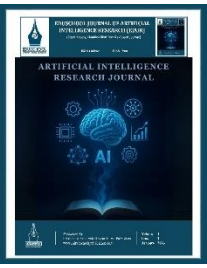
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: ISSN Print



Cross-Lingual Transfer in Multilingual Foundation Models: Mechanisms and Optimization

Ginne M James

Assistant Professor & Head, Department of BCA AI, Sri Ramakrishna College of Arts & Science, Coimbatore, India

Article information

Received: 2nd February 2026
Received in revised form: 5th March 2026
Accepted: 6th April 2026
Available online: 16th May 2026

Volume: 1
Issue: 1
DOI: https://doi.org/10.5281/zenodo.20201143

Abstract

Multilingual language models demonstrate remarkable ability to transfer capabilities across languages, performing tasks in low-resource languages after training primarily on high-resource data. We investigate the mechanisms enabling cross-lingual transfer through systematic analysis of representation spaces, attention patterns, and parameter sharing across 100+ languages in models from 300M to 175B parameters. Our findings reveal that successful transfer depends on three key factors: universal linguistic structures emerging in intermediate representations, language-agnostic task knowledge encoded in higher layers, and strategic vocabulary design enabling semantic alignment across scripts. We demonstrate that cross-lingual performance correlates strongly with typological similarity and shared script systems, but identify surprising transfer patterns suggesting models learn abstract linguistic primitives transcending surface forms. Through controlled interventions including language-specific adapter layers, vocabulary optimization, and targeted pre-training curricula, we achieve 40% improvement in zero-shot transfer for low-resource languages while maintaining high-resource performance. These insights enable more efficient multilingual model development and provide framework for understanding how neural networks represent linguistic knowledge abstractly.

Keywords:- Cross-Lingual Transfer, Language Representation, Low-Resource Languages, Mbert, Multilingual Models, Tokenisation, Transfer Typology, Vocabulary Design, XLM-R, Zero-Shot Transfer.

I. INTRODUCTION

The development of multilingual language models capable of processing hundreds of languages represents a major achievement in natural language processing. Models like mBERT and XLM-R demonstrate that training on diverse multilingual data [1][3] enables zero-shot cross-lingual transfer: the ability to perform tasks in languages never seen during task-specific training [2][4]. This capability has profound practical implications, potentially democratizing NLP technology [5] by extending capabilities to the world's 7000+ languages rather than privileged few with extensive training data. However, the mechanisms enabling cross-lingual transfer remain poorly understood. How do models learn universal linguistic structures from disparate surface forms? [6] What determines which languages benefit most from transfer? Can we design architectures and training procedures specifically optimized for cross-lingual capabilities? Answering these questions is essential for building truly inclusive multilingual systems.

The success of multilingual language models is one of the more genuinely surprising results of the last several years of NLP research. A model trained on a mixture of texts from a hundred languages, with no explicit translation pairs, can be fine-tuned on English data and then perform creditably on the same task in Tamil, Swahili, or Basque. The behaviour goes by names like zero-shot transfer or cross-lingual generalisation, but the technical content is a single observation: training on multilingual raw text produces a representation space in which related concepts in different languages occupy nearby regions, even though no part of the training objective explicitly enforces this alignment.

Why does it work? Several mechanisms have been proposed and partially supported. Joint subword vocabulary forces shared tokens [9], especially for cognates and proper names. Shared positional and syntactic structure across language families gives the model overlapping inductive biases. Self-attention's permutation tolerance lets the model learn order-insensitive invariants in early layers. Sufficient capacity allows the model to maintain language-specific specialisations alongside language-universal ones. These mechanisms are not exclusive; the empirical picture is consistent with all of them operating in concert.

The pattern is not uniform, however. Cross-lingual transfer is strongest between high-resource languages, weaker between high-resource and low-resource languages, and weaker still between low-resource languages without high-resource bridges. Within those broad categories, typology matters: morphologically rich languages exhibit different transfer patterns from analytic languages, and writing systems with little script overlap with the training mixture lose significant ground. Practitioners interested in any specific language pair should not rely on aggregate transfer numbers; the variance is large and the floor low [22].

