

On Multilingual Text Summarization: A Review of Model Architectures, Techniques, and Research Gaps

Charles Munyao Muia¹, Aaron Mogeni Oirere¹, Stephen Kahara Wanjau¹

¹Department of Computer Science, Murang'a University of Technology, 75-10200, Murang'a, Kenya

Article Info

Article history:

Received August 16, 2026

Revised October 09, 2026

Accepted October 15, 2026

Keywords:

Text Summarization

Natural Language Processing

Systematic Literature Review

Low-Resource Languages

Large Language Models

ABSTRACT

The rapid expansion of multilingual digital content has intensified demand for automated text summarization systems that can operate across diverse languages. Yet the architectural and methodological foundations of this domain remain dispersed across individual studies. This paper presents a systematic literature review of multilingual text summarization research in Natural Language Processing, with particular attention to model architectures, multilingual enabling techniques, and existing research gaps. Following a PRISMA-guided review protocol, 47 studies published between 2017 and 2026 were identified, screened, and synthesized from major scholarly databases using predefined eligibility and quality assessment criteria. The analysis shows that the literature is dominated by four dominant architectural paradigms: multilingual sequence-to-sequence models, large language models, adapter-based models, and hybrid architectures, with a recent shift toward large language model-based approaches. Across these studies, multilingual capability is primarily achieved through pre-training and transfer learning, with support from parameter-efficient adaptation and data augmentation. The review further identified four persistent limitation categories: low-resource language underrepresentation; dataset and benchmark constraints; evaluation limitations dominated by ROUGE overreliance; and generalization and scalability challenges, manifested as the curse of multilinguality. These findings indicate that current progress in the field is largely driven by scaling existing models rather than resolving structural multilingual limitations, particularly those related to data availability and model architecture. This paper provides a consolidated view of architectural and methodological trends and offers a structured, informed basis for future research on equitable multilingual summarization, particularly for low-resource languages.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author: Charles Munyao Muia (e-mail: munyao@mut.ac.ke)

1. INTRODUCTION

The rapid growth of digital content across multiple languages has increased the need for Natural Language Processing (NLP) systems that support equitable access to information. Automatic text summarization addresses information overload by generating concise representations of longer documents while preserving salient information [1]. Early summarization research was predominantly concentrated on English-language resources, reflecting the availability of large datasets, linguistic resources, and established evaluation benchmarks. The expansion of multilingual digital content has consequently increased the need for summarization systems capable of operating across diverse linguistic settings [2]. Multilingual text summarization has therefore attracted growing attention due to its potential applications in multilingual news dissemination, information retrieval, knowledge management, and access to information for speakers of less-resourced languages.

The development of neural sequence-to-sequence and Transformer-based architectures has substantially advanced automatic text summarization. Attention-based encoder-decoder models improved the representation of relevant contextual information for summary generation [3], while Transformer

architectures provided more effective mechanisms for modelling contextual relationships and long-range dependencies [4]. The emergence of multilingual pretrained language models subsequently enabled a common model to represent and process multiple languages [5]. These developments have encouraged the use of shared multilingual representations, cross-lingual transfer, parameter sharing, and language-specific adaptation strategies to exploit common linguistic information while accommodating differences between languages.

Despite these advances, multilingual text summarization remains challenging, given that languages differ in linguistic structure, resource availability, corpus representation, and the extent to which knowledge learned in one language can be transferred to another. These differences can result in substantial variation in summarization quality across languages, even when a common architecture is employed [6]. The literature has consequently expanded across diverse architectural configurations, multilingual adaptation strategies, datasets, language settings, and evaluation practices. This growing diversity has produced a fragmented evidence base, making it difficult to establish which architectural patterns recur across multilingual summarization research, how multilingual capability is enabled across different approaches, and which limitations continue to constrain progress, particularly for low-resource languages.

The present study addresses this fragmentation through a Systematic Literature Review of research on multilingual text summarization. Existing reviews have generally examined automatic summarization broadly, multilingual NLP across multiple tasks, or multilingual/cross-lingual summarization from specific perspectives. However, there remains limited systematic synthesis that jointly examines the architectural foundations of multilingual text summarization, the mechanisms used to achieve multilingual capability, and the persistent limitations and research gaps associated with these choices. Bringing these dimensions together within a common evidence base, the study contributes a structured characterization of the architectural development of multilingual summarization and identifies persistent challenges involving low-resource language coverage, datasets and benchmarks, evaluation, multilingual generalization, and scalability. The resulting synthesis also contributes an evidence-based foundation for understanding the current state of multilingual summarization research and for identifying unresolved architectural questions that could warrant subsequent empirical investigation.

The remainder of this paper is organized as follows. Section 2 reviews existing related syntheses and establishes the distinction between the present study and prior reviews. Section 3 presents the methodology used to identify, screen, select, and analyze relevant studies. Section 4 reports the review's findings, focusing on model architectures, multilingual techniques, and identified research gaps. Section 5 discusses the implications of these findings for multilingual text summarization research. Section 6 concludes the paper and outlines directions for further research.

2. RELATED WORK

Existing reviews of text summarization have examined the evolution of summarization methods, modeling paradigms, datasets, and evaluation approaches. Reviews on the broader development of automatic summarization have been useful in establishing the transition from traditional, extractive approaches to neural and Transformer-based approaches [7]. However, their broad task-oriented scope provides limited insight into the architectural issues that arise specifically when summarization is performed across multiple languages. In multilingual summarization, architectural design must additionally address how parameters and representations are shared across languages, how language-specific characteristics are accommodated, and how unequal language resources affect model performance. Consequently, a review of summarization architectures in general does not fully establish which architectural patterns are particularly relevant to multilingual summarization.

Reviews of multilingual NLP have approached the problem from a broader language-processing perspective. Research in this area has examined multilingual representation learning, cross-lingual transfer, and multilingual pretrained models [8][9]. These reviews have provided important insights into how models acquire and transfer shared representations across languages. However, their coverage generally extends across several NLP tasks rather than treating summarization as a distinct analytical problem. As a result, architectural findings on multilingual representation and transfer cannot necessarily be interpreted as evidence for the requirements of multilingual summary generation, in which the model must simultaneously perform language understanding, content selection, compression, and generation.

Research on cross-lingual summarization has also provided more direct coverage of the task and highlighted challenges involving multilingual datasets, language-resource imbalance, cross-lingual transfer, and model development [10][11]. The closer alignment with multilingual summarization makes them more relevant to the present study, but their analytical coverage remains fragmented across specific task formulations, datasets, modeling paradigms, or broader challenges. In particular, the architectural dimensions underlying multilingual capability have not been consistently consolidated across the literature. Issues such as parameter sharing, language-specific components, multilingual pretraining, and adaptation mechanisms are

frequently investigated in individual studies. Yet, the existing review literature provides limited systematic evidence on how these architectural choices collectively characterize the development of multilingual summarization.

Recent developments in multilingual language modelling and summarization have further expanded this design space through large pretrained models, parameter-efficient adaptation, and other mechanisms for transferring knowledge across languages [12][13]. These developments reinforce the need for an architectural synthesis, as improvements reported by individual studies may reflect differences in model capacity, adaptation strategies, language configurations, dataset compositions, or evaluation protocols rather than a common architectural advantage. Broader reviews of multilingual modelling have also tended to treat summarization alongside other NLP tasks [14], while existing summarization-focused reviews have not consistently isolated the architectural consequences of multilinguality [15]. Unlike prior reviews, the present study treats multilingual text summarization as a distinct architectural problem and systematically examines the relationships among model architectures, multilingual techniques, language-resource conditions, and reported research gaps within a single review framework. This distinction allows the review to move beyond summarizing individual models or reproducing broader multilingual NLP surveys and instead establish recurring architectural patterns and unresolved challenges specific to multilingual text summarization.

3. METHODOLOGY

This study adopted a Systematic Literature Review (SLR) approach guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. PRISMA was selected because it provides a transparent, structured, and reproducible process for identifying, screening, selecting, and synthesizing relevant literature [16]. The review process was organized into eight stages: formulation of the research questions, development of the search strategy, definition of eligibility criteria, study selection, quality assessment, data extraction, data synthesis, and analysis.

3.1. Research Questions

The review was guided by research questions developed to support the study objective of examining model architectures, multilingual techniques, and existing research gaps in multilingual text summarization. The questions were defined before the review process and informed the search strategy, eligibility criteria, data extraction, and synthesis procedures.

1. RQ1: What model architectures have been used in multilingual text summarization research?
2. RQ2: What techniques are used to enable multilingual capability in text summarization models across different language settings?
3. RQ3: What limitations and research gaps have been identified in the current multilingual text summarization approaches?

