Bridging the Babel: A Comparative Analysis of Cross-Lingual Models for Multilingual Applications

Abhishek Kumar, Lead Researcher, Valentis

The Landscape of Cross-Lingual Representation

The proliferation of digital information across a multitude of languages has created an urgent need for Natural Language Processing (NLP) systems that can operate beyond monolingual boundaries. This has catalyzed the development of cross-lingual models, which are designed to understand, process, and generate text in multiple languages, often by leveraging knowledge from high-resource languages to enhance capabilities in low-resource ones. This paradigm, known as cross-lingual transfer learning, represents a fundamental shift from training separate, isolated models for each language to creating unified systems that capture universal linguistic properties.[1] The core value of this approach lies in its potential to democratize access to advanced NLP technologies, particularly for the thousands of languages that lack the vast annotated datasets required for traditional model training.[2] At the heart of this capability are models pre-trained on massive multilingual corpora, which learn to map text from different languages into a shared, language-agnostic representational space, thereby enabling the transfer of learned skills across linguistic divides.[1]

Paradigms of Cross-Lingual Transfer

The practical application of cross-lingual models is realized through several distinct learning paradigms, each with its own set of trade-offs between data requirements, computational cost, and performance.

Zero-Shot Cross-Lingual Transfer (ZS-XLT)

Zero-shot cross-lingual transfer is the most direct application of a multilingual model. In this setup, a model is fine-tuned for a specific task using labeled data from a single source language (typically English) and is then applied directly to perform the same task on target languages for which it has seen no labeled examples.[5] The surprising effectiveness of early models like multilingual BERT (mBERT) in ZS-XLT settings was a pivotal moment in multilingual NLP, demonstrating that a model trained without explicit cross-lingual supervision could nonetheless generalize across languages.[6] However, subsequent research has revealed the fragility of this paradigm. ZS-XLT performance often exhibits high variance and is highly dependent on the typological proximity between the source and target languages; performance degrades substantially when transferring to languages that are structurally or genetically distant from the source.[8] This approach is particularly challenged by tasks that involve significant linguistic specificity or cultural nuances, such as hate speech detection, where the performance gap can be considerable.[3]

Few-Shot Cross-Lingual Transfer (FS-XLT)

Recognizing the limitations of the zero-shot approach, few-shot cross-lingual transfer has emerged as a more robust and practical alternative. FS-XLT involves a two-stage fine-tuning process: first, the model is trained on abundant source-language data, and second, it is further fine-tuned on a very small number of labeled examples—often as few as 10 to 100—from the target language.[8] This minimal investment in target-language annotation can yield dramatic performance improvements, often closing a significant portion of the gap with fully supervised models at a negligible cost.[8] For instance, on syntactic tasks like Named Entity Recognition (NER), the addition of a few shots can improve F1 scores by as much as 20 points over a zero-shot baseline.[8] This effectiveness stems from the model's ability to use the few target examples to slightly adjust its internal representations, effectively "calibrating" its latent cross-lingual alignment for the specific task and language. However, this paradigm is not without its challenges; research has shown that FS-XLT performance can be highly sensitive to the selection of the specific few-shot examples, which introduces a significant source of variance and complicates fair and reproducible comparisons between different models and methods.[8]

Prompting and In-Context Learning

With the advent of extremely large language models, a new paradigm of prompting, or in-context learning, has become viable. Instead of updating the model's weights through fine-tuning, this method provides the model with a natural language prompt that includes a task description and a few examples ("shots") of the task being performed. The model is then expected to complete a new, unseen example by leveraging the context provided in the prompt.[11] This is a highly compute-efficient approach as it avoids the need for gradient-based training. Studies have shown that for certain tasks and models, prompting can outperform traditional fine-tuning in few-shot settings, offering a cost-effective method for adapting models to new languages and tasks.[12]

Advanced Transfer Techniques

Beyond these primary paradigms, researchers have developed specialized techniques to further enhance the quality of cross-lingual alignment and transfer.

Code-Switching (CS)

Code-switching is a data augmentation technique designed to explicitly improve the alignment of representations between languages. It involves taking a sentence in the source language and replacing some of its words with their translations in the target language, typically sourced from a bilingual dictionary.[5] This creates synthetic mixed-language sentences that force the model to map words with similar meanings to nearby points in its embedding space. While generally effective, uncontrolled or excessive code-switching can be detrimental, as it may introduce grammatically nonsensical or contextually confusing "dirty samples" that degrade model performance. To address this, more sophisticated methods like Progressive Code-Switching (PCS) have been developed. Inspired by curriculum learning, PCS gradually introduces code-switched examples of increasing difficulty, allowing the model to first learn from simpler, more reliable alignments before moving on to more complex ones, thereby improving the quality of the learned representations.[5]

Translate-Train and Translate-Test

Two other strategies that leverage machine translation are "translate-train" and "translate-test." In the translate-train approach, the entire source-language training dataset is translated into the target language, and the model is then fine-tuned on this translated data. In the translate-test approach, the target-language test data is translated into the source language, and inference is performed using the original source-trained model. The choice between these methods is highly dependent on the quality of the available machine translation systems and the specific characteristics of the task and languages involved.[12]

The evolution of these paradigms reveals a pragmatic trajectory in the field. The initial excitement surrounding the "magic" of pure zero-shot transfer has been tempered by a deeper understanding of its limitations. While ZS-XLT demonstrates that multilingual models learn a powerful, latent cross-lingual space, this space is not perfectly aligned. The remarkable success of FS-XLT shows that this latent alignment can be activated and refined with a surprisingly small amount of target-specific signal. This suggests that for building reliable, real-world multilingual applications, a strategy that incorporates at least a minimal amount of target-language supervision is far more robust than one that relies solely on the often-fragile promise of zero-shot performance.

