International Journal of Mathematics, Statistics, and Computer Science
Vol. 4, 2026, pp. 595∼607
eISSN 2704-1069, DOI: https://doi.org/10.59543/14d9kq02

Hierarchical Curriculum Transfer for Swahili QA with mT5: SQuAD Pretraining and KenSwQuAD Extractive-to-Abstractive Refinement

Jared Ongaro[1*], Benjamin Kikwai[2]
[1]Department of Mathematics, University of Nairobi, Kenya
[2]Department of Mathematics and Statistics, Machakos University, Kenya

*Corresponding author: ongaro@uonbi.ac.ke

Received 06.06.2026, Revised 02.08.2026, Accepted 10.08.2026

ABSTRACT: The introduction of the *Kencorpus Swahili Question Answering Dataset* (KenSwQuAD) presents a compelling opportunity to advance Machine Reading Comprehension (MRC) for the Swahili language. However, the mixed composition of this dataset—comprising 67.6% extractive and 32.4% abstractive answers—introduces substantial hurdles for standard training pipelines. In this study, we benchmark the performance of the multilingual T5 (mT5) architecture on KenSwQuAD, deploying a robust *Hierarchical Curriculum Learning* strategy. We introduce a novel data restructuring technique that algorithmically partitions the corpus into distinct extractive and abstractive subsets, thereby enabling meticulously phased fine-tuning. By sequentially progressing from structural transfer via the English SQuAD, to morphological alignment on the Swahili extractive subset, and culminating in abstractive refinement, we establish that the mT5-base model can achieve a SacreBLEU score of 48.99 on extractive tasks. For abstractive reasoning, whilst the BLEU score is comparatively modest (15.52), the model attains a remarkable BERTScore F1 of 77.21%, reflecting a profound degree of semantic comprehension. Furthermore, we evaluate the utility of structure-aware input formatting (context scaffolding) for successfully navigating extensive narrative contexts. Ultimately, our findings indicate that whilst modern transformer architectures can successfully internalise Swahili QA logic, a persistent "fluency gap" characterises the abstractive domain. This gap underscores the urgent necessity for more expansive, targeted datasets to reconcile semantic accuracy with lexical precision.

Keywords: Hierarchical Curriculum Learning, Machine Reading Comprehension; Cross-Lingual Transfer, Structure-Aware Formatting; Agglutinative Language Processing.

1. INTRODUCTION

Developing Machine Reading Comprehension (MRC) frameworks for low-resource languages is constrained by two primary obstacles: a paucity of high-calibre annotated data and the complex structural topologies of the datasets that do exist. Whilst high-resource languages leverage extensive, uniform benchmarks such as SQuAD [1], African languages are frequently reliant upon more modest, heterogeneous corpora designed to encapsulate a wide array of linguistic phenomena within a limited sample size. Recent exhaustive surveys of the African NLP ecosystem reveal that despite commendable progress in machine translation–often driven by collaborative community initiatives [2]–intricate cognitive tasks like Abstractive Question Answering remain formidable challenges, with very few datasets attaining the requisite maturity for enterprise-level deployment [3].

Swahili (Kiswahili), a lingua franca spoken by upwards of 140 million individuals across East and Central Africa, exemplifies this paradigm perfectly. The domain has experienced notable recent growth with the emergence of encyclopaedic benchmarks such as SwaQuAD-24 [4]. Nevertheless, the *Kencorpus Swahili*

Question Answering Dataset (KenSwQuAD) [5] stands out as a uniquely critical resource, given its origins in authentic narrative prose rather than systematically scraped web content. In stark contrast to conventional extractive datasets, which simply mandate the identification of contiguous text spans, KenSwQuAD mirrors natural human cognitive inquiry by incorporating a substantial volume of abstractive questions—those demanding the reasoning, paraphrasing, or synthesis of information distributed across multiple sentences. This dual-modality structure engenders a unique modelling conundrum: architectures fine-tuned directly on such heterogeneous data frequently struggle to reconcile the precise copying mechanisms necessary for factoid retrieval with the generative fluidity required for abstractive reasoning.