3.2. Search Strategy

The search followed a systematic, reproducible protocol aligned with the PRISMA guidelines and covered six databases: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, and Google Scholar. Keywords were combined to capture variations in terminology around multilingual text summarization. Core search phrases included "multilingual text summarization," "cross-lingual summarization," and "multilingual abstractive summarization," and these were combined with architectural and methodological terms such as "Transformer," "encoder-decoder," "neural architecture," "parameter sharing," and "multilingual pre-training." An identical boolean search string was applied to all databases, adapted only to each database's field syntax: ("multilingual text summarization" OR "cross-lingual summarization") AND ("Transformer" OR "encoder-decoder") AND ("architecture" OR "model"). All searches were conducted between 3rd and 7th June 2026. The searches retrieved 32 records from IEEE Xplore, 28 from ACM Digital Library, 61 from Scopus, 45 from Web of Science, 34 from SpringerLink, and 23 from Google Scholar, giving a total of 223 records.

The search was restricted to publications appearing between January 2017 and May 2026, a period corresponding to the emergence and maturation of transformer-based NLP models. This publication period was selected to capture the development of Transformer-based approaches and recent advances in large language model-based multilingual summarization. This period provides coverage of major developments in the field, while the May 2026 cutoff may exclude studies published after the search period, potentially affecting coverage of the most recent developments. The search process was iterative, with successive keyword refinements across databases to balance breadth of retrieval and relevance, ensuring that key studies were not overlooked while maintaining focus on the target domain. Peer-reviewed journal articles were prioritized to ensure the inclusion of validated, methodologically sound research findings. In addition to peer-

reviewed journal articles, select preprints from established, reputable repositories were considered to capture recent and emerging developments in multilingual text summarization.

3.3. Eligibility Criteria

Eligibility criteria were defined a priori to ensure that the review corpus was both internally coherent and representative of the primary research directions addressed by the three research questions. The eligibility criteria were operationalized as separate inclusion and exclusion conditions applied at each stage of the selection process.

3.3.1. Inclusion Criteria

- Studies addressing multilingual text summarization or explicitly targeting low-resource or zero-shot cross-lingual summarization settings.
- Studies proposing or evaluating model architectures, multilingual techniques, or approaches relevant to multilingual text summarization
- Studies published in English, between 2017 and 2026, in peer-reviewed journals or reputable preprint repositories.

3.3.2. Exclusion Criteria

- Studies focusing exclusively on monolingual summarization with no multilingual or cross-lingual component.
- Studies addressing general multilingual NLP tasks without explicit application to text summarization.
- Studies lacking sufficient methodological detail, empirical evaluation, or accessible full-text availability for analysis, as well as non-English publications.

The restriction to English-language publications was adopted because the review team could reliably assess methodological quality only in English, and multilingual text summarization research reported in international venues is predominantly published in English. This restriction is acknowledged as a potential limitation, particularly because relevant research on low-resource languages may be published in other languages.

3.4. Study Selection

The study selection process followed the PRISMA framework to ensure transparency, reproducibility, and systematic filtering of relevant literature. The selection was conducted in sequential stages, beginning with database search retrieval and ending with the final inclusion of studies for synthesis. Initially, 223 records were identified through the search process conducted on the six scholarly databases. After merging results and scanning for duplicates, 41 duplicate records were excluded, yielding 182 unique studies for screening. The remaining studies underwent title and abstract screening based on the predefined eligibility criteria. At this stage, 103 studies were excluded for irrelevance to multilingual text summarization or for not aligning with the research questions. Subsequently, 79 studies were retained for full-text assessment. During full-text evaluation, 32 studies were excluded guided by the quality assessment criteria. Following this screening process, 47 studies met all inclusion criteria and were retained for final analysis and synthesis in this review.

The 47 included studies were distributed across the 2017–2026 review period, with relatively fewer studies published in the earlier years and increasing publication activity in the later years. Two studies were published in 2017, 2 in 2018, 3 in 2019, 4 in 2020, 5 in 2021, 7 in 2022, 9 in 2023, 7 in 2024, 5 in 2025, and 3 between January and May 2026. This distribution indicates growing research activity in multilingual text summarization during the latter part of the review period, coinciding with the increasing adoption of Transformer-based and large language model approaches. Study screening and quality assessment were performed independently by two reviewers. Any disagreements between the two reviewers were resolved through discussion, and when a consensus could not be reached, a third senior reviewer was consulted to make the final decision. The overall study selection process is summarized using a PRISMA flow diagram in Figure 1.

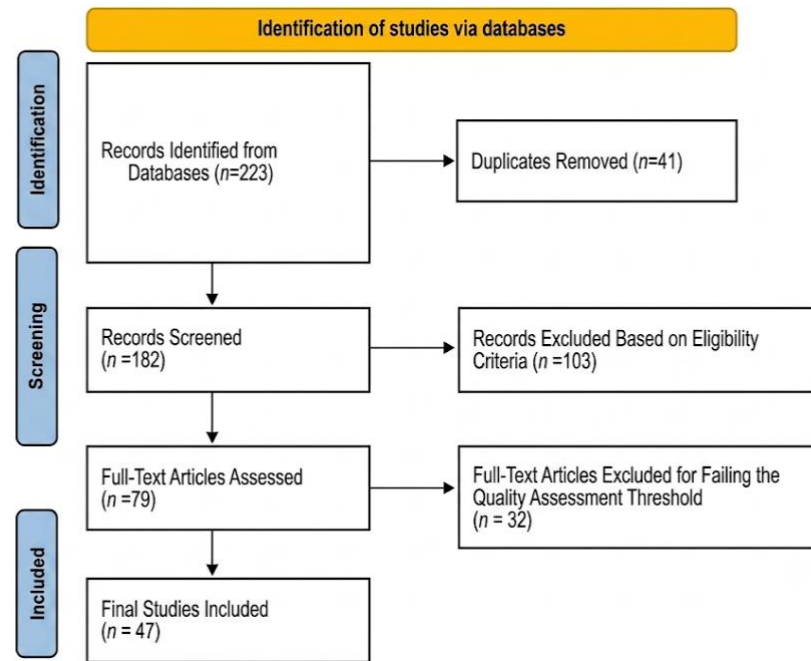


Figure 1. PRISMA Flowchart

3.5. Quality Assessment

A structured quality assessment was conducted to evaluate the methodological consistency and reliability of the studies included in the review. Each study was assessed during the full-text screening stage using five predefined quality assessment criteria aligned with the objectives of the review. The criteria evaluated were: (1) clarity and completeness of the study description, (2) clarity and appropriateness of the research methodology, (3) appropriateness of the experimental or analytical design, (4) transparency of the reported data and evaluation procedures, and (5) adequacy of the evidence supporting the reported findings. Each criterion was scored using a three-point scale: 2 = fully satisfied, 1 = partially satisfied, and 0 = not satisfied. The scores were summed to produce an overall quality score ranging from 0 to 10 for each study. A minimum score of 6 out of 10 was required for a study to satisfy the minimum quality threshold for inclusion in the final synthesis. Using these criteria and the scoring procedure, the 32 studies excluded at the full-text stage were removed for failing to meet the substantive eligibility requirements and/or the predefined minimum quality threshold. The final set of 47 studies, therefore, represents the studies that satisfied both the eligibility requirements and the minimum methodological and evidential quality requirements established for this review.

3.6. Data Extraction

A standardized data extraction form was developed and piloted on a subset of the included studies before full-scale extraction to ensure consistency and relevance to the three research questions. The form captured both study-level metadata and analytical variables required for synthesis. For each of the 47 included studies, the extracted information comprised bibliographic details; model architecture or architectural family; languages addressed and their typological characteristics; multilingual techniques employed; datasets used, with their language and domain coverage; evaluation metrics; key findings relevant to the research questions; and reported limitations. Where relevant information was not explicitly stated but could be reliably inferred from methodological descriptions, it was recorded as inferred, but with appropriate annotation. Missing or ambiguous information was left unrecorded and documented accordingly. All extracted data were systematically organized in a structured spreadsheet maintained throughout the review process with version control to ensure traceability and reproducibility of extraction decisions.

3.7. Data Synthesis

A qualitative, narrative-driven synthesis approach was adopted to analyze the extracted data from the 47 included studies. Thematic synthesis proceeded in three iterative phases. In the first phase, extracted data were organized into an analytical matrix aligned with the three research questions. In the second phase,

cross-cutting themes were identified by examining recurring patterns, divergences, and methodological approaches. In the third phase, these themes were refined into higher-order synthetic narratives, distinguishing between widely supported findings and those that were context-dependent or reported in a limited subset of studies.