Foundational Architectures: A Tale of Two Models

The modern era of cross-lingual learning was inaugurated by the development of large, Transformer-based multilingual models. Among these, Multilingual BERT (mBERT) stands as the pioneering architecture, while XLM-RoBERTa (XLM-R) represents a critical evolutionary step that scaled up the pre-training paradigm to achieve a new state-of-the-art. A comparative analysis of these two models reveals crucial lessons about the ingredients for successful cross-lingual representation learning.

mBERT: The Pioneer of Multilingual Transformers

Released as part of the original BERT suite, Multilingual BERT was the first model to demonstrate the surprising effectiveness of joint pre-training on monolingual corpora for cross-lingual transfer tasks.

- **Architecture:** mBERT employs the standard BERT-base architecture: a 12-layer Transformer encoder with 12 self-attention heads and a hidden state size of 768.[7] This results in a model with approximately 178 million parameters. Notably, due to its large, shared vocabulary, over half of these parameters (more than 92 million) are allocated to the token embedding layer, a design choice that has significant implications for model capacity.[15]
- **Vocabulary:** It utilizes a single WordPiece vocabulary of approximately 119,000 tokens that is shared across all 104 languages.[15] This shared vocabulary was initially thought to be a primary mechanism for its cross-lingual capabilities, as identical subwords in related languages (e.g., "nation" in English and French) would share the same embedding, creating anchor points for aligning the representation spaces.
- **Training Data:** mBERT was pre-trained on the Wikipedia dumps for the 104 languages with the largest Wikipedias at the time of its creation.[18] The choice of languages was based purely on data availability, resulting in a dataset with a severe imbalance; the English Wikipedia, for instance, is orders of magnitude larger than the corpus for a low-resource language like Yoruba.[21] To mitigate this, the training procedure implemented a sampling strategy that up-sampled data from low-resource languages and down-sampled from high-resource ones in an attempt to create a more

balanced exposure during training.[18]

- **Pre-training Objectives:** mBERT was trained using the same two objectives as the original monolingual BERT: Masked Language Modeling (MLM), where the model predicts randomly masked tokens, and Next Sentence Prediction (NSP), a binary classification task to predict if two sentences are consecutive.[14]

Despite its groundbreaking success, subsequent analysis revealed several fundamental limitations. The most significant is the "curse of multilinguality," where the model's fixed capacity is divided among many languages, leading to under-modeling for all of them. Consequently, mBERT's performance on high-resource languages like English is typically inferior to that of a monolingual BERT model of the same size.[19] Furthermore, its reliance on Wikipedia meant that for many low-resource languages, the pre-training data was scarce and of variable quality, leading to poor representations and a dramatic drop in performance.[9] Later research also demonstrated that the NSP objective, once thought to be crucial for learning sentence relationships, was often ineffective or even detrimental to downstream performance.[22]

XLM-R: Scaling Up for a New State-of-the-Art

XLM-RoBERTa was developed to address the key limitations of mBERT by scaling up the training data and refining the pre-training methodology, drawing lessons from the successful monolingual RoBERTa model.

- **Architecture:** XLM-R is also a Transformer encoder-only model but is based on the improved RoBERTa architecture. It was released in base (12 layers, 270M parameters) and large (24 layers, 550M parameters) variants.[25] Key architectural refinements include the complete removal of the NSP objective, focusing solely on MLM, and the use of dynamic masking, where the masking pattern is generated anew for each training epoch.[14] A crucial design choice inherited from the original XLM model was the removal of explicit language embeddings, which forces the model to identify the language from the input tokens themselves and improves its ability to handle code-switched text.[28]
- **Vocabulary:** XLM-R employs a significantly larger shared vocabulary of 250,000 tokens, created using SentencePiece.[25] This larger vocabulary provides better coverage for a wide range of languages, reducing the frequency of out-of-vocabulary words and leading to more efficient tokenization.
- **Training Data:** The most significant departure from mBERT was the training corpus. Instead of Wikipedia, XLM-R was pre-trained on the newly created CC100 dataset, a massive 2.5 terabyte corpus of filtered text from the Common Crawl web scrape, spanning 100 languages.[27] For many low-resource languages, CC100 provides orders of magnitude more data than Wikipedia. For example, the Swahili corpus in CC100 is 1.6 GB, compared to just 0.01 GB in Wikipedia, and the Urdu corpus is 5.7 GB versus 0.01 GB.[32] This vast increase in data was the single most important

factor in XLM-R's success.

The impact of these changes was profound. XLM-R established a new state-of-the-art on a wide range of cross-lingual benchmarks and, crucially, was the first multilingual model to demonstrate that it could be highly competitive with strong monolingual models on their own language-specific tasks.[33]

The evolution from mBERT to XLM-R marks a pivotal moment in the understanding of multilingual models. mBERT's success was a proof of concept, showing that cross-lingual capabilities could emerge from joint monolingual training. However, its performance was fundamentally capped by its limited and imbalanced training data. XLM-R, by keeping the core architecture largely intact but radically scaling the quantity and diversity of the training data, demonstrated that the primary bottleneck was not the model design but the data itself. This shifted the focus of multilingual NLP research away from purely architectural innovations and toward the immense challenges of data curation, filtering, and scaling for the world's languages. The success of XLM-R was a clear signal that progress, especially for the long tail of low-resource languages, was fundamentally a data problem.

Comparative Performance Analysis Across Downstream Tasks

A quantitative comparison of mBERT and XLM-R across standardized cross-lingual benchmarks reveals a clear and consistent performance gap. XLM-R's advancements in pre-training data and methodology translate directly into superior performance on a wide array of downstream NLP tasks, from high-level semantic understanding to token-level structured prediction. The following table summarizes the performance of both models on several key benchmarks, illustrating the scale of XLM-R's improvements.