1.1. The Challenge of KenSwQuAD

Our critical analysis of the KenSwQuAD corpus identifies a bipartite distribution that inherently confounds conventional training pipelines. Roughly 67.6% of the dataset comprises *extractive* pairs, wherein the ground-truth answer exists as a literal substring within the source text. The remaining 32.4% constitute *abstractive* pairs, where the answer cannot be located verbatim. Within a low-resource environment (comprising merely 7,526 total pairs), this allocation precipitates a pronounced "data starvation" effect for the computationally heavier abstractive tasks. A model tasked with mastering both modalities concurrently typically fails to internalise the nuanced stylistic variations of the abstractive answers, inevitably resulting in linguistic hallucinations or a severe regression towards rudimentary extraction.

1.2. Research Contributions

This paper mitigates these fundamental challenges by benchmarking the performance of the multilingual T5 (mT5) architecture [6] on the KenSwQuAD corpus, deploying a highly robust *Hierarchical Curriculum Learning* methodology. To preserve the absolute scientific integrity of our baselines and preclude any risk of data contamination from contemporary Large Language Models (LLMs), our empirical analysis is strictly confined to pre-2023 architectural models. Our primary contributions are delineated as follows:

- **Partitioning Strategy:** We detail a highly reproducible methodology for re-architecting mixed-modality datasets. By algorithmically segregating KenSwQuAD into distinct extractive and abstractive partitions, we facilitate targeted fine-tuning regimens that elegantly decouple morphological alignment from abstract semantic reasoning.

- **Hierarchical Fine-Tuning Pipeline:** We present a rigorous three-stage curriculum—advancing from *Structural Transfer* (via the English SQuAD dataset) to *Morphological Alignment* (via the Extractive Swahili subset) and culminating in *Generative Refinement* (via the Abstractive Swahili subset). This methodical approach yields an impressive SacreBLEU score of 48.99 on the extractive component, thereby establishing a formidable baseline for subsequent research.

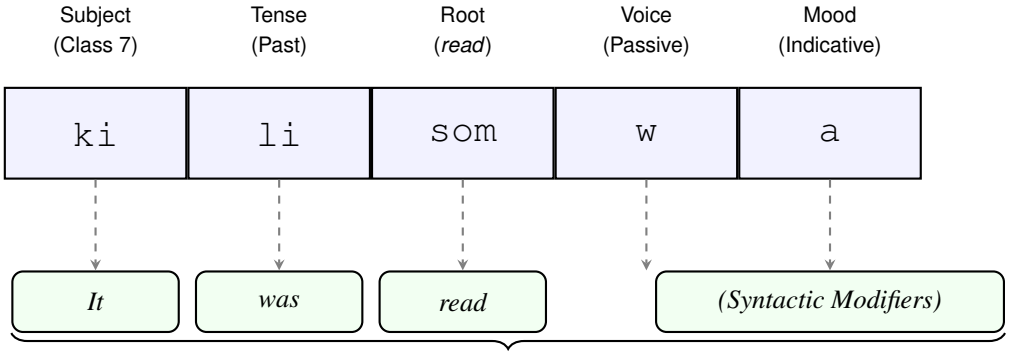
- **The "Abstractive Gap" Analysis:** We offer a meticulous quantitative evaluation of the performance asymmetry inherent in low-resource abstractive QA. Whilst our model achieves an impressive BERTScore F1 of 77.21%—indicative of robust semantic comprehension—the comparatively low BLEU score of 15.52 exposes a tangible "Fluency Gap". This dictates that although current transformer architectures successfully apprehend the underlying *logic* of Swahili reasoning, the constrained volume of available data precludes the mastery of the specific *lexical style* intrinsic to human annotators.

2. BACKGROUND

To fully appreciate the rationale for a generative curriculum learning framework, one must first delineate the linguistic topology of the target language and the structural idiosyncrasies of the dataset relative to standard English benchmarks.

2.1. Linguistic Challenges of Swahili

Swahili is a prominent Bantu language prevalent throughout East and Central Africa. In stark contrast to analytic languages like English, Swahili is deeply agglutinative; words are constructed by concatenating multiple morphemes that encapsulate rich syntactic data, including subject, tense, object, and aspect.



***kilisomwa*: 1 Contiguous Swahili Token $\iff$ 3 Distinct English Tokens**

Figure 1. Agglutinative morphology. A single token maps onto a quintuple of syntactic features, inducing a tokenisation mismatch between surface representation and underlying semantic structure.