To distinguish reported evidence from interpretation, the synthesis first describes evidence extracted directly from the included studies. The authors' interpretation is presented separately when identifying broader patterns, relationships, or implications across the studies. Interpretive statements were based on the aggregated evidence and were not presented as findings directly reported by individual studies. In the taxonomy tables, the counting rules were as follows: Architecture classes were treated as mutually exclusive, so their counts sum to the 47 included studies. Technique and limitation categories were treated as non-exclusive; therefore, a study was counted once for each category it reported. Sub-patterns within a category were likewise counted as non-exclusive when a study reported multiple techniques. Unless otherwise stated, all percentages were calculated using the 47-study corpus as the denominator.

4. RESULTS

This section presents the results of the systematic literature review of 47 studies. The findings were organized into a taxonomy linking the main dimensions of multilingual text summarization. The taxonomy connects model architectures with multilingual techniques used to support multilingual capability, while datasets and benchmarks provide the basis for training and evaluation. These dimensions are further linked to the limitations and research gaps identified in the reviewed studies, particularly those concerning low-resource language coverage, dataset limitations, evaluation, and generalization. This organization provides a consistent framework for synthesizing the findings across the three research questions.

4.1. Model Architectures for Multilingual Text Summarization

In answering the first research question, four distinct architecture classes employed in multilingual text summarization were identified. They included: multilingual sequence-to-sequence models, Adapter-based and parameter-efficient models, large language model (LLM)-based approaches, and Hybrid architectures. To ensure consistent classification, the four architectural classes were defined based on the primary architectural mechanism reported in each study. Seq2seq models used encoder–decoder architectures; hybrid models combined a main architecture with an additional structural, retrieval, graph, or knowledge-based component; adapter-based models used parameter-efficient adaptation methods such as adapters, LoRA, or prefix-tuning; and LLM-based approaches used large-scale generative pretrained models as the primary framework. Where categories overlapped, studies were classified according to their primary architectural contribution. Table 1 presents a structured taxonomy of the identified architecture classes, study counts, and representative model types or approaches, as extracted from the included literature.

Table 1. Model architectures for multilingual text summarization

Architecture Class	Count (n = 47)	Representative Models / Approaches
Multilingual Sequence-to-Sequence Models	19 (40%)	mBART, mT5, multilingual BART variants, encoder–decoder Transformers
Hybrid Architectures	7 (15%)	Retrieval-augmented models, graph-enhanced Transformers, knowledge-infused summarizers
Adapter-Based / Parameter-Efficient Models	10 (22%)	Adapter-tuned Transformers, LoRA-based summarizers, prefix-tuning configurations
Large Language Model (LLM)-Based Approaches	11 (23%)	Instruction-tuned LLMs, generative foundation models, supervised fine-tuning

4.1.1. Multilingual Sequence-to-Sequence Models

Multilingual sequence-to-sequence (seq2seq) models constituted the most frequently reported architecture class, documented in 19 of the 47 reviewed studies. All 19 studies reported using Transformer-based encoder–decoder frameworks as the primary model backbone. Of these, 12 studies employed mBART or its multilingual extension mBART-50 as the base architecture [17]. Five studies (26%) reported using mT5 in encoder–decoder configurations. The remaining 2 studies reported custom multilingual Transformer encoder–decoder architectures trained from scratch on domain-specific multilingual corpora. With respect to input–output language configurations, 14 of the 19 studies (74%) reported monolingual summarization across

multiple languages, with each language evaluated independently. The remaining 5 studies reported cross-lingual summarization settings in which the input document language differed from the output summary language. The number of languages addressed per study within this architecture class ranged from 2 to 45. Seventeen of the 19 studies conducted evaluations exclusively on high-resource language pairs, a common practice claimed in [18]. Two studies explicitly reported degradation in cross-lingual generalization when the architecture was applied to low-resource target languages outside the pre-training distribution.

4.1.2. Hybrid Architectures

Hybrid architectures were reported in 7 of the 47 reviewed studies (15%). Studies in this class described systems that combined a neural sequence-to-sequence backbone with one or more supplementary structural or retrieval-based modules. Three distinct hybrid configurations were documented across the reviewed literature. Retrieval-augmented summarization architectures were reported in 3 of the 7 studies. These studies reported incorporating a document retrieval component, proposed in [19], that supplied candidate passages to the generation module as auxiliary context, with retrieved passage counts ranging from top-3 to top-10 per input document. Graph-enhanced Transformer configurations were reported in an additional 3 studies (43%). All three studies reported appending graph neural network (GNN) layers or dependency-parse-based graph encoders to the Transformer encoder to capture inter-sentential and cross-lingual structural relationships before decoding [20]. Two of the three graph-based studies applied the architecture to long-document summarization tasks with input lengths exceeding 1,024 tokens. In one study, 14% of the hybrid class reported a knowledge-infused summarization architecture that integrates a multilingual knowledge graph into the encoder via entity-linking and relation-embedding layers [21]. All hybrid architecture studies reported increased computational overhead relative to standard seq2seq baselines as a documented characteristic of their configurations.

4.1.3. Adapter-Based / Parameter-Efficient Models

Adapter-based and parameter-efficient model architectures were reported in 10 of the 47 reviewed studies (21%). Studies in this class described configurations in which large pretrained multilingual models were adapted for summarization tasks by updating only a restricted subset of parameters, while the base model weights were held frozen [22]. Three parameter-efficient adaptation strategies were identified: adapter module insertion, Low-Rank Adaptation (LoRA), and prefix-tuning. Adapter module insertion was the most frequently documented approach, reported in 5 of the 10 studies (50%). All five studies reported inserting bottleneck feed-forward adapter layers at each Transformer block. The proportion of trainable parameters relative to total model parameters was explicitly reported in 4 of these studies, with values ranging from 0.5% to 3.2%. LoRA-based adaptation was documented in 4 of the 10 studies. These studies described decomposing weight update matrices of Transformer attention layers into low-rank matrix products during fine-tuning. Reported LoRA rank hyperparameters across these studies ranged from $r = 4$ to $r = 16$. Prefix-tuning was reported in 1 study, which described prepending a sequence of trainable continuous prompt vectors to the input of each Transformer layer to condition generation on a target-language summarization objective. Across all 10 studies in this class, the reported target-language counts ranged from 3 to 30 languages. Eight of the 10 studies applied these configurations in low-resource language settings, where full fine-tuning was reported to be computationally infeasible.

4.1.4. Large Language Model (LLM)-Based Approaches

The reviewed studies show that recent LLM-based summarization approaches increasingly rely on large pretrained generative models combined with different adaptation and prompting strategies. Of the 11 LLM-based studies, five investigated instruction-tuned models in zero-shot or few-shot settings, four used supervised fine-tuning of models such as mBART-large and mT5-XXL, and two examined both zero-shot prompting and supervised approaches. This indicates a shift from conventional multilingual encoder-decoder architectures toward instruction-based and generative approaches that can perform summarization with limited task-specific training. The use of zero-shot and few-shot prompting is particularly relevant to multilingual summarization because it provides a mechanism to extend summarization capabilities to languages with limited annotated training data. However, the reviewed studies also indicate that the effectiveness of these approaches remains dependent on the language coverage and capabilities of the underlying pretrained models.

Evaluation practices in LLM-based studies remain a significant limitation. Eight of the 11 studies reported concerns about hallucinations or the reliability of generated summaries, while six identified limitations associated with the use of ROUGE for evaluating summaries in non-English and morphologically complex languages [23]. Only three studies incorporated human evaluation, indicating that automatic metrics continue to dominate the assessment of LLM-generated multilingual summaries. These findings suggest that

although instruction tuning, zero-shot/few-shot prompting, and supervised fine-tuning have expanded the applicability of LLMs to multilingual summarization, current evaluation practices do not yet fully capture factuality, fluency, relevance, and language-specific quality. More robust evaluation frameworks that combine automated metrics with human assessment are therefore needed to reliably assess the performance of LLM-based multilingual summarization systems.

4.2. Techniques for Enabling Multilingual Capability in Text Summarization

To address research question two on the techniques used to enable multilingual capability in text summarization models, the analysis drew from the same corpus of 47 studies included in this review. Across the reviewed literature, studies reported a range of techniques for facilitating multilingual functionality in summarization systems. These techniques operate at different levels of the model pipeline, including pre-training, fine-tuning, representation learning, parameter adaptation, and data preparation. The identified techniques were grouped into five categories for structured synthesis. It is important to note that the total frequency counts exceed 47 ($n = 71$ aggregate mentions), as 29 of 47 studies (62%) reported applying two or more technique categories within a single system. However, percentages are calculated relative to the total included study corpus ($n = 47$), as presented in Table 2.