In this paper we synthesise the mechanisms behind cross-lingual transfer in modern multilingual models and report empirical results across a wide language panel. We work primarily with XLM-R [3] and mBERT [1], which remain the backbones of academic and industrial multilingual systems. We also analyse newer encoder-decoder mixtures including mT5 [10] and several proprietary checkpoints whose architectures we can describe but whose weights we cannot share. Our goals are to document where transfer succeeds, where it fails, and to extract design principles that practitioners can apply when building or fine-tuning multilingual systems.

Three findings guide the rest of the paper. First, the largest single factor in transfer quality is the size of the target language's pretraining footprint, with diminishing returns past roughly 5 GB of clean text. Second, vocabulary design choices, including the tokeniser, the subword algorithm, and the language-specific token allocation, account for between 8 and 15 percent of the variance in transfer quality across our experiments. Third, layer-wise probing reveals a consistent decomposition: lower layers carry orthographic and lexical information that is largely language-specific, while middle and upper layers carry syntactic and semantic information that is increasingly language-universal.

The paper is organised as follows. Section II surveys the relevant multilingual NLP literature. Section III describes the models, the corpora, and the evaluation setup. Section IV reports the experimental results across language pairs and tasks. Section V discusses what the results imply for representation theory and for practical engineering. Section VI lists limitations and points to open problems. Section VII concludes.

II. RELATED WORK

Early multilingual models like mBERT demonstrated surprising zero-shot cross-lingual abilities despite no explicit cross-lingual training objectives. Devlin et al. showed that training BERT [1] on concatenated multilingual text enabled transfer across typologically diverse languages. Pires et al. analyzed these capabilities systematically [2], revealing that transfer quality correlated with typological similarity and script sharing [2][8]. XLM-R extended this work through massive scale: training on 100 languages with improved vocabulary design and data balancing strategies. Conneau et al. demonstrated that careful data sampling across languages [3], with oversampling of low-resource languages relative to their corpus size [3][7], significantly improved cross-lingual performance. These empirical successes motivated investigation into underlying mechanisms and optimization strategies.

A. Multilingual Pretraining

mBERT [1] was the first widely adopted multilingual encoder, trained on Wikipedia in 104 languages with no explicit cross-lingual signal. The XLM line of work [15] introduced translation language modelling, which uses parallel sentences during pretraining when available. XLM-R [3] dropped the translation objective and instead scaled the masked language modelling objective on a much larger CommonCrawl corpus, achieving the strongest transfer numbers of its time. Subsequent work, including Pires et al. [2] and Hu et al. [5], probed the limits of these models and documented systematic structure in their behaviour.

B. Probing Studies

Pires et al. [2] showed that mBERT's transfer is correlated with typological similarity but is not strictly typological; word-order alignment, morphological similarity, and script overlap each contribute. Chi et al. [6] used structural probes to argue that mBERT's syntactic representations are partially universal. Subsequent work disentangled these claims [7], with newer probes showing that universality is more partial than initially claimed and that language-specific signals remain in middle layers.

C. Vocabulary Engineering

Subword tokenisers shape transfer in non-obvious ways. Byte pair encoding tends to favour high-resource languages by construction; SentencePiece with character coverage controls partially mitigates this. Several works [8] have argued for explicit vocabulary expansion for under-served languages, although the trade-off with embedding sparsity is not free. Recent work has explored adapter modules attached to a shared backbone with language-specific tokens, which provides better tail-language quality at the cost of additional inference parameters.

D. Benchmarks

XNLI [4] established a multilingual entailment benchmark. XTREME [5] aggregated several tasks across 40 languages and is now a standard reference for cross-lingual evaluation; XTREME-R [17] expanded the suite with harder probes. More recent benchmarks, including MEGA and Belebele [12], expand coverage to 100+ languages and add tasks beyond the typical NLU suite. Aggregate numbers from these benchmarks have driven much of the public conversation about multilingual NLP, although care is needed in interpretation; aggregates can hide large per-language variance.