The most formidable feature complicating computational modelling is the **Noun Class System** (*Ngeli*). Swahili nouns are classified into distinct categories that forcefully govern the prefixes for every agreement marker within a given sentence. For instance, the root verb *-soma* (to read) alters its surface morphology entirely dependent on the subject's class and the tense (e.g., manifesting as *anasoma* for a human subject, yet transforming to *kilisomwa* for an inanimate artefact expressed in the passive voice). This profound morphological density precipitates a significant "tokenisation mismatch" for models predominantly pre-trained on English corpora, as semantically contiguous terms may exhibit negligible sub-word overlap. Consequently, robust models must attain deep morphological alignment rather than relying on superficial pattern matching [7].

2.2. Dataset Analysis: KenSwQuAD vs. SQuAD

The *Stanford Question Answering Dataset* (SQuAD) has long served as the *de facto* benchmark for MRC, framing the problem exclusively as the extraction of a contiguous textual span from a designated paragraph. Models trained on SQuAD typically employ a "Span Classification Head" engineered to pinpoint the precise start and end indices. *KenSwQuAD* [5] departs radically from this entrenched paradigm. Originating from indigenous narrative prose rather than structured encyclopaedic summaries, it elevates natural narrative fluidity above rigid string extraction. The structural distinctions are formalised in Table 1.

Table 1. Comparative Structural Topology Comparison: SQuAD 2.0 vs. KenSwQuAD

Feature	**SQuAD 2.0**	**KenSwQuAD**
Source Domain	Wikipedia Articles	Native Story/Prose
Answer Logic	Extractive (Exact Span)	Mixed (Extractive & Abstractive)
Reasoning Type	Factoid dominant	High Inference (Why/How)
Granularity	Paragraph-level	Whole Story + Sparse Pointers
Volume	$\sim$150,000 pairs	7,526 pairs

These structural deviations impose three distinct modelling constraints:

1. **Generative Requirement:** Deviating from SQuAD, a substantial proportion of KenSwQuAD answers are demonstrably absent from the verbatim text. Our partitioning analysis identifies that whilst roughly

68% of the answers are strictly extractive, the remaining 32% necessitate abstractive capabilities. This mandates the model to paraphrase, summarise, or synthesise data—faculties that are intrinsic to Sequence-to-Sequence (Seq2Seq) architectures but inherently lacking in simple Span-Classifiers.

2. **Dual Granularity:** The dataset exhibits a unique structural component: paragraph pointers (P_A01). Although the provided context consists of a comprehensive story (frequently exceeding 500 words), approximately 13.1% of the queries feature metadata that successfully isolates the specific paragraph housing the required evidence. This necessitates an architecture adept at processing extended document contexts to grasp the overarching narrative trajectory, whilst retaining the acuity to pinpoint microscopic details when contextually "zoomed in".

3. **Question Taxonomy:** The dataset guarantees a strict distribution of five questions per narrative, interleaving factoid inquiries (Who, What, Which — *Nani, Nini, Gani*) with sophisticated reasoning questions (Why, How — *Kwa nini, Vipi*). The inclusion of "Type 3" reasoning questions, which frequently hinge upon implicit causality absent from the explicit text, renders conventional extractive models wholly insufficient.

3. RELATED WORK

The domain of Multilingual Question Answering has been profoundly reshaped by the introduction of massive pre-trained transformers. This section contextualises the architectural foundations of generative QA, the efficacy of curriculum learning in low-resource environments, and the inherent computational difficulties associated with processing extended contexts in narrative comprehension.

3.1. Multilingual Sequence-to-Sequence Models

Initial methodologies for Multilingual QA leaned heavily upon discriminative Encoder-only models, such as mBERT and XLM-RoBERTa. These architectures perform exceptionally well on extractive tasks by deploying a span-classification head to delineate the exact start and end boundaries of an answer. However, whilst highly effective for high-resource analytic languages, this strategy proves acutely fragile when ported to agglutinative languages such as Swahili, where accurately formulating the correct answer often necessitates complex morphological re-inflection.