Model	Benchmark	Metric	Average Score	Source Snippets
mBERT	XNLI	Accuracy	66.3%	[35]
XLM-R	XNLI	Accuracy	80.9%	[28]
mBERT	MLQA	F1 Score	57.7%	[38]
XLM-R	MLQA	F1 Score	70.7% (+13%)	[33]

mBERT	CoNLL-NER (Zero-Shot)	F1 Score	78.52%	[32]
XLM-R	CoNLL-NER (Zero-Shot)	F1 Score	80.94% (+2.42%)	[32]
mBERT	TyDi QA-GoldP (Zero-Shot)	F1 Score	56.4%	[38]
XLM-R	TyDi QA-GoldP (Zero-Shot)	F1 Score	65.1%	[39]

Natural Language Inference (XNLI)

The Cross-lingual Natural Language Inference (XNLI) benchmark is a rigorous test of a model's semantic reasoning capabilities, requiring it to determine whether a "premise" sentence entails, contradicts, or is neutral with respect to a "hypothesis" sentence across 15 languages. On this task, the performance difference between the models is stark. XLM-R achieves an average accuracy of 80.9%, a massive improvement of +14.6% over mBERT's 66.3%.[33] This substantial gain is particularly pronounced in low-resource languages, where XLM-R outperforms previous models by 15.7% in Swahili and 11.4% in Urdu.[33] This result underscores the critical role of large-scale, diverse training data in building robust cross-lingual semantic representations.

Extractive Question Answering (MLQA, XQuAD, TyDi QA)

Extractive Question Answering (QA) tasks require a model to identify and extract the specific span of text from a given context that answers a question. This tests both comprehension and localization abilities. On the Multilingual Question Answering (MLQA) benchmark, XLM-R achieves an average F1 score that is 13 percentage points higher than mBERT's.[33] Similarly, on the Typologically Diverse Question Answering (TyDi QA) Gold Passage task, which is specifically designed to challenge models with linguistic phenomena not typically found in English, XLM-R again demonstrates superior generalization. In a zero-shot setting, mBERT scores an average F1 of 56.4% [38], whereas XLM-R achieves 65.1%.[39] These results indicate that XLM-R's richer representations enable it to more accurately understand questions and locate answers across a wider range of linguistic contexts.

Structured Prediction (NER)

Named Entity Recognition (NER) is a fundamental structured prediction task that involves identifying and classifying entities like persons, organizations, and locations in text. While still significant, the performance gap between the models is narrower on this task. In a zero-shot transfer setting on the CoNLL-2002/2003 datasets, XLM-R achieves an average F1 score of 80.94%, a +2.42% improvement over mBERT's 78.52%.[32] This suggests that while both models learn effective representations for this more syntactically-oriented, token-level task, the enhanced quality of XLM-R's embeddings provides a consistent, albeit smaller, advantage.[40]

The XTREME Benchmark

The Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark provides a comprehensive evaluation across nine distinct tasks and 40 typologically diverse languages, offering a holistic view of a model's cross-lingual generalization ability.[37] The results on this benchmark confirm the trends observed in individual tasks. On the official leaderboard, the XLM-R Large model achieves an average score of 68.2, significantly outperforming mBERT, which scores 59.6.[37] This consistent superiority across a wide range of tasks and languages solidifies XLM-R's position as a more powerful and versatile cross-lingual encoder.

The pattern of these results is revealing. The magnitude of XLM-R's performance improvement over mBERT appears to be directly proportional to the semantic complexity of the task at hand. The largest gains are seen in NLI and QA, tasks that demand deep reasoning about meaning and relationships between sentences. The smallest, though still notable, gain is in NER, a task that can often rely more heavily on local context and syntactic cues. This pattern suggests a crucial conclusion: while more and better data improves performance across the board, its most profound impact is on the development of a rich, nuanced, and well-aligned *semantic* representation space. mBERT's training on Wikipedia was sufficient to learn many of the universal syntactic patterns needed for tasks like NER. However, it was the sheer scale and diversity of XLM-R's training on the Common Crawl that was necessary to build the deep semantic understanding required for high-level cross-lingual reasoning. This implies a strategic consideration for practitioners: while mBERT may remain a computationally cheaper and viable option for simpler, token-level tasks in well-supported languages, any application requiring genuine semantic understanding across a diverse language set necessitates the use of more advanced models like XLM-R.

Beyond the Benchmarks: The Impact of Linguistic Diversity

While standardized benchmarks provide essential quantitative comparisons, a deeper understanding of multilingual models requires an analysis of how they perform across the vast landscape of human linguistic diversity. Factors such as language family, typological features, and script play a crucial role in determining the efficacy of cross-lingual transfer, often revealing limitations that are not apparent from aggregate scores alone.

The "Curse of Multilinguality"

A fundamental challenge in designing massively multilingual models is the "curse of multilinguality".[28] This concept describes the inherent trade-off between the number of languages a model supports and its per-language capacity. For a model with a fixed number of parameters, each additional language dilutes the capacity available for every other language. Research on XLM-R demonstrated this effect empirically: as the number of languages in pre-training increased from 7 to 100 for a fixed-size model, performance on low-resource languages initially improved due to positive transfer, but beyond a certain point, the overall performance on all languages began to degrade.[28] This degradation occurs because the model's finite parameters are forced to represent an increasingly diverse set of linguistic phenomena, leading to under-modeling. While this curse can be partially mitigated by significantly increasing the model's parameter count (e.g., using XLM-R Large instead of Base), it remains a key constraint in the development of truly universal language models.[28]

Linguistic Proximity vs. Lexical Overlap

Early hypotheses about mBERT's success centered on the idea that its shared vocabulary was the primary mechanism for transfer, with overlapping subwords acting as anchors to align different languages.[7] However, a compelling body of evidence has since shown that structural and typological similarity between languages is a far more significant factor than simple lexical overlap.[10]