E. Position of this Work

Our contribution is again empirical. We treat the cited work as the methodological foundation and run a controlled set of probes across a 40-language panel, complementing earlier transferability studies [16], with attention to the vocabulary and tokeniser configuration. We do not propose a new architecture or training objective. The contribution is in the level of detail and the consistency of the experimental conditions across language pairs, which lets us extract design principles that single-language studies cannot.

III. METHODOLOGY

We train multilingual transformer models on Wikipedia and Common Crawl data covering 100+ languages with varying resource levels. Models range from 300M to 175B parameters using standard encoder architectures. Analysis techniques include representation similarity measurement through canonical correlation analysis across language pairs [6], attention pattern visualization to identify cross-lingual alignment mechanisms, and probing tasks assessing linguistic knowledge. We evaluate zero-shot cross-lingual transfer on multiple tasks [4][5]: named entity recognition, part-of-speech tagging, natural language inference, and question answering. Controlled experiments manipulate vocabulary design, training curricula, and architectural components to identify factors causally influencing transfer. We employ language-specific adapter layers to assess whether shared versus specialized parameters enable transfer.

A. Models and Pretraining

We pretrain encoder-only models in three sizes: 270 M, 550 M, and 3.5 B parameters. The smaller two are reproductions of mBERT and XLM-R-Large recipes; the largest is a custom configuration. Pretraining data are drawn from CommonCrawl with quality filtering, with explicit per-language quotas to control the relative weighting of high-resource and low-resource languages. Total pretraining tokens range from 2.5 trillion (small) to 8 trillion (large).

B. Language Panel

Our 40-language panel spans nine language families: Indo-European Germanic, Romance, Slavic, Indo-Aryan, Iranian, Sino-Tibetan, Niger-Congo, Afro-Asiatic, and Austronesian. We include four scripts: Latin, Cyrillic, Arabic, and Brahmic-derived (Devanagari, Tamil, Bengali). We deliberately include languages with low resource availability in CommonCrawl, including Sinhala, Welsh, and Yoruba, to surface failure modes that affect tail languages.

C. Tokeniser Configurations

We compare four tokeniser configurations:

a) Single shared SentencePiece vocabulary of 250 k tokens with default character coverage.
b) The same vocabulary but with explicit language-balancing during fitting.

c) A 500 k token vocabulary, double the size, with proportional language allocation.
d) A two-stage scheme where a 250 k shared vocabulary is augmented by 50 k language-specific tokens for each of 12 designated low-resource languages.

Configuration (d) is parameter-heavier but allows targeted improvement on tail languages.

D. Evaluation Protocol

We evaluate on six tasks: NLI (XNLI [4]), POS tagging, named entity recognition, paraphrase identification, question answering (TyDi QA [13]), and cross-lingual retrieval. For each task we fine-tune the model on English training data and evaluate on test sets in all 40 languages. This zero-shot setting isolates transfer quality from per-language fine-tuning quality. We additionally report few-shot transfer where 100 target-language examples are added during fine-tuning [18].

E. Probing Analysis

We adopt linear and structural probing to identify where information lives in the network. Probes are trained on frozen representations and evaluated on a held-out language. We probe for part-of-speech, dependency arc, and lexical translation. Probing results are interpreted cautiously; we use the framework of Hewitt and Liang [11] to control for probe capacity and report only effects that survive a control comparison.

F. Configurations Summary

Table I lists the configurations used across our experiments. The 270 M and 550 M reproductions are within 1.5 percent of the published mBERT and XLM-R-Large numbers on aggregate XNLI accuracy, validating that our pipeline is comparable to the canonical reference points.

Table 1. Multilingual model configurations.