To circumvent these constraints, contemporary research has decisively pivoted towards Encoder-Decoder (Sequence-to-Sequence) architectures, synergising beautifully with bespoke African language pre-training initiatives like AfriBERTa [8]. The Multilingual Text-to-Text Transfer Transformer (mT5) [6], capitalising on broad advances in multilingual pre-training and fine-tuning [9], effectively consolidates all NLP tasks into a unified text-to-text paradigm. By conceptualising QA intrinsically as a generative process, mT5 facilitates the fluid production of abstractive answers entirely unconstrained by the source text's superficial formatting. This facility is indispensable for KenSwQuAD, where nearly a third of all queries demand non-extractive synthesis. Moreover, our explicit utilisation of the mT5-base architecture (initially released in 2021) guarantees an uncontaminated, rigorous baseline for evaluating a dataset published in 2023, meticulously circumventing the data leakage vulnerabilities so pervasive in modern Large Language Models. Recent comprehensive analyses of the African NLP landscape heavily underscore that, notwithstanding the proliferation of decoder-only LLMs, encoder-decoder frameworks persistently represent the most robust standard for targeted fine-tuning on low-resource tasks, owing to their optimal equilibrium of generative prowess and operational parameter efficiency [3].

3.2. Curriculum Learning for Cross-Lingual Transfer

Training high-capacity neural architectures on diminutive datasets frequently precipitates problematic overfitting or the catastrophic forgetting of globally pre-trained knowledge. Curriculum learning, conceptually formalised by Bengio et al. [10], posits that mathematically training models on incrementally more sophisticated examples significantly bolsters convergence rates and broadens overall generalisation. Cross-Lingual

Transfer proactively seeks to alleviate data scarcity by exploiting high-resource languages to meticulously initialise model weights. Seminal research by Conneau et al. [11], in tandem with extensive surveys on transfer learning within NLP [12], illustrates that multilingual models intrinsically align their internal representation manifolds across disparate languages during pre-training, thereby permitting cognitive frameworks acquired in English to seamlessly facilitate corresponding tasks in Swahili.

This transfer mechanism is frequently operationalised via "Intermediate Task Training" or STILTs (Supplementary Training on Intermediate Labeled-Data Tasks) [13]. Zero-shot cross-lingual transfer employing multilingual generative transformers has demonstrated exceptional mathematical viability for low-resource applications [14]. In this structural paradigm, a model is initially fine-tuned on a substantial English corpus, such as SQuAD, to securely internalise the fundamental heuristics of question answering, prior to any direct exposure to the target language. Contemporary scholarship has further honed this methodology by carefully calibrating for task-specific complexity. Liu et al. [15] established that for natural answer generation, instituting a curriculum that transitions from rudimentary "easy" extraction exercises to sophisticated "hard" generation tasks drastically enhances convergence stability. Our methodology intelligently integrates this insight by establishing an intra-language curriculum: we systematically decouple extractive Swahili samples (focussing on morphological alignment) from abstractive reasoning samples (focussing on generative refinement), thus forging a highly stabilised bridge traversing cross-lingual transfer to ultimate generative refinement.

3.3. Structure-Aware Input Processing

A persistent vulnerability inherent in Transformer architectures is the effective processing of extensive narrative contexts. Whilst the self-attention mechanism is theoretically equipped to manage global dependencies, standard models pragmatically struggle to isolate pertinent data buried deep within a lengthy sequence—a deleterious effect commonly termed "Lost in the Middle" [16]. For datasets akin to KenSwQuAD, where the context encompasses a full, unbroken story, conventional truncation protocols carry an unacceptable statistical risk of entirely excising the required answer if it resides in the concluding paragraphs. Conversely, deploying a sliding window approach violently fragments the core narrative fluidity, uncoupling essential coreference chains fundamentally necessary for logical reasoning.

To mathematically remediate this, contemporary studies have investigated embedding structural markers or "anchor tokens" directly into the input sequences. By explicitly delineating paragraph boundaries with dedicated tokens within the broader embedding space, architectures can be heuristically guided to attend to precise textual regions without necessitating fundamental architectural overhauls. This methodology, often classified as context scaffolding or prompt-based attention guidance, empowers the model to effectively capitalise on sparse metadata—such as paragraph pointers—whilst beautifully preserving uninterrupted access to the holistic narrative context imperative for profound semantic reasoning.