Pioneering experiments demonstrated that mBERT could effectively transfer knowledge between languages with zero shared vocabulary. This was shown by training a model on an artificially created "Fake-English" (where every character was systematically shifted in Unicode to eliminate overlap) and successfully transferring to real languages like Spanish and Hindi.[22] Similarly, strong transfer was observed between Hindi and Urdu, two languages that are structurally very similar but use entirely different scripts (Devanagari and Arabic), resulting in minimal lexical overlap at the subword level.[7] Conversely, when the structural properties of a language, such as word order, were artificially scrambled during pre-training, cross-lingual transfer performance dropped significantly.[22] This body of work strongly suggests that multilingual models learn an abstract, language-agnostic representation of syntax and grammar, which is the true vehicle for knowledge transfer.[6]

This leads to a more profound understanding of how these models function. The success of cross-lingual transfer points to the emergence of a latent "interlingua"—a shared, abstract representational space that captures universal properties of language structure, independent of their surface lexical forms. This interlingua is not explicitly designed but emerges as a consequence of the model's optimization objective (e.g., MLM) across a diverse set of languages. The model discovers that the most efficient way to predict masked words in many different languages is to develop an internal representation that captures their common underlying grammatical principles. However, this emergent interlingua is not perfect or unbiased. Its "native language" is effectively a weighted average of the high-resource languages in its training data. As shown in the table below, this data is heavily skewed towards the Indo-European family. Consequently, transferring knowledge to another Indo-European language is akin to translating between two closely related dialects of this interlingua, a relatively simple task. In contrast, transferring to a typologically distant language from a different family requires a much larger and more difficult representational leap, explaining the performance disparities observed in practice.

Language Family and Script Analysis

The distribution of languages in pre-training corpora is heavily skewed, which directly impacts model performance across different language families. The following table provides a breakdown of the language coverage of mBERT, XLM-R, and mT5, classified by major language families as defined by Ethnologue.[46]

Language Family	# Languages in mBERT (104)	# Languages in XLM-R (100)	# Languages in mT5 (101)
Indo-European	46	55	45
Uralic	4	4	4
Turkic	6	6	6
Afro-Asiatic	4	4	5

Sino-Tibetan	4	3	3
Austronesian	8	8	8
Niger-Congo	3	2	4
Dravidian	4	4	4
Kra-Dai	1	2	1
Austro-Asiatic	1	1	1
Japonic	1	1	1
Koreanic	1	1	1
Other/Isolates	21	9	18

This data makes the imbalance clear: Indo-European languages constitute roughly half of the languages in these models, a massive over-representation compared to their global distribution.[48] This skew is a primary reason why transfer performance is consistently higher between European languages and degrades significantly for typologically distant pairs, such as transferring from English (Indo-European) to Arabic (Afro-Asiatic) or Japanese (Japonic).[44] For languages that were entirely unseen during pre-training, factors like script type and language family become the most crucial predictors of transfer success, as the model must rely on its generalized knowledge of linguistic structures learned from related languages.[50]

The Next Generation: Addressing Foundational Limitations

The successes and shortcomings of mBERT and XLM-R paved the way for a new generation of multilingual models designed to tackle their foundational limitations. Two prominent examples, mT5 and XLM-V, represent distinct but complementary evolutionary paths: mT5 expands the architectural paradigm to new task domains, while XLM-V refines the core representational quality by addressing the critical vocabulary bottleneck.

mT5: The Text-to-Text Multilingual Paradigm

Multilingual T5 (mT5) extends the "Text-to-Text Transfer Transformer" (T5) framework to a massively multilingual context, fundamentally changing the architectural approach from its encoder-only predecessors.

- **Architecture:** Unlike mBERT and XLM-R, mT5 is a full encoder-decoder Transformer model.[51] This architecture is inherently suited for generative tasks. mT5 unifies all NLP problems into a text-to-text format, where the model takes a textual input (e.g., a sentence to be translated, a question to be answered) and is trained to generate a textual output.[54] This flexible framework allows a single model to perform classification, question answering, summarization, and translation without task-specific architectural modifications. The model was released in a range of sizes, from mT5-Small (300 million parameters) up to mT5-XXL (13 billion parameters).[39]
- **Training Data:** mT5 was pre-trained on the multilingual C4 (mC4) corpus, a new dataset created from the Common Crawl web scrape that covers 101 languages.[39] Like CC100, mC4 provides a massive and diverse source of text, far exceeding the scale of Wikipedia.
- **Performance:** At its largest scales, mT5 achieves state-of-the-art performance on many multilingual benchmarks, demonstrating the power of its generative, text-to-text approach.[39] However, its performance is highly dependent on model size. At smaller scales (e.g., mT5-Base), it can underperform similarly-sized encoder-only models like XLM-R on discriminative tasks, likely because its parameter budget is split between both an encoder and a decoder.[56]

XLM-V: Overcoming the Vocabulary Bottleneck

While mT5 focused on architectural evolution, XLM-V addressed a more fundamental representational issue that plagued all previous multilingual models: the vocabulary bottleneck.