Table 2. Techniques for enabling multilingual capability in text summarization

Technique Category	Count (n = 47)	Representative Methods
Multilingual pre-training	18 (38%)	Sequence-to-sequence denoising, masked language modeling, causal language modeling
Cross-Lingual Transfer Learning	14 (30%)	Zero-shot transfer, few-shot cross-lingual fine-tuning, language-agnostic encoder alignment
Shared Representation Learning	11 (23%)	Shared multilingual vocabularies (SentencePiece/BPE), language-agnostic encoders, joint embedding spaces
Parameter-Efficient Adaptation Methods	16 (34%)	LoRA, prefix-tuning, adapter modules, prompt tuning
Data-Centric & Augmentation Techniques	12 (26%)	Back-translation, parallel corpus construction, cross-lingual data augmentation, and synthetic summary generation

4.2.1. Multilingual Pre-training

Multilingual pre-training was the most frequently reported technique category, documented in 18 of the 47 reviewed studies (38%). Studies in this category reported employing large-scale pretrained multilingual models as backbone architectures for summarization, leveraging representations acquired during pretraining on multilingual corpora. Thirteen of the 18 studies in this category reported using mBART or mBART-50 as the primary pretrained model. Four studies reported the application of mT5, a multilingual variant of the T5 model pretrained with a text-to-text denoising objective across 101 languages. One study [24] reported using XLM-R as the encoder component in a summarization pipeline. With respect to pre-training objectives, 11 of the 18 studies (61%) reported sequence-to-sequence denoising as the primary pre-training strategy. Four studies reported masked language modeling (MLM) as the backbone pretraining objective, which was subsequently adapted for generation [25]. Three studies reported causal language modeling objectives within instruction-tuned or generative model configurations. Studies in this category further reported that pre-training corpus composition varied across the reviewed models, with some explicitly noting limited coverage of low-resource languages in the pre-training data as a characteristic of the employed backbones. The reported target-language counts per study ranged from 2 to 45 languages in this category.

4.2.2. Cross-Lingual Transfer Learning

Cross-lingual transfer learning was reported in 14 of the 47 reviewed studies (30%). Studies in this category described the transfer of summarization capabilities from one or more source languages, predominantly English, to one or more target languages with limited or no target-language training data. Three distinct transfer configurations were observed across the reviewed literature: zero-shot transfer, few-shot transfer, and supervised cross-lingual fine-tuning. Zero-shot cross-lingual transfer was reported in 6 of the 14 studies in this category. These studies reported fine-tuning models exclusively on English-language

summarization data and subsequently evaluating the models on non-English target languages without any target-language supervision. Few-shot transfer configurations were reported in 4 studies, in which a small number of target-language training examples, ranging from 50 to 500 instances across studies, were incorporated during adaptation. Supervised cross-lingual fine-tuning, as explored in [26], in which models pretrained on multilingual corpora were subsequently fine-tuned on parallel or target-language-specific summarization datasets, was reported in the remaining 4 studies. Across the 14 studies, 10 (71%) reported English as the sole transfer source language. Two studies reported multi-source transfer configurations involving more than one high-resource language. The reported target language sets ranged from a single low-resource language to evaluations spanning up to 12 typologically diverse languages.

4.2.3. Shared Representation Learning

Shared representation learning techniques were reported in 11 of the 47 reviewed studies (23%). Studies in this category described approaches in which a unified representational space was constructed to process inputs and produce outputs across multiple languages, without maintaining language-specific model components. Two principal shared representation strategies covered in [27] were identified: shared subword tokenization and joint multilingual embedding spaces. Shared subword tokenization, primarily via Byte-Pair Encoding (BPE) or SentencePiece unigram language model vocabularies, was reported in 9 of the 11 studies as the mechanism for achieving multilingual vocabulary coverage. These studies reported vocabulary sizes ranging from 32,000 tokens to 250,000 tokens, with larger vocabulary configurations reported in studies covering broader language sets. Joint multilingual embedding space learning, in which language-agnostic sentence or token representations are trained using contrastive or alignment objectives, was proposed in [28] and reported in 4 studies, with two of these studies reporting shared tokenization alongside joint embedding training as co-occurring techniques. Studies in this category reported the use of shared representations in both encoder-only and encoder-decoder configurations. Six of the 11 studies (55%) reported shared encoder representations with language-specific or shared decoder modules. The remaining 5 studies reported fully shared encoder-decoder architectures in which no language-specific parameters were maintained, or in which language identity tokens were prepended to guide output generation.

4.2.4. Parameter-Efficient Adaptation Methods

Parameter-efficient adaptation methods were the second most frequently reported technique category, documented in 16 of the 47 reviewed studies (34%). Studies in this category reported techniques that enable the adaptation of large pretrained multilingual models for summarization tasks or specific languages by modifying only a restricted subset of model parameters, rather than performing full fine-tuning [29]. Across all 16 studies in this category, the reported proportions of trainable parameters were consistently below 5% of the total number of model parameters, with exact values explicitly stated in 9 of them. Four distinct parameter-efficient adaptation approaches were identified across the reviewed studies: adapter modules, Low-Rank Adaptation (LoRA), prefix-tuning, and prompt tuning. Adapter module insertion was the most frequently reported parameter-efficient technique, documented in 7 of the 16 studies. These studies reported inserting bottleneck feed-forward adapter layers at each Transformer block, with trainable parameter counts ranging from 0.5% to 3.2% of the total model parameters. LoRA-based adaptation was reported in 5 studies, which described decomposing the weight update matrices of attention layers into low-rank products during fine-tuning. Reported rank values across LoRA-based studies ranged from $r = 4$ to $r = 16$. Prefix-tuning was reported in 3 studies, which described prepending a sequence of trainable continuous vectors to the input of each Transformer layer to condition the model on target-language or task-specific summarization behavior. One study reported using prompt tuning as a soft-prompt adaptation strategy applied exclusively to the input embedding layer.

4.2.5. Data-Centric & Augmentation Techniques

Data-centric and augmentation techniques were reported in 12 of the 47 reviewed studies (26%). Studies in this category described techniques applied at the data pipeline level to expand the availability of multilingual training data for summarization. Four augmentation strategies were identified across the reviewed literature: back-translation, parallel corpus construction, cross-lingual data augmentation via machine translation, and synthetic summary generation. Back-translation was the most frequently reported data augmentation technique, documented in 5 of the 12 studies. These studies described translating monolingual target-language documents into a high-resource source language for model training, or generating target-language summaries by translating source-language summaries produced by a pretrained model, as documented in [30]. Parallel corpus construction involving the alignment of document-summary pairs across language pairs was reported in 4 studies. Two of these studies reported constructing new multilingual summarization datasets as a primary contribution alongside the proposed technique.

Cross-lingual data augmentation via neural machine translation (NMT) systems was reported in 3 studies, which described translating existing monolingual summarization training corpora into one or more target languages to create synthetic multilingual training sets. Synthetic summary generation, in which a pretrained summarization model is used to produce pseudo-reference summaries for unannotated multilingual documents, according to [31], was reported in 2 studies. One study reported the combined use of back-translation and synthetic summary generation within a single data preparation pipeline. In all 12 studies in this category, the reported augmented dataset sizes ranged from approximately 5,000 synthetic instances to over 200,000 augmented document–summary pairs.

4.3. Limitations and Research Gaps in Multilingual Text Summarization

In the third research question, the review examined the limitations and research gaps reported in current multilingual text summarization research. The extracted findings were synthesized into four broad categories: low-resource language coverage, dataset and benchmark limitations, evaluation limitations, and generalization and scalability challenges. As was the case in Section 4.2, several studies reported limitations spanning multiple categories. Aggregate counts, therefore, exceed 47, but all percentages were calculated with respect to the total included corpus ($n = 47$). Table 3 presents the taxonomy of the identified limitations and research gaps.

Table 3. Limitations and Research Gaps in Multilingual Text Summarization

Gap / Limitation Category	Count ($n = 47$)	Key Sub-Patterns
Low-Resource Language Coverage	31 (66%)	English-centric bias; under-represented script coverage; low-resource exclusion
Dataset and Benchmark Limitations	27 (57%)	News-domain bias; corpus imbalance; no unified benchmark
Evaluation Limitations	34 (72%)	ROUGE over-reliance; scarce human evaluation; protocol inconsistency
Generalization and Scalability Challenges	29 (62%)	Curse of multilinguality; cross-domain degradation; computational cost

4.3.1. Low-Resource Language Coverage

Low-resource language coverage was the most broadly reported limitation category, documented in 31 of the 47 reviewed studies (66%). Language-wise analysis shows that the reviewed literature remains strongly concentrated on high-resource languages, particularly English, Chinese, French, German, and Spanish. Among the 31 studies reporting limitations related to low-resource language coverage, 24 were associated with the dominance of high-resource pre-training or evaluation languages, while 9 specifically reported an English-centric development or evaluation bias. This concentration limits the extent to which multilingual summarization models are evaluated across linguistically diverse settings. In particular, 12 studies reported limited representation or absence of non-Latin-script languages, including Arabic, Hindi, Amharic, and languages from Southeast Asia. These findings indicate that multilingual coverage remains uneven, with substantial attention given to languages for which larger training and evaluation resources are already available.