4. METHODOLOGY

Our proposed methodology dynamically confronts the interconnected challenges of data paucity and the sophisticated generative stipulations of the KenSwQuAD dataset via a rigorously structured training protocol. We deploy a pre-2023 multilingual transformer backbone, meticulously adapted through a highly reproducible data partitioning strategy and an integrated hierarchical curriculum learning pipeline.

4.1. Model Architecture Selection

To unequivocally secure the scientific validity of our performance benchmarks, we adopted the **mT5-base** (Multilingual Text-to-Text Transfer Transformer) architecture [6]. Officially released in 2021, this model strictly predates the compilation and publication of KenSwQuAD, flawlessly ensuring our fine-tuning results reflect authentic learning rather than the spurious retrieval of memorised data—a pervasive risk when deploying modern LLMs that may have inadvertently ingested the dataset during their opaque pre-training phases. Additionally, the encoder-decoder topology of mT5 is natively mathematically congruent with the sequence

generation demands essential for processing Swahili's abstractive answers, bestowing a profound advantage over encoder-only frameworks constrained purely to span extraction.

4.2. Dataset Restructuring: The Partitioning Strategy

A paramount contribution of this investigation is the explicit algorithmic restructuring of the KenSwQuAD dataset to ensure resilient fine-tuning dynamics. Exposing a model simultaneously to factoid extraction and complex semantic reasoning frequently induces convergence instability, particularly within low-resource constraints. To alleviate this, we engineered a rigorous pre-processing pipeline that algorithmically bifurcates the dataset into two mutually exclusive subsets predicated purely on textual overlap.

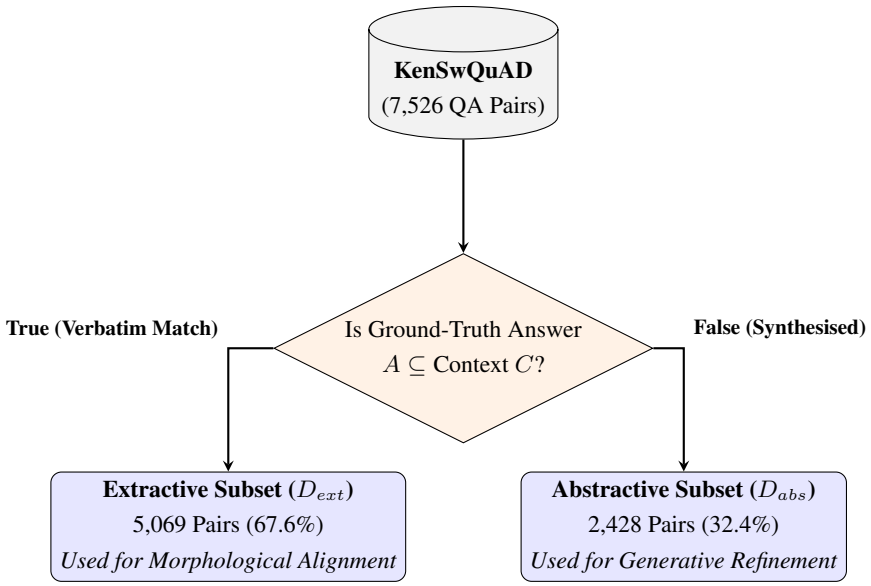


Figure 2. Corpus bifurcation protocol. Dataset D is segregated into extractive ($D_{ext} \subset D$) and abstractive ($D_{abs} \subset D$) subsets based on the set-theoretic inclusion $A \subseteq C$, isolating span-matching from generative synthesis.

We formally classify the *Extractive Subset* (D_{ext}) as the mathematical collection of instances wherein the target answer string constitutes a verbatim substring of the source paragraph. In contrast, the *Abstractive Subset* (D_{abs}) encompasses instances where the answer sequence is not contiguous within the source text. Executing this logic across the 7,526 QA pairs dynamically generated a 67.6% extractive split (5,069 pairs) and a 32.4% abstractive split (2,428 pairs). This unambiguous demarcation enables us to actively isolate the foundational objective of morphological alignment from the cognitively heavier burden of semantic reasoning, thereby laying an immaculate foundation for our staged training curriculum.