- **The Problem:** The "vocabulary bottleneck" refers to the critical limitation that arises from using a relatively small, fixed-size vocabulary (e.g., 250k tokens for XLM-R) to represent over 100 languages.[57] As model parameter counts scaled into the billions, vocabulary size remained stagnant. This forces the model to represent the vast lexical diversity of the world's languages with an average of only a few thousand unique tokens per language. The consequences are severe: frequent and inefficient over-tokenization (representing single words with many subword tokens), which creates longer, harder-to-process sequences, and a poor quality of representation, especially for morphologically rich languages and low-resource languages that are underrepresented in the vocabulary.[57]
- **The Solution:** XLM-V introduces an innovative method for constructing a very large multilingual

vocabulary. Instead of creating a single, homogenized vocabulary, it de-emphasizes token sharing between languages with little lexical overlap. The process involves training monolingual tokenizers, clustering languages based on lexical similarity, and then building a combined vocabulary of 1 million tokens with capacity intelligently allocated to each language cluster to ensure sufficient coverage.[57]

- **Architecture and Data:** Critically, XLM-V uses the exact same XLM-R base architecture and the same CC100 training data.[53] This experimental design isolates the vocabulary as the sole innovation, allowing for a direct measurement of its impact.
- **Performance:** The results are striking. XLM-V outperforms XLM-R on every downstream task it was evaluated on. The gains are particularly dramatic for low-resource language benchmarks, such as an +11.2% F1 improvement on MasakhaNER (an NER task for African languages) and a +5.8% accuracy improvement on AmericasNLI (a natural language inference task for indigenous American languages).[57]

The development of these next-generation models illuminates two distinct and highly valuable directions for the future of multilingual NLP. mT5's success with its encoder-decoder architecture demonstrates the viability of a unified, generative framework for a wide range of cross-lingual tasks, significantly expanding the *scope* of what a single multilingual model can do. In parallel, XLM-V's breakthrough performance proves that the *quality* of the underlying representations in existing models was being severely constrained by a flawed vocabulary design. By showing that a better vocabulary alone—with no changes to architecture or training data—could unlock such significant gains, particularly for historically marginalized languages, XLM-V identified a more fundamental bottleneck. This suggests that the next major leap in multilingual model performance will likely come from a synthesis of these two approaches: a model that combines a flexible encoder-decoder architecture with a scaled and intelligently constructed vocabulary.

The Sociotechnical Dimension: Cultural and Social Bias

Multilingual language models are not purely technical artifacts; they are sociotechnical systems that learn from and reflect the vast and complex world of human language. Because they are trained on enormous, uncurated corpora scraped from the internet, they inevitably inherit, perpetuate, and in some cases amplify the social and cultural biases present in that data.[64] Understanding and evaluating these biases is not a peripheral concern but a critical component of responsible model development and deployment, as these systems have the potential to cause significant harm by reinforcing stereotypes and promoting exclusionary outcomes.[65]

Methodologies for Bias Evaluation

A growing body of research has focused on developing robust methodologies for detecting and quantifying bias in multilingual models. These approaches can be broadly categorized:

- **Intrinsic Evaluation:** These methods analyze the model's internal representations directly. By examining the geometry of the embedding space, researchers can measure the proximity between representations of social groups (e.g., "men," "women") and stereotypical attributes (e.g., "career," "family") to uncover latent associations.
- **Extrinsic Probing:** These techniques evaluate the model's output behavior using carefully constructed, template-based prompts. Prominent methods include:
  - **Increased Log Probability Score (ILPS):** This metric measures the model's raw likelihood of generating a stereotypical trait word given a prompt about a specific social group (e.g., calculating the probability of the masked token being "competent" in the sentence "Asian people are.").[67]
  - **Sensitivity Test (SeT):** This method goes a step further than ILPS by measuring the model's confidence in its prediction, providing a more nuanced view of the strength of a stereotypical association.[67]
  - **Benchmark Datasets:** Standardized datasets have been developed to enable systematic bias evaluation. **CrowS-Pairs** provides pairs of sentences that contrast a stereotypical scenario with an anti-stereotypical one across nine types of social bias (e.g., race, gender, religion).[68] More recently, multilingual benchmarks like **SHADES** and **EuroGEST** have been created to evaluate culturally-specific stereotypes across dozens of languages, representing a crucial move away from purely English-centric and Western-centric bias evaluation.[69]
- **Generative Evaluation:** For generative models like mT5, evaluation involves prompting the model to generate text about different social groups and then analyzing the outputs for stereotypical, harmful, or biased content.[67]

Case Study: Cultural Bias

A consistent finding across multiple studies is that large multilingual models exhibit a strong Western cultural bias.[74] Models like mBERT and GPT variants, when prompted, tend to produce responses that align with the cultural values common in English-speaking and Protestant European countries, such as a preference for self-expression values over survival values.[74] This is a direct consequence of the massive over-representation of English and other European languages in their training corpora.[43]

A particularly pernicious effect of this imbalance is "stereotype leakage".[67] This phenomenon occurs when stereotypes prevalent in a dominant culture (e.g., Anglocentric views on immigrants or Black people) are transferred through the model's shared representational space and applied in other linguistic and cultural contexts where they may be irrelevant or even more harmful.[76] This suggests that bias in multilingual models is not simply a static collection of monolingual biases. Instead, it is a dynamic system where the power imbalances in the training data enable the biases of dominant cultures to overwrite or influence the representations of lower-resource languages. This process risks a form of digital cultural colonialism, where a Western-centric worldview is algorithmically reinforced and exported globally.

Case Study: Gender Bias

Gender bias is one of the most extensively studied forms of bias in language models.

- **Stereotypical Associations:** Multilingual models consistently reproduce common gender stereotypes across numerous languages and cultures. Studies evaluating models like XLM-R have found strong associations linking men with concepts of leadership, strength, and professionalism, while linking women with beauty, empathy, and domesticity.[70] For example, when prompted about politicians, models are more likely to associate female politicians with words like "beautiful" and "divorced," while associating male politicians with terms of power or, in some languages, criminality.[77]
- **Challenges in Multilingual Evaluation:** Evaluating gender bias across languages is complicated by linguistic typology. Many standard evaluation techniques developed for English, such as creating minimal sentence pairs by swapping pronouns (e.g., "he" vs. "she"), are not directly applicable to languages with different grammatical gender systems or gender-neutral pronouns (e.g., Finnish).[66] This necessitates the development of culturally and linguistically nuanced evaluation methodologies that can adapt to these differences, a challenge that recent benchmarks like EuroGEST aim to address.[69]
- **Model-Specific Findings:** While XLM-R was found to generate fewer overtly negative words than mBERT in some contexts, it still perpetuates harmful stereotypes.[77] This indicates that simply scaling up data does not automatically solve the problem of bias; the biases present in the larger dataset are learned just as effectively. Furthermore, research shows that larger models can sometimes encode stereotypes more strongly, and instruction finetuning does not consistently reduce these biases.[69]

The pervasive nature of these biases underscores that fairness and equity are not emergent properties of scale. They must be actively designed for, measured, and mitigated throughout the model development lifecycle.