The analysis also identified language-specific challenges beyond simple differences in the number of languages represented. Six studies reported vocabulary fragmentation or tokenization-related difficulties affecting languages such as Finnish, Turkish, Swahili, and Yoruba, while four studies associated the under-representation of low-resource languages in pre-training data with lower automatic evaluation scores [32]. However, the latter relationship was reported as an association rather than a causal effect. Overall, the findings suggest that the low-resource language gap involves multiple factors, including limited training resources, inadequate representation of non-Latin scripts, and difficulties in representing morphologically diverse languages. Consequently, conclusions about multilingual summarization performance remain considerably stronger for high-resource languages than for low-resource languages.

4.3.2. Dataset and Benchmark Limitations

Dataset and benchmark limitations were reported in 27 of the 47 reviewed studies (57%). The dataset-wise analysis indicates that the evidence base is also concentrated around a relatively limited range of datasets and domains. Of the 27 studies reporting dataset and benchmark limitations, 21 identified a

predominance of news-oriented data, with datasets such as XL-Sum, MLSum, WikiLingua, and CNN/DailyMail appearing among the datasets used in the reviewed studies [33]. Furthermore, 15 of these 21 studies relied primarily on a single dataset, limiting opportunities to assess whether model performance remains consistent across different domains and data characteristics. This concentration is particularly important for multilingual summarization because performance observed in a single domain may not adequately reflect models' ability to summarize content across different languages, topics, and document types.

Language coverage and benchmark standardization were also identified as important dataset-level limitations. Six studies reported skewed language distributions, while eight studies used or discussed datasets that covered fewer than 10 languages. In addition, 13 studies reported the absence of a standardized multilingual benchmark, making direct comparison between models and approaches more difficult [34]. Five studies also identified issues related to noise or inconsistent data quality, while three developed new datasets because existing resources were considered insufficient for their research objectives. Taken together, these findings show that dataset limitations extend beyond the availability of multilingual corpora to include uneven language representation, dependence on individual datasets, differences in data quality, and the lack of standardized benchmarks for evaluating multilingual summarization systems.

4.3.3. Evaluation Limitations

Evaluation limitations constituted the most frequently documented limitation category in the reviewed corpus, reported in 34 of the 47 studies (72%). Over-reliance on ROUGE-based automatic evaluation metrics was the dominant pattern, observed in 30 of the 34 studies in this category. Of these, 18 studies (60% of the ROUGE-reliant subset) reported ROUGE-1, ROUGE-2, and ROUGE-L as the sole evaluation metrics, with no supplementary metric or human assessment included. Fourteen studies explicitly acknowledged that ROUGE scores exhibit reduced sensitivity to lexical and morphological variation across languages, a limitation reported in [35], whereby n-gram overlap-based scoring may inadequately capture summary quality in morphologically rich or agglutinative target languages.

Human evaluation was reported in only 11 of the 47 reviewed studies (23%). Of these, 7 studies (64% of the human-evaluation subset) employed human assessment as a supplementary measure alongside automatic metrics, while 4 studies reported human evaluation as the primary assessment protocol. Evaluation criteria employed in human assessment varied across studies: fluency and coherence were assessed in all 11 studies, factual consistency in 6 studies, and relevance in 4 studies. Inconsistent evaluation protocols across the reviewed corpus were reported as a limitation in 19 studies. These studies documented variability in evaluated language sets, summary length constraints, tokenization procedures for metric computation, and test split definitions across the literature, with 8 studies explicitly stating that these inconsistencies prohibited direct numerical comparison of reported scores with prior work [36]. Five studies additionally reported that BERTScore and METEOR were employed as supplementary automatic metrics to partially address ROUGE limitations, though cross-study comparability remained constrained by inconsistent metric configurations.

4.3.4. Generalization and Scalability Challenges

Generalization and scalability challenges were reported in 29 of the 47 reviewed studies (62%). Within this category, the curse of multilinguality was the most consistently documented recurring phenomenon, reported explicitly in 22 of the 29 studies (76%). The reviewed literature also identified the curse of multilinguality as a scalability challenge, whereby increasing the number of supported languages can reduce per-language performance under fixed model capacity. These studies reported that as the number of languages supported by a single model increased, per-language task performance declined, a pattern attributed to fixed-capacity model parameters being distributed across an expanding set of languages [37]. Fifteen of the 22 studies documenting the curse of multilinguality reported this degradation in the context of models supporting more than 15 languages simultaneously. Seven studies (32% of this subset) reported that the degradation was most pronounced for low-resource languages with limited representation in the pre-training corpus, consistent with the study in [2] about low-resource coverage limitations.

Cross-domain generalization limitations were reported in 17 of the 29 studies (59%) in this category. Fourteen of these studies (82% of the cross-domain subset) reported that models trained and evaluated exclusively on news-domain corpora exhibited performance degradation when applied to out-of-domain text types, including scientific abstracts, legal documents, and conversational transcripts [38]. Cross-lingual generalization to unseen language families was reported as a limitation in 13 studies (45% of this category), with 8 of these studies reporting that zero-shot transfer to typologically distant languages yielded substantially lower automatic metric scores than transfer within the same language family. Computational constraints were explicitly reported as a scalability limitation in 18 of the 29 studies. Of these, 11 studies (61% of the computation-limited subset) reported that full fine-tuning of large multilingual models,

particularly those exceeding 600M parameters, was computationally infeasible under the reported experimental resource constraints, with GPU memory limitations cited as the primary constraint in 7 studies. Four studies also reported that multilingual scaling in extending a trained summarization model to additional languages post-training introduced catastrophic forgetting of previously learned language representations, with [39] identifying this as a distinct scalability challenge.

5. DISCUSSION

This section synthesizes findings across the reviewed literature in direct correspondence with the three research questions guiding this review. The discussion is organized thematically rather than source-by-source to cover the dominant trends, methodological patterns, and unresolved challenges that characterize the current state of multilingual text summarization research. The subsections that follow interpret the results of each research question before integrating the findings into a unified cross-cutting synthesis.

5.1. Architectural Trends in Multilingual Text Summarization

The architectural distribution across the reviewed studies indicates a field that is still in transition rather than one that has converged on a stable design paradigm. Multilingual sequence-to-sequence models remain the dominant approach, largely reflecting the influence of the Transformer encoder-decoder framework and its adaptation to multilingual settings through pretrained models such as mBART and mT5 [40]. Their continued prevalence suggests that most research still relies on architectures that explicitly encode multilinguality through design choices made during pre-training. At the same time, the growing presence of large language model-based approaches, particularly in recent studies, signals a shift in how multilingual summarization capabilities are achieved. Instead of being explicitly engineered through multilingual training objectives, capability is increasingly derived from scale and broad data exposure, with models applied in zero-shot and few-shot settings.

This shift highlights an important difference in how multilinguality is represented across architecture classes. Sequence-to-sequence models rely on structured multilingual pre-training, in which language balance and shared representations are deliberately constructed [41]. In contrast, large language models exhibit multilingual ability as an emergent property, making their performance more sensitive to the distribution of languages in pre-training data and often favoring high-resource languages. This helps explain the consistent evaluation bias toward high-resource language pairs observed across the literature. Meanwhile, adapter-based models and hybrid architectures occupy more targeted roles within the landscape. Adapter-based approaches are mainly used where computational efficiency and low-resource adaptation are required. At the same time, hybrid architectures extend sequence-to-sequence models with retrieval or structured components to improve factuality and long-document handling [42]. Together, these patterns suggest that the field is being shaped by a trade-off between scalability and equitable multilingual coverage, where increases in model capability do not necessarily translate into balanced performance across languages.

5.2. Effectiveness of Techniques Supporting Multilingual Capability

The techniques used to enable multilingual capability in the reviewed studies show a strong convergence toward a shared methodological pipeline rather than a diverse set of independent strategies. Multilingual pre-training and parameter-efficient adaptation form the core of this pipeline and are frequently used together. In most cases, pre-training provides a shared multilingual representation space, while adaptation methods such as fine-tuning, LoRA, or adapter modules adjust the model to specific languages or tasks without requiring full retraining. This separation of roles makes the approach computationally efficient, as per [43]. Still, it also means that multilingual performance is largely constrained by the properties of the pre-training data, which are not controlled at the level of individual studies.