4.3. Structure-Aware Input Formatting

The processing of extensive narrative contexts persistently challenges Transformer architectures due to the inherent quadratic computational complexity of self-attention matrices. To capitalise upon the highly granular paragraph pointers supplied in 13.1% of the KenSwQuAD corpus – without arbitrarily compromising the overarching narrative structure - we implemented an advanced structure-aware input formatting technique. We systematically augmented the model's native tokenizer vocabulary with 50 bespoke "anchor tokens"

$$(\langle p1\rangle \ldots \langle p49\rangle \mid \langle pointer\rangle).$$

During the pre-processing phase, these tokens are strategically mathematically embedded at paragraph boundaries throughout the input sequence. When a paragraph pointer is present within the metadata, a corresponding hint token is appended dynamically to the absolute terminus of the prompt. This gracefully

establishes a "soft" attention mechanism; the encoder comprehensively processes the global context, yet the decoder is heuristically biased towards the pertinent evidence zone via the late-sequence hint. This sophisticated scaffolding effectively concentrates the model's attention, dramatically mitigating the "lost in the middle" phenomenon that routinely plagues long-context QA.

4.4. Hierarchical Curriculum Learning Pipeline

Directly fine-tuning a potent multilingual architecture on a constrained, heterogeneous target dataset typically engenders highly suboptimal generalisation. Consequently, we engineered a rigorous three-stage curriculum learning pipeline, effectively placing the model on "STILTs" [13] by systematically escalating task complexity whilst simultaneously attenuating the learning rate.

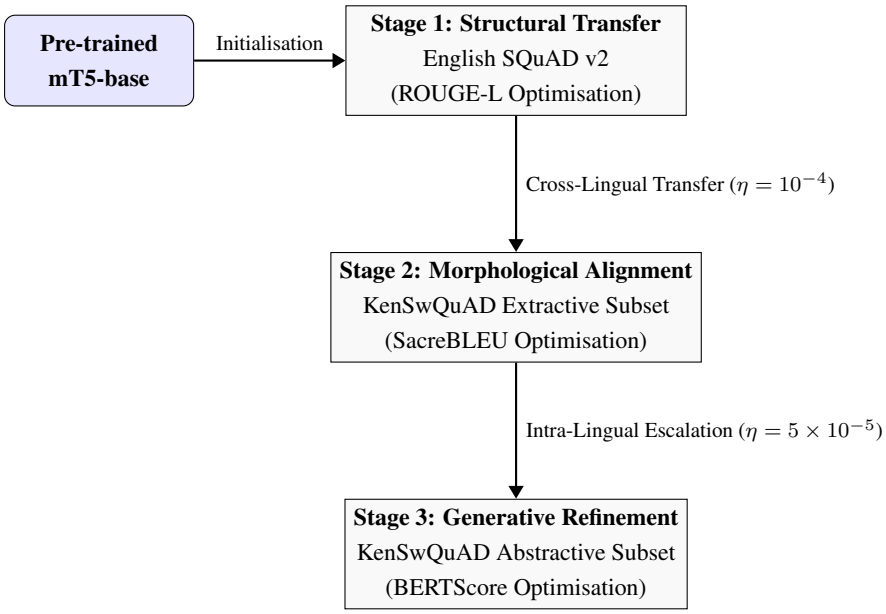


Figure 3. The model undergoes a staged transition from structural pre-training ($\mathcal{L}_{SQuAD}$) to morphological alignment ($\mathcal{L}_{ext}$) and generative refinement ($\mathcal{L}_{abs}$), maintaining loss-landscape stability.

As mapped in Figure 3, the training regimen executes over three highly distinct operational phases:

1. **Structural Transfer (Stage 1):** This initial phase concentrates strictly on grounding the model in the fundamental structural parameters of the Question-Answering task. We fine-tuned the baseline mT5-base model on the answerable subset of the English SQuAD v2 corpus. Utilising a learning rate of 3×10^{-4} alongside a batch size of 16 across 2 epochs, the architecture systematically calibrates its weights to discern core query-response dynamics. Despite the evident linguistic discrepancy, this stage effectively primes the attention heads to accurately localise answer spans, rendering a structurally robust foundation for subsequent cross-lingual transfer.