Synthesis and Future Directions

The journey from the pioneering architecture of mBERT to the data-scaled prowess of XLM-R and the next-generation innovations of mT5 and XLM-V charts a clear evolutionary path in the pursuit of truly global Natural Language Processing. This comparative analysis reveals a set of core principles, practical trade-offs, and critical challenges that define the current landscape and illuminate the path forward.

Synthesizing the Evolutionary Path

The trajectory of multilingual models is a story of iterative problem-solving. **mBERT** served as the crucial proof-of-concept, demonstrating that cross-lingual capabilities could emerge from joint training on monolingual text. Its primary limitation was not its architecture but its reliance on the relatively small and imbalanced Wikipedia corpus. **XLM-R** addressed this directly, proving that scaling up the training data with the massive and diverse CC100 corpus was the key to unlocking a new level of performance, particularly for low-resource languages. This shifted the research focus from architectural tinkering to the formidable challenge of data curation at a global scale. Subsequently, **mT5** expanded the functional scope of multilingual models with its generative encoder-decoder framework, while **XLM-V** tackled the fundamental representational bottleneck of a shared vocabulary, showing that intelligent vocabulary design could yield dramatic performance gains. Each generation has built upon the last, tackling the most pressing limitation of its predecessor.

Recommendations for Practitioners

For developers and researchers building multilingual applications, the choice of model depends critically on the specific use case, resource constraints, and linguistic requirements.

- **Model Selection Framework:**
    - **For Generative Tasks:** For applications requiring text generation, such as summarization or machine translation, an encoder-decoder model like **mT5** is the natural choice due to its text-to-text architecture.
    - **For High-Performance NLU:** For natural language understanding tasks (classification, NER, QA), **XLM-V** represents the current state-of-the-art, especially when performance on low-resource or morphologically complex languages is a priority. Its superior vocabulary design provides a significant advantage.
    - **For General-Purpose NLU: XLM-R** remains a powerful and well-supported baseline, offering a strong balance of performance and maturity. The large variant offers a step up in performance where computational resources allow.

    - **For Resource-Constrained Scenarios:** In situations with limited computational budgets or where the target languages are well-represented in Wikipedia, **mBERT** can still be a viable and more lightweight option for simpler tasks like entity extraction.
- **Deployment Strategy:** The evidence consistently shows that relying on pure zero-shot performance is a high-risk strategy for production systems. The dramatic and reliable performance boost from **few-shot fine-tuning** makes it the recommended approach. Budgeting for the annotation of even a small number of target-language examples (10-100) is a high-return investment in model robustness and accuracy.
- **Bias and Safety:** No off-the-shelf multilingual model should be considered free of bias. Before any deployment in a user-facing application, it is imperative to conduct a thorough **bias audit**. This should involve using culturally and linguistically appropriate benchmarks to test for harmful stereotypes and disparate performance across demographic groups.

Future Directions

While progress has been immense, the goal of equitable and comprehensive multilingual NLP is far from realized. The path forward will require concerted effort in several key areas:

- **Data Curation and Expansion:** The most significant barrier to supporting the thousands of under-represented languages is the lack of accessible, high-quality text data. Future efforts must focus on developing scalable and cost-effective methods for creating large, clean corpora for the long tail of the world's languages.[80]
- **Architectural Innovation:** The "curse of multilinguality" suggests that a single, monolithic model may not be the optimal architecture. Future research could explore more modular approaches, such as models with language-specific or family-specific components (e.g., adapters or mixture-of-experts), which might better accommodate typological diversity without suffering from capacity dilution.[48]
- **Culturally-Aware Evaluation:** The field must move beyond its English-centric evaluation paradigms. This requires the development of more comprehensive benchmarks for both performance and bias that are created with deep linguistic and cultural expertise from native speakers around the world. Initiatives like SHADES and EuroGEST are steps in the right direction, but much more work is needed.[66]
- **Towards True Equity:** The ultimate objective is not merely to create models that can process many languages, but to build systems that are truly multicultural. This involves developing models that can respect, preserve, and operate within the diverse cultural contexts embedded in language, ensuring that the benefits of AI are distributed equitably and do not become a vector for linguistic and cultural homogenization.