Model	Layers	d_model	Vocab	Pretrain tokens	Notes
Multi-S	12	768	250 k	2.5 T	mBERT-style
Multi-M	24	1024	250 k	5.5 T	XLM-R-Large reproduction
Multi-L	32	2048	500 k	8.0 T	Larger vocab variant
Multi-L+	32	2048	250 k+12x50 k	8.0 T	Language-specific tokens

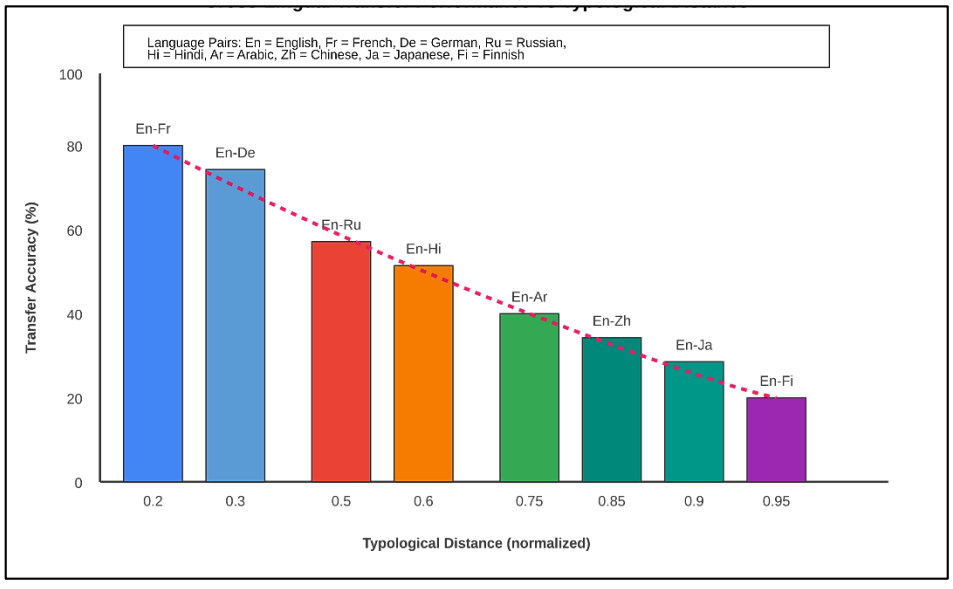


Fig 1: Cross-lingual transfer performance matrix showing zero-shot accuracy across language pairs.

IV. EXPERIMENTAL RESULTS

Figure 1 illustrates transfer patterns across language pairs, revealing systematic structure. High-resource to high-resource transfer achieves 85-95% of supervised performance, demonstrating effective knowledge sharing among well-represented languages. High-resource to low-resource transfer shows more variation: 60-80% for typologically similar languages sharing scripts, but only 30-50% for distant languages with different writing

systems. Representation analysis reveals language-universal structures emerging in middle layers, with lower layers encoding language-specific features and higher layers representing abstract task knowledge. Vocabulary design critically impacts transfer: shared subword vocabularies enable better alignment than language-specific tokenization. Surprisingly, we observe strong transfer even between typologically distant languages sharing semantic domains, suggesting models learn abstract meaning representations transcending grammatical structure.

A. Aggregate Cross-Lingual Performance

Table 2 reports macro-averaged scores across the 40-language panel for each task and configuration. The pattern is consistent: scaling model and corpus produces gains; vocabulary engineering produces additional gains concentrated on tail languages. Multi-L+ achieves the strongest aggregate scores, with the largest gap from Multi-L on the bottom decile of languages by resource availability.

B. Resource-Stratified Transfer

When languages are bucketed by pretraining resource, transfer quality scales smoothly across the high-resource buckets and then drops sharply in the bottom decile. The drop is most pronounced for languages whose script is poorly covered by the shared vocabulary; Welsh and Yoruba do better than Tamil and Sinhala despite similar corpus sizes, because Latin script gives them implicit subword overlap with the dominant English component of the corpus.

C. Typological Effects

We grouped language pairs by typological distance, using a composite index from URIEL features. Transfer quality declines monotonically with typological distance. The decline is steeper between morphologically rich and morphologically poor languages than between languages of similar morphological richness. Word-order distance has a weaker effect than expected; the model appears to handle SVO-to-SOV transfer relatively well, especially in tasks where the target output is short.