Cross-lingual transfer learning operates in a closely related yet distinct manner, explicitly leveraging high-resource languages to support low-resource ones [44]. Across the reviewed studies, this approach relies heavily on English as the primary source of transfer, suggesting that transfer is often driven more by data availability than by a balanced multilingual design strategy. While effective in expanding coverage to low-resource settings, this dependence introduces a structural bias that is difficult to correct during evaluation, particularly in zero-shot and few-shot scenarios [45]. Shared representation learning and data augmentation techniques play a supporting role in this pipeline. Shared vocabularies enable cross-lingual processing within a single model, while augmentation methods such as back-translation are primarily used to mitigate the lack of parallel or annotated data [46]. Their widespread use highlights a persistent constraint in the field: not only in modeling capacity but also in the availability and structure of multilingual training data.

Across all technique categories, the findings suggest that multilingual summarization systems operate as layered pipelines, in which each stage relies on assumptions established earlier in the Process. This dependency is relatively stable in high-resource languages. Still, it becomes fragile in low-resource settings,

where data scarcity and linguistic imbalance undermine the effectiveness of both pre-training and transfer-based strategies. The key limitation, therefore, lies not in individual techniques but in the uneven conditions under which they operate, which directly shape the reliability of multilingual performance across languages.

5.3. Persistent Challenges and Research Gaps

The limitations identified across the reviewed studies form a tightly coupled system rather than isolated methodological issues. Low-resource language underrepresentation, dataset bias, evaluation constraints, and generalization difficulties reinforce one another through shared dependencies on data availability and model design assumptions [47][48]. The strong concentration of evaluation on high-resource languages creates a feedback loop in which model development, benchmarking, and reported performance are all shaped by the same limited linguistic coverage [49]. This imbalance is further reinforced by representational choices such as shared subword vocabularies, which tend to favor high-resource and typologically similar languages while introducing inefficiencies for morphologically rich languages. In parallel, dataset limitations, particularly the dominance of news-domain corpora and the lack of standardized multilingual benchmarks, restrict the generalizability of reported findings and make cross-study comparisons methodologically unstable.

These structural issues are compounded at the evaluation level, where reliance on ROUGE-based metrics limits the ability to capture semantic adequacy across linguistically diverse outputs, particularly for morphologically complex languages [50][51]. Although some studies supplemented ROUGE with semantically oriented measures such as BERTScore and METEOR, their use remained limited in the reviewed literature. Similarly, human evaluation was reported in only a minority of studies, with assessments focusing on dimensions such as fluency, coherence, factual consistency, and relevance. The limited and inconsistent use of these complementary evaluation approaches is particularly concerning for large language model-based systems, where hallucinations and factual reliability may not be adequately captured by lexical-overlap metrics. A more comprehensive evaluation framework should therefore combine lexical, semantic, factual-consistency, and human-centered assessment to provide a more reliable measure of multilingual summary quality.

Within this broader context, the curse of multilinguality emerges as the central scalability constraint, in which increasing the number of supported languages degrades per-language performance under fixed model capacity [52]. Although parameter-efficient adaptation methods partially mitigate this trade-off through selective specialization, as observed in [53], they do not resolve the underlying trade-off between linguistic coverage and representational capacity, indicating that the core limitation is architectural and data-structural rather than purely methodological. These unresolved challenges point to several concrete technical directions for future research. These include developing better representations and adaptation methods for low-resource languages, creating more balanced and standardized multilingual benchmarks, designing evaluation frameworks that combine semantic, factual, and human assessment, and developing scalable architectures that reduce the trade-off between language coverage and per-language performance. Addressing these problems could improve the robustness, fairness, and generalizability of multilingual summarization systems.

5.4. Overall Synthesis of the Findings

Integrating the findings across all three research questions reveals a persistent structural limitation in multilingual text summarization. The methods that currently dominate the field, including large pretrained models, transfer learning from high-resource languages, and evaluation on news-based benchmarks, are largely optimized for a limited subset of global languages. In contrast, low-resource and typologically diverse languages remain systematically underserved. This imbalance is reflected across all levels of the research pipeline. Architecturally, the shift from seq2seq models to large language models improves performance in high-resource settings but also amplifies pre-existing representational biases inherited from training data [54]. At the methodological level, the reliance on pre-training and transfer learning means that multilingual capability is largely data-driven rather than explicitly designed, which inherently favors languages with abundant textual resources [55]. At the evaluation level, the dominance of ROUGE-based metrics and news-domain datasets further narrows what is considered meaningful progress, as these setups fail to capture important dimensions such as factual consistency, morphological variation, and cross-domain robustness.

This structural imbalance is reinforced by the curse of multilinguality, which highlights a fundamental scalability trade-off in unified multilingual models. Increasing language coverage often reduces per-language performance at fixed model capacity. Although parameter-efficient adaptation techniques partially mitigate this issue, their effectiveness remains constrained by the quality and balance of pretrained representations [56]. A further unifying constraint is dataset dependency, as the availability of suitable multilingual corpora strongly shapes the field's progress. The concentration of resources in high-resource

languages and the lack of standardized benchmarks limit both training diversity and evaluation reliability, making it difficult to separate genuine methodological improvements from dataset-driven effects [57]. Cross-study numerical comparisons should therefore be interpreted with caution because reported performance values are obtained across different datasets, language combinations, model configurations, training conditions, and evaluation protocols. A higher score reported by one study cannot necessarily be interpreted as evidence of a superior approach when the underlying experimental conditions differ. Collectively, these findings suggest that future research will depend less on isolated architectural or algorithmic innovations, but more on addressing structural inequalities in data availability, evaluation design, and multilingual benchmarking infrastructure.

6. CONCLUSION

This study synthesized evidence from 47 studies on multilingual text summarization, focusing on model architectures, multilingual enabling techniques, and associated limitations and research gaps. For RQ1, the findings show that multilingual sequence-to-sequence models remain the dominant architectural paradigm, while adapter-based methods and large language models are increasingly emerging as scalable alternatives. For RQ2, multilingual capability is primarily achieved through large-scale pre-training and cross-lingual transfer learning, supported by parameter-efficient adaptation, shared representation learning, data augmentation, and prompting strategies. In RQ3, the review identifies that despite these advances, persistent challenges remain, particularly in achieving balanced performance across languages, handling morphological diversity, ensuring factual consistency, and developing robust multilingual evaluation frameworks. The primary contribution of this review is a structured, evidence-based characterization of these architectural trends, methodological strategies, and systemic limitations, establishing a synthesis-level foundation for the field.

This review is subject to a methodological limitation arising from the inherent heterogeneity of studies that span diverse architectures, language settings, datasets, evaluation protocols, and task formulations, as is the case in multilingual text summarization. While this diversity is essential for capturing the breadth of the field, it also implies that the findings are interpretive rather than directly comparable through objective quantitative aggregation. To this end, a key unresolved question remains the lack of a principled basis for determining the optimal architectural design for multilingual text summarization models. Existing approaches largely rely on empirical comparisons within limited language settings, which may not generalize to broader multilingual distributions. Addressing this gap calls for future work that moves beyond isolated model comparisons toward an empirical analysis of cross-lingual transfer dynamics. Controlled evaluation frameworks that explicitly link architectural design choices to language diversity and data availability also form a future research direction.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article, as no datasets were generated or analyzed during the current study

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

- [1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: [10.1007/s11042-022-13428-4](https://doi.org/10.1007/s11042-022-13428-4).
- [2] M. Munaf, H. Afzal, K. Mahmood, and N. Iltaf, "Low resource summarization using pre-trained language models," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 10, Art. no. 141, pp. 1–19, Oct. 2024, doi: [10.1145/3675780](https://doi.org/10.1145/3675780).
- [3] H. Zhang, P. S. Yu, and J. Zhang, "A systematic survey of text summarization: From statistical methods to large language models," *ACM Comput. Surv.*, vol. 57, no. 11, Art. no. 277, 2025, doi: [10.1145/3731445](https://doi.org/10.1145/3731445).
- [4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 7871–7880, doi: [10.18653/v1/2020.acl-main.703](https://doi.org/10.18653/v1/2020.acl-main.703).
- [5] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2021, pp. 483–498, doi: [10.18653/v1/2021.naacl-main.41](https://doi.org/10.18653/v1/2021.naacl-main.41).
- [6] T. Hasan *et al.*, "XL-Sum: Large-scale multilingual abstractive summarization for 44 languages," in *Findings Assoc. Comput. Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4693–4703, doi: [10.18653/v1/2021.findings-acl.413](https://doi.org/10.18653/v1/2021.findings-acl.413).
- [7] K. Huang *et al.*, "A survey on large language models with multilingualism: Recent advances and new frontiers," *Artif. Intell. Rev.*, vol. 59, Art. no. 146, 2026, doi: [10.1007/s10462-026-11534-5](https://doi.org/10.1007/s10462-026-11534-5).