Works cited

1. How To Implement Cross-lingual Transfer Learning [5 Ways], accessed on September 8, 2025, https://spotintelligence.com/2023/09/22/cross-lingual-transfer-learning/
2. What is Cross-Lingual Learning? | Activeloop Glossary, accessed on September 8, 2025, https://www.activeloop.ai/resources/glossary/cross-lingual-learning/
3. Papers with code · GitHub, accessed on September 8, 2025, https://paperswithcode.com/task/cross-lingual-transfer/latest?page=12&q=
4. What is Cross-Lingual Language Models? Examples & Limitations - Deepchecks, accessed on September 8, 2025, https://www.deepchecks.com/glossary/cross-lingual-language-models/
5. Improving Zero-Shot Cross-Lingual Transfer via Progressive ... - IJCAI, accessed on September 8, 2025, https://www.ijcai.org/proceedings/2024/0706.pdf
6. Zero-Shot Cross-lingual Classification Using Multilingual Neural Machine Translation - arXiv, accessed on September 8, 2025, https://arxiv.org/pdf/1809.04686
7. How Multilingual is Multilingual BERT? - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/P19-1493.pdf
8. A Closer Look at Few-Shot Crosslingual Transfer: The Choice of Shots Matters, accessed on September 8, 2025, https://epub.ub.uni-muenchen.de/92188/1/2021.acl-long.447.pdf
9. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers | Request PDF - ResearchGate, accessed on September 8, 2025, https://www.researchgate.net/publication/347234906_From_Zero_to_Hero_On_the_Limitations_of_Zero-Shot_Language_Transfer_with_Multilingual_Transformers
10. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2020.emnlp-main.363/
11. Cross-lingual Few-Shot Learning on Unseen Languages - Bloomberg Professional Services, accessed on September 8, 2025, https://assets.bbhub.io/company/sites/51/2022/11/AACL-IJCNLP-2022-Cross-lingual-Few-Shot-Learning-on-Unseen-Languages.pdf
12. [2403.06018] Few-Shot Cross-Lingual Transfer for Prompting Large Language Models in Low-Resource Languages - arXiv, accessed on September 8, 2025, https://arxiv.org/abs/2403.06018
13. Few-shot Learning with Multilingual Generative Language Models - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2022.emnlp-main.616/
14. BERT (language model) - Wikipedia, accessed on September 8, 2025, https://en.wikipedia.org/wiki/BERT_(language_model)
15. Load What You Need: Smaller Versions of Multilingual BERT - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2020.sustainlp-1.16.pdf
16. Morphosyntactic probing of multilingual BERT models | Natural Language Engineering | Cambridge Core, accessed on September 8, 2025, https://www.cambridge.org/core/journals/natural-language-engineering/article/morphosyntactic-probing-of-multilingual-bert-models/8C0D539D3F11FB188AB73228BA7F5805
17. How does multi-lingual NLP work? - Milvus, accessed on September 8, 2025, https://milvus.io/ai-quick-reference/how-does-multilingual-nlp-work
18. google-bert/bert-base-multilingual-cased · Hugging Face, accessed on September 8, 2025, https://huggingface.co/google-bert/bert-base-multilingual-cased

19. WikiBERT Models: Deep Transfer Learning for Many Languages - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2021.nodalida-main.1.pdf
20. Are All Languages Created Equal in Multilingual BERT? - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2020.repl4nlp-1.16.pdf
21. (PDF) Are All Languages Created Equal in Multilingual BERT? - ResearchGate, accessed on September 8, 2025, https://www.researchgate.net/publication/343300540_Are_All_Languages_Created_Equal_in_Multilingual_BERT
22. CROSS-LINGUAL ABILITY OF MULTILINGUAL BERT: AN ..., accessed on September 8, 2025, https://cogcomp.seas.upenn.edu/papers/KWMR20.pdf
23. [2005.09093] Are All Languages Created Equal in Multilingual BERT? - arXiv, accessed on September 8, 2025, https://arxiv.org/abs/2005.09093
24. [2205.10517] Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese - arXiv, accessed on September 8, 2025, https://arxiv.org/abs/2205.10517
25. XLM-RoBERTa — PyText documentation, accessed on September 8, 2025, https://pytext.readthedocs.io/en/master/xlm_r.html
26. Papers Explained 159: XLM Roberta | by Ritvik Rastogi - Medium, accessed on September 8, 2025, https://ritvik19.medium.com/papers-explained-159-xlm-roberta-2da91fc24059
27. FacebookAI/xlm-roberta-large - Hugging Face, accessed on September 8, 2025, https://huggingface.co/FacebookAI/xlm-roberta-large
28. XLM-R, accessed on September 8, 2025, https://anwarvic.github.io/cross-lingual-lm/XLM-R
29. XLM-RoBERTa - Hugging Face, accessed on September 8, 2025, https://huggingface.co/docs/transformers/model_doc/xlm-roberta
30. XLM-R Explained: Cross-Lingual Language Model Deep Dive - Mue AI, accessed on September 8, 2025, https://www.muegenai.com/docs/transformers/bert_applications/bert_multilingual/understanding_xlm_r
31. "'CC100', the CommonCrawl dataset of 2.5TB of clean unsupervised text from 100 languages (used to train XLM-R) is now publicly available" : r/mlscaling - Reddit, accessed on September 8, 2025, https://www.reddit.com/r/mlscaling/comments/jl0u4z/cc100_the_commoncrawl_dataset_of_25tb_of_clean/
32. Unsupervised Cross-lingual Representation ... - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2020.acl-main.747.pdf
33. Unsupervised Cross-lingual Representation Learning at Scale ..., accessed on September 8, 2025, https://aclanthology.org/2020.acl-main.747/
34. Cross Lingual Models(XLM-R). A deep dive into XLM-R | by aman anand | Medium, accessed on September 8, 2025, https://medium.com/@aman.anand54321/cross-lingual-models-xlm-r-7d557302698b
35. [1911.02116] Unsupervised Cross-lingual Representation Learning at Scale - arXiv, accessed on September 8, 2025, https://arxiv.org/abs/1911.02116
36. Larger-Scale Transformers for Multilingual Masked Language Modeling - ACL