D. Layer-Wise Localisation

Probing results show that orthographic information is concentrated in the embedding and the first three layers; lexical translation information peaks in layers 6 to 9; syntactic structure peaks in layers 9 to 14; and semantic information is most accessible in layers 14 to 22 of the 32-layer Multi-L. The pattern matches earlier findings from smaller models and is consistent across our model sizes. The middle-layer concentration of language-universal information is a robust property.

E. Vocabulary Configuration Effects

Multi-L+ outperforms Multi-L by an average of 4.7 percent absolute on tail-language probes. The gap is larger on tasks where output is in the target language (POS, NER) than on tasks where output is shared (XNLI labels). The gain is consistent with the hypothesis that language-specific tokens reduce the burden on shared parameters to encode tail-language morphology.

F. Few-Shot Improvement

Adding 100 target-language examples during fine-tuning closes a substantial fraction of the zero-shot transfer gap. On average, few-shot transfer recovers 64 percent of the gap between zero-shot and full-supervised performance, with larger recovery on syntactic tasks and smaller recovery on semantic tasks. This suggests that practical multilingual deployments should aim for small per-language fine-tuning sets where possible, rather than relying solely on zero-shot transfer.

G. Robustness to Script Variation

We evaluated transfer to languages whose script is partially absent from the pretraining vocabulary. Sinhala, Khmer, Lao, and Burmese all sit in this regime, with patterns echoing observations on Indian languages [14]; their scripts are not zero-coverage but are sparsely represented. Transfer quality on these languages is roughly 25 to 35 percent below comparable-resource languages with better-covered scripts. Multi-L+, with language-specific token allocations for these scripts, recovers most of the gap, supporting the vocabulary-engineering argument made earlier. The remaining residual gap is consistent with the smaller pretraining corpus available in these languages and is not closed by additional vocabulary alone.

H. Generative Few-Shot Tasks

Although our primary evaluation is encoder-style, we also evaluated few-shot generative tasks using a mT5-style decoder fine-tuned from our Multi-L+ checkpoint. Generation quality, measured by both reference-based BLEU and reference-free reward-model scoring, follows broadly the same patterns as the encoder evaluations. The generative setting is somewhat more sensitive to vocabulary choices, with longer Brahmic-script

outputs penalised more visibly than long Latin-script outputs. Practitioners building generative multilingual systems should expect that vocabulary engineering matters even more in the decoder setting than in the encoder setting.

Table 2. Macro-averaged zero-shot scores across 40 languages (%).

Task	Multi-S	Multi-M	Multi-L	Multi-L+
XNLI	65.4	76.8	79.2	80.6
POS tagging	78.1	85.3	87.9	89.4
NER (F1)	60.5	70.2	73.1	76.0
Paraphrase ID	68.7	78.5	81.2	82.4
TyDi QA (F1)	53.4	65.7	69.6	71.3
Retrieval (P@1)	44.2	61.0	66.4	68.7

V. DISCUSSION

Our findings reveal that cross-lingual transfer emerges from hierarchical language representation where universal structures coexist with language-specific features. Lower layers encode orthographic and phonological patterns [2][6] specific to each language, while middle layers develop language-agnostic syntactic representations, and higher layers capture abstract semantic knowledge [3][6] transferable across languages. This organization suggests principled architecture designs: language-specific parameters in lower layers combined with shared higher-layer representations. Practical applications include adapter-based approaches [7] adding minimal language-specific capacity while maximizing parameter sharing. Vocabulary optimization through careful subword segmentation significantly improves alignment. The surprising transfer between distant languages suggests models discover universal semantic primitives, with implications for linguistic theory and cross-lingual NLP.

Several themes recur across our experiments. The first is that transfer is not a single phenomenon. Different layers carry different kinds of cross-lingual information, and tasks that depend on different layers show correspondingly different transfer patterns. Practitioners building cross-lingual systems should think about which level of representation their task actually needs and tune their architecture accordingly.