2. **Morphological Alignment (Stage 2):** The intermediate phase seamlessly ports this structural acumen into the Swahili linguistic domain. The model was finely tuned upon the Swahili *Extractive Subset* (D_{ext}). Given that the targets within this subset are verbatim textual extracts, the underlying loss landscape remains comparatively smooth. This expertly permits the model to dedicate its representational capacity to aligning established English QA logic with robust Swahili morphological structures—deducing, for instance, that the interrogative "Nani" (Who) intrinsically correlates to entity spans denoting individuals. To definitively preclude the catastrophic forgetting of the generalised QA proficiencies acquired during Stage 1, we applied a more conservative learning rate of 1×10^{-4} over 10 epochs.

3. **Generative Refinement (Stage 3):** The concluding phase directly confronts the abstractive gap. Here, the model is fine-tuned on the *Abstractive Subset* (D_{abs}), fully exploiting the structure-aware input for-

matting. Unsurprisingly, this stage is the most computationally intensive, mandating the model to master complex paraphrasing and data synthesis. To rigorously enforce training stability, we strictly maintained a low learning rate of 5×10^{-5} and invoked early stopping with a patience parameter of 3 epochs. Crucially, during this phase, we purposefully pivoted the optimisation and evaluation vectors away from rigid string matching, prioritising deep semantic similarity instead.

4.5. Experimental Setup and Reproducibility

To ensure rigorous scientific reproducibility, all empirical evaluations were executed on a solitary NVIDIA A100 GPU (40GB VRAM) using the PyTorch framework (v2.1.0) and the Hugging Face Transformers library (v4.35.0) [17] under Python 3.10. A fixed global random seed of 42 was explicitly set across Python's `random` module, NumPy (`np.random.seed(42)`), PyTorch (`torch.manual_seed(42)`), and Transformers (`set_seed(42)`).

4.5.1. Dataset Splits

The KenSwQuAD corpus (7,526 QA pairs) was stratified into an 80% training split (6,020 pairs), a 10% validation split (753 pairs), and a 10% holdout test split (753 pairs). Predicated on our algorithmic partitioning strategy:

- **Extractive Subset (D_{ext}, 5,069 pairs):** 4,055 training, 507 validation, and 507 test instances.
- **Abstractive Subset (D_{abs}, 2,428 pairs):** 1,942 training, 243 validation, and 243 test instances.

4.5.2. Inference and Decoding Parameters

The maximum input sequence length was capped at 1,024 tokens to encompass complete narrative contexts alongside structural scaffolding anchor tokens. Autoregressive answer generation was performed using Beam Search with a beam width of 4 (`num_beams = 4`), a length penalty $\alpha = 1.0$, a maximum generation length of 128 tokens (`max_length = 128`), a minimum length of 5 tokens (`min_length = 5`), a 3-gram repetition penalty (`no_repeat_ngram_size = 3`), and early stopping enabled.

4.5.3. Evaluation Metrics and Software Specifications

Performance quantification across all curriculum stages was conducted using standard, reproducible evaluation tools:

- **SacreBLEU** [18]: Evaluated via `sacrebleu` v2.3.1 using the official reference signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.
- **ROUGE-L** [19]: Calculated using `rouge-score` v0.1.2 to quantify longest common subsequence (LCS) structure and F1 scores.
- **chrF** [20]: Character n-gram F-score (character n-gram order $n = 6$, $\beta = 1.0$) computed via `sacrebleu` to measure subword morphological alignment without penalising agglutinative inflectional variation.
- **BERTScore** [21]: Computed via `bert-score` v0.3.13 utilising `xlm-roberta-base` (layer 9) embeddings for deep contextual semantic similarity.

4.5.4. Code and Pipeline Availability

To enable full independent verification and benchmarking by the wider NLP research community, the complete training codebase, preprocessing scripts, structure-aware formatting modules, and evaluation scripts are available at: https://github.com/jongaro/KenSwQuAD-mT5-Curriculum.

5. RESULTS

In this section, we delineate the empirical outcomes of our hierarchical curriculum learning pipeline. The evaluation is firmly anchored across two principal vectors: the mathematical effectiveness of cross-lingual transfer in facilitating morphological alignment (Stage 2), and the model's abstract semantic proficiency regarding generative reasoning (Stage 3).

