efficient systems, this degradation may arise from limited adapter capacity, imbalanced tokenization, or insufficient language-specific training data. Fairness-oriented evaluation should report per-language performance distributions, track performance disparities across resource levels, and include qualitative error analysis in languages that fall below operational thresholds. Without such granularity, system developers may optimize for average accuracy while leaving marginalized language communities underserved.

Robustness in multilingual fine-tuning also depends on the stability of parameter updates. Small parameter budgets can reduce overfitting, but they can also produce adapters that are brittle under distribution shift, code-switching, or dialectal variation. Regularization, data augmentation, and multilingual consistency constraints can improve robustness, but they require system-level evaluation beyond standard benchmark accuracy. The operational question is whether a parameter-efficient model maintains acceptable behavior across monolingual, cross-lingual, and code-switched inputs over time. Continuous evaluation and

model monitoring should therefore be treated as part of the fine-tuning lifecycle rather than as a final acceptance test.

7. Governance, Policy, and Sustainability Implications

Governance of multilingual language technologies must address the risks of encoding harmful biases and reproducing unequal language access. The critique of large language models as stochastic parrots underscores the importance of documentation, accountability, and careful deployment when systems produce fluent but ungrounded or harmful outputs [18]. Parameter-efficient fine-tuning does not eliminate these risks; it changes their locus. A small adapter can steer a large multilingual model toward sensitive or unsafe content just as effectively as full fine-tuning in some cases. Governance frameworks must therefore track adapter provenance, training data, and intended use, and they must require auditing of language-specific behavior.

Foundation model reports have highlighted the broad societal implications of large-scale model development and deployment [19]. In multilingual contexts, these implications include linguistic surveillance, cultural homogenization, and the concentration of AI capacity in institutions that control large pretrained models. Parameter-efficient adaptation can lower the barrier for smaller institutions to build language-specific applications, but only if base models and adaptation tooling are openly available and well documented. Policy interventions should encourage shared multilingual backbones, transparent adapter registries, and community participation in language resource development.

Energy and policy considerations are equally central to sustainability [20]. Although parameter-efficient methods reduce the marginal cost of fine-tuning compared with full model training, the aggregate environmental footprint depends on how many adapters are trained, how often they are retrained, and how efficiently inference is provisioned. A well-governed multilingual system should include carbon accounting for training and serving, strategies for adapter reuse, and procurement requirements that favor energy-efficient infrastructure. In addition, policies should support data collection for low-resource languages so that efficiency gains do not simply reinforce existing resource asymmetries.

Finally, auditability and institutional accountability require that parameter-efficient modules be treated as versioned artifacts with clear compatibility metadata. Model cards for adapters can document language coverage, training data limitations, evaluation results, and known failure modes. Such practices align with broader movements toward algorithmic accountability and can help multilingual systems remain trustworthy as they scale.

8. Conclusion

Parameter-efficient fine-tuning is not merely a technical convenience for multilingual natural language understanding; it is a structural intervention that reshapes how language technology is built, deployed, and governed. Adapter-based methods, low-rank updates, prefix-based approaches, and selective parameter updates each offer distinct trade-offs among transfer quality, modularity, serving efficiency,Parameter-efficient fine-tuning is not merely a technical convenience for multilingual natural language understanding; it is a structural intervention that reshapes how language technology is built, deployed, and governed. Adapter-based methods, low-rank updates, prefix-based approaches, and selective parameter updates each offer distinct trade-offs among transfer quality, modularity, serving efficiency, robustness, and linguistic fairness. The multilingual setting intensifies these trade-offs because linguistic diversity, uneven resource availability, typological distance, and evaluation asymmetries interact with parameter budgets in ways that are not captured by aggregate task

accuracy alone. A modular adapter ecosystem can allow a shared multilingual backbone to serve many languages and tasks without full model replication, but it requires careful governance of adapter compatibility, provenance, and performance across languages. Low-rank updates reduce memory and communication overhead, yet they may underfit languages whose structural properties diverge sharply from those heavily represented in the pretrained model. Selective parameter updates preserve cross-lingual capacity but depend on the pretrained encoder already containing useful multilingual structure. The operational value of these methods therefore lies not in any single technique, but in an integrated design philosophy that treats language as a first-class system variable.

Multilingual natural language understanding systems are embedded in broader infrastructures of data collection, model sharing, evaluation, and institutional accountability. Parameter-efficient fine-tuning can lower the cost of language-specific adaptation, thereby enabling more inclusive participation by smaller research groups, public institutions, and language communities. At the same time, efficiency gains can obscure the continued concentration of power in organizations that control large pretrained models and the compute resources required to train them. Governance mechanisms must ensure that parameter-efficient modules are documented, versioned, and audited with attention to language-specific harms and limitations. Evaluation should move beyond average cross-lingual scores toward per-language reporting, robustness testing under code-switching and dialectal variation, and monitoring of performance drift over time. Energy accounting should cover adapter training, hyperparameter search, and serving overhead, not only the marginal reduction compared with full fine-tuning. By treating parameter efficiency as part of a broader system of linguistic fairness, sustainability, and institutional responsibility, researchers and practitioners can build multilingual natural language understanding systems that are both computationally tractable and socially accountable.

References

1. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, 2790–2799. PMLR.

2. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. Association for Computational Linguistics.

4. Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 7654–7673. Association for Computational Linguistics.

5. Ben Zaken, E., Goldberg, Y., & Ravfogel, S. (2022). BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Proceedings of the 60th

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 1–9. Association for Computational Linguistics.

6. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In 2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE) (pp. 29-32). IEEE.

7. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996–5001. Association for Computational Linguistics.

8. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Proceedings of the 37th International Conference on Machine Learning, 4411–4421. PMLR.

9. Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., & Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019.

10. Ansell, A., Ponti, E. M., Pfeiffer, J., Ruder, S., Glavaš, G., Vulić, I., & Korhonen, A. (2022). MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In Findings of the Association for Computational Linguistics: EMNLP 2022, 4762–4781. Association for Computational Linguistics.

11. Üstün, A., Bisazza, A., Bouma, G., & van Noord, G. (2020). UDapter: Language adaptation for universal dependencies. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2196–2206. Association for Computational Linguistics.

12. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2021). AdapterHub: A framework for adapting transformers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 46–54. Association for Computational Linguistics.

13. Bapna, A., & Firat, O. (2019). Simple, scalable adaptation for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 1538–1548. Association for Computational Linguistics.

14. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems 36, 10088–10115. Curran Associates.

15. Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 3118–3135. Association for Computational Linguistics.

16. Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In Proceedings of the 58th Annual Meeting of the

Association for Computational Linguistics, 4623–4637. Association for Computational Linguistics.

17. Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 10215–10245. Association for Computational Linguistics.

18. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. Association for Computing Machinery.

19. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

20. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. Association for Computational Linguistics.