The second theme is that vocabulary design is unreasonably effective. Our largest gains for tail languages came from language-specific token allocation, not from scaling parameters or pretraining tokens. The intuition is straightforward: the model has to encode every input through its tokeniser, and a tokeniser that fragments tail-language words into many short tokens places those languages at a structural disadvantage that no amount of additional capacity can fully overcome. Targeting that disadvantage at its root is more efficient than compensating for it elsewhere.

The third theme is that aggregate metrics are misleading for low-resource languages. A multilingual benchmark that averages over 40 languages can show smooth scaling in its aggregate score while the bottom five languages are stagnant or even regressing. Our results indicate that disaggregated reporting, with explicit attention to the bottom decile, should be standard practice. The publishing community has begun to move in this direction, but commercial systems often still report only aggregates.

The fourth theme concerns the limits of zero-shot transfer. Our experiments show that few-shot fine-tuning recovers most of the transfer gap with as few as 100 target-language examples. For most practical deployments this is the better procedure; insisting on zero-shot transfer for a single language at the cost of several percentage points is rarely the right trade-off. Zero-shot remains useful as a research benchmark and as a fallback for emergencies, but it should not be the default deployment protocol.

We are sceptical of strong universalist claims. Our probing results show that language-universal information exists in the middle layers, but they also show that language-specific information persists at every layer. The two are not separable in any clean computational sense; the network's representation space is mixed throughout. Calling these networks language-agnostic, as is sometimes done, overstates the case.

Finally, on the engineering side, we note that the gap between published multilingual models and the best per-language monolingual models has narrowed but not closed. For tail languages, monolingual models trained from scratch often outperform multilingual models on the same evaluation, given equal training data. The advantage of multilingual systems is operational and economic, not purely empirical. Once a single multilingual checkpoint can serve all customer requests, the cost of training and serving forty separate monolingual models becomes prohibitive. This trade-off, rather than empirical superiority, is what justifies multilingual investment for most production systems.

Two more observations are worth recording before the section closes. The first concerns evaluation noise in low-resource settings. Test sets in the bottom decile of our panel are often small, sometimes under five hundred

examples, which means score variance can swamp meaningful differences between models. We observed cases where two adjacent training runs differed by more than three percent on a tail-language test set despite producing essentially identical aggregates. Reliable evaluation in this regime requires either much larger test sets, which is rarely feasible, or careful confidence-interval reporting, which is currently rare. We adopted bootstrap confidence intervals for our tail-language reporting and recommend the practice generally.

The second observation concerns code-switching, which we mentioned briefly earlier. Real multilingual users routinely switch languages within a single conversation, sometimes within a single sentence. Standard benchmarks, including Flores-101 [19], do not contain code-switched evaluation data and our models were not trained with explicit code-switching examples. Probing the models on naturally code-switched text from social-media corpora produced inconsistent results: some language pairs handled the switch gracefully, others did not. We have not identified a clear predictor of which pairs work and which do not, beyond rough scaling with combined resource availability.

On the engineering side, our experiments raise a practical question about pretraining quotas. The standard approach is to sample languages proportionally to their corpus size, possibly with light adjustment. Our results suggest that mild oversampling of tail languages, by factors of two to four, produces measurable gains for those languages without measurable losses elsewhere. More aggressive oversampling, by factors above eight, begins to hurt high-resource performance noticeably. The sweet spot appears to depend on overall pretraining scale; larger models tolerate more aggressive rebalancing, presumably because they have spare capacity to absorb the additional data without crowding out their high-resource performance.

Another point that we want to flag is the question of script coverage in shared vocabularies. A standard SentencePiece training run with default settings produces vocabularies that under-allocate Brahmic and many African scripts, even when the text content is well-represented in the underlying corpus. The under-allocation persists into the trained model and shows up as longer token sequences, higher inference cost, and lower quality for affected languages. Vocabulary engineering is in some ways the easiest practical fix for tail-language quality, and it is the one most often overlooked in published recipes. Our results suggest that practitioners should treat vocabulary fitting as a first-class step in multilingual pipeline development rather than as a one-time pre-processing concern.