- [8] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, vol. 603, Art. no. 128255, Oct. 2024, doi: [10.1016/j.neucom.2024.128255](https://doi.org/10.1016/j.neucom.2024.128255).
- [9] X. Zhang, "The path of multilingual neural machine translation: Cross-language transfer and domain adaptation," *Procedia Comput. Sci.*, vol. 279, pp. 327–336, 2026, doi: [10.1016/j.procs.2026.03.241](https://doi.org/10.1016/j.procs.2026.03.241).
- [10] B. Nath and Y. Gulzar, "Zero-shot English–Assamese neural machine translation via pivot-based cross-lingual embedding alignment and transfer learning," *Sci. Rep.*, vol. 16, Art. no. 13732, 2026, doi: [10.1038/s41598-026-44209-w](https://doi.org/10.1038/s41598-026-44209-w).
- [11] G. Park, J. Park, and H. Lee, "Cross-lingual summarization for low-resource languages using multilingual retrieval-based in-context learning," *Appl. Sci.*, vol. 15, no. 14, Art. no. 7800, Jul. 2025, doi: [10.3390/app15147800](https://doi.org/10.3390/app15147800).
- [12] A. Ghabi and H. Hamam, "Architectural deep dive into large language models," in *Generative AI and Large Language Models: Opportunities, Challenges, and Applications*, vol. 1, A. Koubaa, A. Ammar, L. Ghouti, W. Boulila, and B. Benjdira, Eds. Cham, Switzerland: Springer Nature Switzerland, 2025, pp. 65–105, doi: [10.1007/978-3-031-90573-5_3](https://doi.org/10.1007/978-3-031-90573-5_3).
- [13] J. Wang *et al.*, "A survey on cross-lingual summarization," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 1304–1323, 2022, doi: [10.1162/tacl_a_00520](https://doi.org/10.1162/tacl_a_00520).
- [14] K. P. Sharma *et al.*, "A systematic review on text summarization: Techniques, challenges, opportunities," *Expert Syst.*, vol. 42, no. 4, Art. no. e13833, 2025, doi: [10.1111/exsy.13833](https://doi.org/10.1111/exsy.13833).
- [15] M. Parikh, A. Gandhi, and R. Tripathi, "Approaches for summarization of articles across multiple languages: A review," in *Proc. 2025 Int. Conf. Mach. Learn. Auton. Syst. (ICMLAS)*, 2025, pp. 1150–1154, doi: [10.1109/ICMLAS64557.2025.10968896](https://doi.org/10.1109/ICMLAS64557.2025.10968896).
- [16] A. C. Tricco *et al.*, "Updating the PRISMA reporting guideline for scoping reviews: A scoping review," *J. Clin. Epidemiol.*, vol. 196, Art. no. 112314, 2026, doi: [10.1016/j.jclinepi.2026.112314](https://doi.org/10.1016/j.jclinepi.2026.112314).
- [17] I. G. A. I. U. Putri, I. N. P. Trisna, and N. K. D. Rusjayanthi, "Abstractive text summarization to generate Indonesian news highlight using transformers model," *J. Inf. Syst. Informatics*, vol. 7, no. 2, pp. 1248–1263, Jun. 2025, doi: [10.51519/journalisi.v7i2.1082](https://doi.org/10.51519/journalisi.v7i2.1082).
- [18] S. Doddapaneni, M. S. U. R. Khan, D. Venkatesh, R. Dabre, A. Kunchukuttan, and M. M. Khapra, "Cross-lingual auto evaluation for assessing multilingual LLMs," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, Vienna, Austria, 2025, pp. 29297–29329, doi: [10.18653/v1/2025.acl-long.1419](https://doi.org/10.18653/v1/2025.acl-long.1419).
- [19] S. Ajay Mukund and K. S. Easwarakumar, "Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation," *Symmetry*, vol. 17, no. 5, Art. no. 633, May 2025, doi: [10.3390/sym17050633](https://doi.org/10.3390/sym17050633).
- [20] R. Pasunuru, M. Liu, M. Bansal, S. Ravi, and M. Dreyer, "Efficiently summarizing text and graph encodings of multi-document clusters," in *Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2021, pp. 4768–4779, doi: [10.18653/v1/2021.naacl-main.380](https://doi.org/10.18653/v1/2021.naacl-main.380).
- [21] B. Dong, C. Bu, Y. Zhu, S. Ji, and X. Wu, "Simplified multi-view graph neural network for multilingual knowledge graph completion," *Front. Comput. Sci.*, vol. 19, no. 7, Art. no. 197324, 2024, doi: [10.1007/s11704-024-3577-3](https://doi.org/10.1007/s11704-024-3577-3).
- [22] G. Balde, S. Roy, M. Mondal, and N. Ganguly, "Learning faster with better tokens: Parameter-efficient vocabulary adaptation for specialized text summarization," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, San Diego, CA, USA, 2026, pp. 45071–45086, doi: [10.18653/v1/2026.acl-long.2088](https://doi.org/10.18653/v1/2026.acl-long.2088).
- [23] I. Mondshine, T. Paz-Argaman, and R. Tsarfaty, "Beyond n-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, 2025, pp. 19019–19035, doi: [10.18653/v1/2025.acl-long.932](https://doi.org/10.18653/v1/2025.acl-long.932).
- [24] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 8440–8451, doi: [10.18653/v1/2020.acl-main.747](https://doi.org/10.18653/v1/2020.acl-main.747).
- [25] S. Rothe, J. Maynez, and S. Narayan, "A thorough evaluation of task-specific pretraining for summarization," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 140–145, doi: [10.18653/v1/2021.emnlp-main.12](https://doi.org/10.18653/v1/2021.emnlp-main.12).
- [26] Y. Ye *et al.*, "CC-Tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria, 2025, pp. 19036–19051, doi: [10.18653/v1/2025.acl-long.933](https://doi.org/10.18653/v1/2025.acl-long.933).
- [27] J. Daðason and H. Loftsson, "Subword tokenization for low- and medium-resource languages: A systematic evaluation," in *Proc. 35th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2026, pp. 5684–5692, doi: [10.24963/ijcai.2026/633](https://doi.org/10.24963/ijcai.2026/633).
- [28] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, Dublin, Ireland, 2022, pp. 878–891, doi: [10.18653/v1/2022.acl-long.62](https://doi.org/10.18653/v1/2022.acl-long.62).
- [29] L. Wang *et al.*, "Parameter-efficient fine-tuning in large language models: A survey of methodologies," *Artif. Intell. Rev.*, vol. 58, no. 8, Art. no. 227, 2025, doi: [10.1007/s10462-025-11236-4](https://doi.org/10.1007/s10462-025-11236-4).
- [30] J. Nicolas, V. Gunawan, A. Sutanto, and Hidayatullah, "Evaluation of the effect of back-translation direction on machine translation quality for low-resource Sundanese-English language," *Procedia Comput. Sci.*, vol. 269, pp. 784–796, 2025, doi: [10.1016/j.procs.2025.09.021](https://doi.org/10.1016/j.procs.2025.09.021).
- [31] M. Nadăș, L. Dioșan, and A. Tomescu, "Synthetic data generation using large language models: Advances in text and code," *IEEE Access*, vol. 13, pp. 134615–134633, 2025, doi: [10.1109/ACCESS.2025.3589503](https://doi.org/10.1109/ACCESS.2025.3589503).
- [32] J. Karthikeyan, M. Y. Sayed, Y. Waykar, V. Sathya, K. P. R., and M. Suseela Grace Padma, "Natural language processing for multilingual and low-resource languages," in *Proc. 2025 IEEE 7th Int. Conf. Comput., Commun. Autom. (ICCCA)*, 2025, pp. 1–7, doi: [10.1109/ICCCA66364.2025.11325195](https://doi.org/10.1109/ICCCA66364.2025.11325195).