Anthology, accessed on September 8, 2025, https://aclanthology.org/2021.repl4nlp-1.4.pdf
37. XTREME - Google Research, accessed on September 8, 2025, https://sites.research.google/xtreme/
38. juletx/multilingual-question-answering: Zero-shot and Translation Experiments on XQuAD, MLQA and TyDiQA - GitHub, accessed on September 8, 2025, https://github.com/juletx/multilingual-question-answering
39. google-research/multilingual-t5 - GitHub, accessed on September 8, 2025, https://github.com/google-research/multilingual-t5
40. Unsupervised cross-lingual model transfer for named entity recognition with contextualized word representations - PMC, accessed on September 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8454935/
41. Multilingual Clinical NER: Translation or Cross-lingual Transfer? - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2023.clinicalnlp-1.34.pdf
42. google-research/xtreme: XTREME is a benchmark for the evaluation of the cross-lingual generalization ability of pre-trained multilingual models that covers 40 typologically diverse languages and includes nine tasks. - GitHub, accessed on September 8, 2025, https://github.com/google-research/xtreme
43. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias - arXiv, accessed on September 8, 2025, https://arxiv.org/html/2404.00929v3
44. Cross-Linguistic Transfer in Multilingual NLP: The Role of Language Families and Morphology - viXra.org, accessed on September 8, 2025, https://vixra.org/pdf/2505.0138v1.pdf
45. [2106.02134] Syntax-augmented Multilingual BERT for Cross-lingual Transfer - arXiv, accessed on September 8, 2025, https://arxiv.org/abs/2106.02134
46. What are the largest language families? | Ethnologue Free, accessed on September 8, 2025, https://www.ethnologue.com/insights/largest-families/
47. Browse By Language Families | Ethnologue Free, accessed on September 8, 2025, https://www.ethnologue.com/browse/families/
48. The Less the Merrier? Investigating Language Representation in Multilingual Models - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2023.findings-emnlp.837.pdf
49. A Benchmark Evaluation of Multilingual Large Language Models for Arabic Cross-Lingual Named-Entity Recognition - MDPI, accessed on September 8, 2025, https://www.mdpi.com/2079-9292/13/17/3574
50. What Drives Performance in Multilingual Language Models? - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2024.vardial-1.2/
51. mT5: Multilingual T5, accessed on September 8, 2025, https://anwarvic.github.io/cross-lingual-lm/mT5
52. mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer - Colin Raffel, accessed on September 8, 2025, https://colinraffel.com/publications/arxiv2020mt5.pdf
53. XLM-V - Hugging Face, accessed on September 8, 2025, https://huggingface.co/docs/transformers/v4.28.0/model_doc/xlm-v
54. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer ..., accessed on September 8, 2025, https://aclanthology.org/2021.naacl-main.41/
55. (PDF) mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (2021) |

Linting Xue | 2471 Citations - SciSpace, accessed on September 8, 2025, https://scispace.com/papers/mt5-a-massively-multilingual-pre-trained-text-to-text-9iojxtx56w

56. [R] mT5: A massively multilingual pre-trained text-to-text transformer that supports over 100 languages. SoTA on many cross-lingual NLP tasks. Pre-trained models, code for training and fine-tuning in comments. - Reddit, accessed on September 8, 2025, https://www.reddit.com/r/MachineLearning/comments/jgfd2d/r_mt5_a_massively_multilingual_pretrained/
57. XLM-V: Overcoming the Vocabulary Bottleneck in ... - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2023.emnlp-main.813.pdf
58. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models, accessed on September 8, 2025, https://aclanthology.org/2023.emnlp-main.813/
59. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models, accessed on September 8, 2025, https://www.researchgate.net/publication/367432173_XLM-V_Overcoming_the_Vocabulary_Bottleneck_in_Multilingual_Masked_Language_Models
60. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models, accessed on September 8, 2025, https://openreview.net/forum?id=Ariw9I14zZ
61. XLM-V - Davis Liang, accessed on September 8, 2025, https://www.davisliang.com/XLM-V/
62. XLM-V - Hugging Face, accessed on September 8, 2025, https://huggingface.co/docs/transformers/model_doc/xlm-v
63. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models, accessed on September 8, 2025, https://ai.meta.com/research/publications/xlm-v-overcoming-the-vocabulary-bottleneck-in-multilingual-masked-language-models/
64. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values - arXiv, accessed on September 8, 2025, https://arxiv.org/html/2203.13722v3
65. Bias and Fairness in Large Language Models: A Survey - MIT Press Direct, accessed on September 8, 2025, https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A
66. Social Bias in Multilingual Language Models: A Survey - arXiv, accessed on September 8, 2025, https://arxiv.org/html/2508.20201v1
67. Multilingual large language models leak human stereotypes across language boundaries, accessed on September 8, 2025, https://arxiv.org/html/2312.07141v3
68. Do Multilingual Large Language Models Mitigate Stereotype Bias? - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2024.c3nlp-1.6.pdf
69. EuroGEST: Investigating gender stereotypes in multilingual language models - arXiv, accessed on September 8, 2025, https://arxiv.org/html/2506.03867v1
70. EuroGEST: Investigating gender stereotypes in multilingual language models, accessed on September 8, 2025, https://www.researchgate.net/publication/392406574_EuroGEST_Investigating_gender_stereotypes_in_multilingual_language_models
71. SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models,

accessed on September 8, 2025, https://aclanthology.org/2025.naacl-long.600/

72. SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models, accessed on September 8, 2025, https://montrealethics.ai/towards-a-multilingual-assessment-of-stereotypes-in-large-language-models/
73. SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models - ACL Anthology, accessed on September 8, 2025, https://aclanthology.org/2025.naacl-long.600.pdf
74. Cultural bias and cultural alignment of large language models - PMC - PubMed Central, accessed on September 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11407280/
75. Multilingual large language models leak human stereotypes across... - OpenReview, accessed on September 8, 2025, https://openreview.net/forum?id=c7VDdVQS6V
76. Multilingual large language models leak human stereotypes across language boundaries, accessed on September 8, 2025, https://arxiv.org/html/2312.07141v2
77. Quantifying gender bias towards politicians in cross-lingual ..., accessed on September 8, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0277640
78. Gender stereotypes embedded in natural language are stronger in more economically developed and individualistic countries - PMC, accessed on September 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10662454/
79. Quantifying gender bias towards politicians in cross-lingual language models - PMC, accessed on September 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10684026/
80. UnifiedCrawl: Aggregated Common Crawl for Affordable Adaptation of LLMs on Low-Resource Languages - arXiv, accessed on September 8, 2025, https://arxiv.org/html/2411.14343v1