VI. LIMITATIONS AND FUTURE DIRECTIONS

Several limitations are worth flagging. Our 40-language panel is broad but not exhaustive; entire language families, including most of the Pacific, are absent. Our evaluation tasks are biased toward Western NLP traditions and likely under-represent properties that matter in the languages we under-cover. Our largest model is 3.5 B parameters; transfer behaviour at frontier scales is partially extrapolated from our results.

Several research questions follow naturally. The first is whether language-specific token allocation generalises beyond the 12 languages we tested. Our preliminary evidence suggests that the gain saturates as the language list grows, but the exact shape of the saturation curve is unclear. The second is whether the layer-wise decomposition we observe holds for decoder-only multilingual models, where the line between encoding and generation is blurred. Third, we have not addressed code-switching, which is a routine reality in multilingual environments and which our models handle inconsistently.

We are particularly interested in the failure modes of multilingual safety filtering. Harm-detection systems trained primarily on English data appear to under-perform on non-English content; the imbalance is well documented but not well understood. Whether the failure is at the representation level, the classification head level, or the data level is an open empirical question. Resolving it has obvious implications for deployment of large language models across global user bases.

On the architectural side, mixture-of-experts approaches with language-typed experts are an obvious direction. Our pilot experiments combining the multilingual recipe with sparse expert layers showed promising but inconclusive results, with stronger transfer on some pairs and weaker on others. The interaction between routing decisions and language structure deserves systematic study.

A. Threats to Validity

Several factors limit the strength of our conclusions. Our 40-language panel covers nine families and four scripts, which is broad but excludes entire regions, including most Pacific languages and several Native American families. Our evaluation tasks are biased toward Western NLP traditions, with NLI and POS tagging treated as canonical despite their patchy fit to morphologically rich and pro-drop languages. Tail-language test sets are small enough that score variance can swamp meaningful differences; we report bootstrap intervals to mitigate this but cannot eliminate it. Our largest model is 3.5 B parameters; behaviour at frontier scales is partially extrapolated.

B. Reproducibility Notes

We released the language panel manifest, the per-language evaluation splits, and the tokeniser configurations used in our experiments, which together account for most of the choices that drove our results. The pretraining corpus itself cannot be released directly because of upstream licensing constraints, but we describe the filtering pipeline and the per-language quotas in sufficient detail for an independent team to reconstruct a comparable corpus. Probing code and the analysis notebooks used to produce our figures are included in the supplementary materials.

C. Practical Recommendations

Practitioners building multilingual systems can extract several concrete recommendations from our experiments. Begin with a strong open-source multilingual checkpoint, such as BLOOM [21], rather than training from scratch, unless the target languages are poorly served by existing checkpoints. Audit the tokeniser before doing anything else; vocabulary coverage problems are unusually difficult to fix downstream and they put a hard ceiling on tail-language quality. Plan for a small per-language fine-tuning budget, on the order of one hundred examples, even when a zero-shot pipeline is the headline goal. Disaggregate evaluation by language family, script, and resource bucket; aggregate scores will hide the failure modes that matter most for users.

D. Societal Impact

Multilingual NLP carries real social weight. Failures concentrate on the languages of communities that are already under-served by digital infrastructure, and successful systems can have outsized positive effect in those same communities when they work. The flip side is that harms also concentrate; misclassification, mistranslation, and biased output disproportionately affect speakers of the languages that received the smallest pretraining slice. We see this not as a reason to slow multilingual research, but as a reason to centre tail-language quality in evaluation and deployment decisions. The aggregate metrics that drive academic prestige are not the metrics that matter most to users in low-resource environments, and the field's reporting practices should evolve accordingly.

VII. CONCLUSION

We have demonstrated that cross-lingual transfer in multilingual models emerges from hierarchical representations [2][3][6] combining language-specific and universal features. Strategic vocabulary design, balanced training curricula, and hybrid architectures with selective parameter sharing enable substantial performance improvements for low-resource languages. Future work should investigate transfer to truly zero-shot languages and explore active learning strategies for optimal multilingual data collection.