- [33] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 39–57, 2024, doi: [10.1162/tacl_a_00632](https://doi.org/10.1162/tacl_a_00632).
- [34] M. Bhandari, P. N. Gour, A. Ashfaq, P. Liu, and G. Neubig, "Re-evaluating evaluation in text summarization," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., 2020, pp. 9347–9359, doi: [10.18653/v1/2020.emnlp-main.751](https://doi.org/10.18653/v1/2020.emnlp-main.751).
- [35] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating summarization evaluation," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 391–409, 2021, doi: [10.1162/tacl_a_00373](https://doi.org/10.1162/tacl_a_00373).
- [36] Y. Liu, A. Fabbri, P. Liu, Y. Zhao, L. Nan, R. Han, S. Han, S. Joty, C.-S. Wu, C. Xiong, and D. Radev, "Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, Toronto, Canada, 2023, pp. 4140–4170, doi: [10.18653/v1/2023.acl-long.228](https://doi.org/10.18653/v1/2023.acl-long.228).
- [37] U. Shaham, M. Elbayad, V. Goswami, O. Levy, and S. Bhosale, "Causes and cures for interference in multilingual translation," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 1: Long Papers, Toronto, Canada, 2023, pp. 15849–15863, doi: [10.18653/v1/2023.acl-long.883](https://doi.org/10.18653/v1/2023.acl-long.883).
- [38] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 4623–4637, doi: [10.18653/v1/2020.acl-main.421](https://doi.org/10.18653/v1/2020.acl-main.421).
- [39] N. Muennighoff *et al.*, "Scaling data-constrained language models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, doi: [10.52202/075280-2191](https://doi.org/10.52202/075280-2191).
- [40] A. D. Kotkar, R. S. Mahadik, P. G. More, and S. A. Thorat, "Comparative analysis of transformer-based large language models (LLMs) for text summarization," in *Proc. 2024 1st Int. Conf. Adv. Comput. Emerg. Technol. (ACET)*, 2024, pp. 1–7, doi: [10.1109/ACET61898.2024.10730348](https://doi.org/10.1109/ACET61898.2024.10730348).
- [41] M. La Quatra and L. Cagliero, "BART-IT: An efficient sequence-to-sequence model for Italian text summarization," *Future Internet*, vol. 15, no. 1, Art. no. 15, 2023, doi: [10.3390/fi15010015](https://doi.org/10.3390/fi15010015).
- [42] R. Aharoni, S. Narayan, J. Maynez, J. Herzig, E. Clark, and L. Lapata, "Multilingual summarization with factual consistency evaluation," in *Findings Assoc. Comput. Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada, 2023, pp. 3562–3591, doi: [10.18653/v1/2023.findings-acl.220](https://doi.org/10.18653/v1/2023.findings-acl.220).
- [43] O. Razuvayevskaya, B. Wu, J. A. Leite, F. Heppell, I. Srba, C. Scarton, K. Bontcheva, and X. Song, "Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification," *PLOS ONE*, vol. 19, no. 5, Art. no. e0301738, 2024, doi: [10.1371/journal.pone.0301738](https://doi.org/10.1371/journal.pone.0301738).
- [44] Y. Bai, Y. Gao, and H. Huang, "Cross-lingual abstractive summarization with limited parallel resources," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL) and 11th Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, vol. 1: Long Papers, 2021, pp. 6910–6924, doi: [10.18653/v1/2021.acl-long.538](https://doi.org/10.18653/v1/2021.acl-long.538).
- [45] R. Hada *et al.*, "Are large language model-based evaluators the solution to scaling up multilingual evaluation?" in *Findings Assoc. Comput. Linguistics: EACL 2024*, St. Julian's, Malta, 2024, pp. 1051–1070, doi: [10.18653/v1/2024.findings-eacl.71](https://doi.org/10.18653/v1/2024.findings-eacl.71).
- [46] R. Wijayanti, M. L. Khodra, K. Surendro, and D. H. Widyantoro, "Learning bilingual word embedding for automatic text summarization in low-resource language," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 4, pp. 224–235, 2023, doi: [10.1016/j.jksuci.2023.03.015](https://doi.org/10.1016/j.jksuci.2023.03.015).
- [47] A. Kaushik, S. H. Attri, and R. S. Jha, "Exploring text summarization techniques: A review of current challenges and future directions," in *Proc. 2024 2nd Int. Conf. Disruptive Technol. (ICDT)*, 2024, pp. 289–295, doi: [10.1109/ICDT61202.2024.10489243](https://doi.org/10.1109/ICDT61202.2024.10489243).
- [48] D. Das, S. Guha, and B. Semaan, "How do datasets, developers, and models affect biases in a low-resourced language? The case of the Bengali language," in *Proc. 2026 ACM SIGCAS/SIGCHI Conf. Comput. Sustain. Soc. (COMPASS)*, 2026, pp. 340–357, doi: [10.1145/3811242.3819100](https://doi.org/10.1145/3811242.3819100).
- [49] H. Schroeder, D. Roy, and J. Kabbara, "Just put a human in the loop? Investigating LLM-assisted annotation for subjective tasks," in *Findings Assoc. Comput. Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria, 2025, pp. 25771–25795, doi: [10.18653/v1/2025.findings-acl.1323](https://doi.org/10.18653/v1/2025.findings-acl.1323).
- [50] F. Franzon and C. Zanini, "The entropy of morphological systems in natural languages is modulated by functional and semantic properties," *J. Quant. Linguistics*, vol. 30, no. 1, pp. 42–66, 2023, doi: [10.1080/09296174.2022.2063501](https://doi.org/10.1080/09296174.2022.2063501).
- [51] J. Z. Forde, R. Zhang, L. Sutawika, A. F. Aji, S. Cahyawijaya, G. I. Winata, M. Wu, C. Eickhoff, S. Biderman, and E. Pavlick, "Re-evaluating evaluation for multilingual summarization," in *Proc. 2024 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Miami, FL, USA, 2024, pp. 19476–19493, doi: [10.18653/v1/2024.emnlp-main.1085](https://doi.org/10.18653/v1/2024.emnlp-main.1085).
- [52] T. A. Chang, C. Arnett, Z. Tu, and B. K. Bergen, "When is multilinguality a curse? Language modeling for 250 high- and low-resource languages," in *Proc. 2024 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, FL, USA, 2024, pp. 4074–4096, doi: [10.18653/v1/2024.emnlp-main.236](https://doi.org/10.18653/v1/2024.emnlp-main.236).
- [53] J. Pfeiffer *et al.*, "Lifting the curse of multilinguality by pre-training modular transformers," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, WA, USA, 2022, pp. 3479–3495, doi: [10.18653/v1/2022.naacl-main.255](https://doi.org/10.18653/v1/2022.naacl-main.255).

- [54] V. Yogarajan, G. Dobbie, and T. T. Keegan, "Debiasing large language models: Research opportunities," *J. Roy. Soc. New Zealand*, vol. 55, no. 2, pp. 372–395, 2025, doi: [10.1080/03036758.2024.2398567](https://doi.org/10.1080/03036758.2024.2398567).
- [55] H. A. Pandya and B. S. Bhatt, "Does learning from language family help? A case study on a low-resource question-answering task," *Natural Lang. Process.*, vol. 31, no. 2, pp. 375–392, 2025, doi: [10.1017/nlp.2024.13](https://doi.org/10.1017/nlp.2024.13).
- [56] L. Xu, H. Xie, S. J. Qin, X. Tao, and F. L. Wang, "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 6, pp. 6107–6126, Jun. 2026, doi: [10.1109/TPAMI.2026.3657354](https://doi.org/10.1109/TPAMI.2026.3657354).
- [57] T. R. McIntosh *et al.*, "Inadequacies of large language model benchmarks in the era of generative artificial intelligence," *IEEE Trans. Artif. Intell.*, vol. 7, no. 1, pp. 22–39, 2026, doi: [10.1109/TAI.2025.3569516](https://doi.org/10.1109/TAI.2025.3569516).

BIOGRAPHIES OF AUTHORS



Charles Munyao Muia is a PhD Computer Science student at Murang'a University of Technology. He earned his BSc and MSc in Computer Science from Murang'a University of Technology in 2021 and 2024, respectively. He is a student member of the Association for Computing Machinery (ACM) and a professional member of the International Association of Engineers (IAENG). His research focuses on Artificial Intelligence, particularly Natural Language Processing (NLP) and Large Language Models (LLMs). He can be contacted at email: munyao@mut.ac.ke



Aaron Mogeni Oirere is a senior lecturer and the Chair of the Department of Computer Science at the Murang'a University of Technology. He received his BSc. Computer Science from Periyar University, India, in 2007, MSc. Computer Science from Bharathiar University, India, in 2010, and Ph.D. Computer Science from Dr. Babasaheb Ambedkar Marathwada University, India, in 2016. His research focuses on computational linguistics and Natural Language Processing. He can be contacted at email: amogeni@mut.ac.ke



Stephen Kahara Wanjau holds a PhD in Computer Science from Murang'a University of Technology (MUT), an M.Sc. in Computer Systems (JKUAT), an M.A. in Organizational Development (USIU-A), and a B.Sc. in Information Sciences (Moi). Currently serving as the Director of Performance Management and Standards at MUT, he is a distinguished ICT strategist and certified QMS/ISMS Lead Auditor with over fifteen years of experience in operational excellence and information security. His research and professional practice focus on leveraging Artificial Intelligence, Machine Learning, Network Security, and Distributed Systems. A DAAD-UNILEAD alumnus and member of the Association of Computing Practitioners-Kenya (ACPK). Dr. Kahara bridges the gap between rigorous international standards and impactful technological innovation to drive institutional development. He can be contacted at email: skahara@mut.ac.ke