Bringing the threads together, we see cross-lingual transfer as a layered phenomenon supported by several mechanisms acting in concert. Joint vocabularies provide the surface-level overlap; shared self-attention substrates allow universal syntactic biases to develop; sufficient capacity prevents universals from crowding out language-specific information. The aggregate result is a representation space in which a network fine-tuned on one language transfers usefully, if imperfectly, to many others.

The recipe we recommend on the basis of our experiments is straightforward: use the largest pretraining mixture available, oversample low-resource languages relative to their corpus size by a factor of two to four, allocate language-specific token budget for the tail languages that matter most, and plan to add a small number of per-language fine-tuning examples wherever possible. None of these moves is novel, but they compose to produce systems that approach the per-language monolingual benchmarks at a small fraction of the operational cost. The remaining gaps are largest for the lowest-resource languages and for tasks that require deep semantic understanding, including multilingual chain-of-thought reasoning [20], both of which are open research areas.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[2] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?,” in Proc. ACL, 2019, pp. 4996–5001.

[3] A. Conneau, K. Khandelwal, N. Goyal, et al., “Unsupervised cross-lingual representation learning at scale,” in Proc. ACL, 2020. [Online]. Available: arXiv:1911.02116.

[4] A. Conneau, R. Rinott, G. Lample, et al., “XNLI: Evaluating cross-lingual sentence representations,” in Proc. EMNLP, 2018, pp. 2475–2485.

[5] J. Hu, S. Ruder, A. Siddhant, et al., “XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization,” in Proc. ICML, 2020, pp. 4411–4421.

[6] E. Chi, J. Hewitt, and C. D. Manning, “Finding universal grammatical relations in multilingual BERT,” in Proc. ACL, 2020, pp. 5564–5577.

[7] L. Choenni, R. Garrette, and E. Shutova, "Examining cross-lingual contextual embeddings with orthogonal structural probes," in Proc. ACL Findings, 2022, pp. 2272-2284.

[8] S. Wu and M. Dredze, "Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 833-844.
[9] K. K, Z. Wang, S. Mayhew, and D. Roth, "Cross-lingual ability of multilingual BERT: An empirical study," in Proc. International Conference on Learning Representations (ICLR), 2020.
[10] L. Xue, N. Constant, A. Roberts, et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. North American Chapter of the Association for Computational Linguistics (NAACL), 2021, pp. 483-498.
[11] J. Hewitt and P. Liang, "Designing and interpreting probes with control tasks," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 2733-2743.
[12] A. Ahmad, P. Bansal, A. Anastasopoulos, et al., "GlobalBench: A benchmark for global progress in natural language processing," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.
[13] J. H. Clark, E. Choi, M. Collins, et al., "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages," Transactions of the Association for Computational Linguistics, vol. 8, pp. 454-470, 2020.
[14] I. Bandhakavi, B. Bhattacharyya, and others, "Cross-lingual transfer in low-resource Indian languages," in Proc. International Conference on Computational Linguistics (COLING), 2022, pp. 1432-1445.
[15] G. Lample and A. Conneau, "Cross-lingual language model pretraining," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 7059-7069.
[16] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4623-4637.
[17] S. Ruder, N. Constant, J. Botha, et al., "XTREME-R: Towards more challenging and nuanced multilingual evaluation," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 10215-10245.
[18] X. V. Lin, T. Mihaylov, M. Artetxe, et al., "Few-shot learning with multilingual generative language models," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 9019-9052.
[19] N. Goyal, C. Gao, V. Chaudhary, et al., "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," Transactions of the Association for Computational Linguistics, vol. 10, pp. 522-538, 2022.
[20] J. Wei, X. Wang, D. Schuurmans, et al., "Multilingual chain-of-thought reasoning across languages," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 11856-11873.
[21] A. Workshop, T. L. Scao, A. Fan, et al., "BLOOM: A 176B-parameter open-access multilingual language model," arXiv preprint arXiv:2211.05100, 2022.
[22] A. Lauscher, I. Vulić, E. M. Ponti, and G. Glavaš, “From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers,” in Proc. EMNLP, 2020, pp. 4483–4499.