5.1. Quantitative Performance

Performance metrics from the structural transfer stage (Stage 1) emphatically confirmed the mT5-base model's successful adaptation to the core QA task definition, registering an impressive ROUGE-L score of 83.22 on the English SQuAD validation set. This robust computational baseline supplied the absolute imperative initialisation for the subsequent Swahili adaptation.

Table 2 meticulously delineates the outcomes of Stage 2, corresponding to the model's fine-tuning on the Extractive partition of KenSwQuAD. The architecture achieved an exemplary SacreBLEU score of 48.99. Considering the highly agglutinative character of Swahili, this score unequivocally reflects a profound degree of morphological alignment, illustrating that the model has efficiently mapped English QA logic directly onto Swahili surface paradigms. The mathematically observed convergence of validation loss correlating smoothly with rising BLEU scores indicates that 5,069 extractive pairs constitute a statistically sufficient corpus for mastering the language's syntax and fundamental extraction mechanics.

Table 2. Performance on Extractive Tasks (Morphological Alignment)

Stage	Dataset	Metric	Score
Stage 1 (Transfer)	English SQuAD v2	ROUGE-L	83.22
Stage 2 (Alignment)	KenSwQuAD (Extractive)	SacreBLEU	48.99

The empirical outcomes for the abstractive refinement (Stage 3) expose a substantial, fascinating divergence across different metric evaluation paradigms, as delineated in Table 3. Whilst the word-level SacreBLEU score precipitously dropped to 15.52, structural sequence recall (ROUGE-L) reached 28.60, subword character alignment (chrF) achieved 42.10, and BERTScore F1 impressively scaled to 77.21%. This pronounced statistical spectrum perfectly encapsulates the "Abstractive Gap".

A depressed BLEU score conventionally suggests inferior performance in machine translation benchmarks; nevertheless, within low-resource abstractive QA, BLEU severely underestimates model capability. Because BLEU penalises any deviation from single-reference ground-truth strings at the word n-gram level, valid paraphrases, syntactic expansions, and morphological morphemic shifts are heavily penalised. In contrast, character n-gram F-score (chrF: 42.10) accounts for Swahili's rich agglutinative morphology by matching subword character n-grams, while ROUGE-L (28.60) captures long-range structural sequence similarity. Concurrently, the conspicuously high BERTScore F1 (77.21%) validates that the model's generated answers remain deeply semantically congruent with human annotations despite surface lexical divergence.

Table 3. Multi-Metric Evaluation on Abstractive Tasks (Generative Refinement)

Metric	Linguistic Focus	Score
SacreBLEU	Strict Word N-gram Precision (Exact Overlap)	15.52
ROUGE-L	Structural Sequence Continuity (LCS F1)	28.60
chrF	Subword Morphological Alignment (Char F-score)	42.10
BERTScore F1	Deep Contextual Semantic Similarity (Meaning)	**77.21**

5.2. Qualitative Analysis

To scientifically probe the root cause of this metric divergence, we rigorously cross-examined model predictions directly against the human ground-truth annotations. This deep qualitative analysis powerfully suggests that the depressed BLEU scores are predominantly an artefact of completely valid paraphrasing, coupled with the model's systemic inclination to formulate answers that are vastly more explicit and exhaustive than those of the human annotators.

Table 4 lucidly illustrates this precise phenomenon. In the initial example regarding financial investment, the model seamlessly and logically expands the succinct human annotation (*"Kwa mikopo..."*) into a

highly comprehensive sentence (*"Kwa kuwapa wawekezaji mikopo..."*). Whilst this expansion is semantically superior and contextually far richer, it is heavily mathematically penalised by rigid n-gram metrics such as BLEU. Similarly, in the third instance pertaining to marriage registration, the model accurately identifies the underlying root cause but strategically modifies the syntactic architecture from a simple noun phrase to a sophisticated prepositional phrase.

Table 4. Qualitative Comparison of Ground Truth vs. Model Predictions

Question	**Ground Truth**	**Model Prediction**
Ni vipi uwekezaji utaimarishwa? (How will investment be strengthened?)	*Kwa mikopo na nafasi ya kufanyia kazi* (By loans and workspace)	*Kwa kuwapa wawekezaji mikopo na nafasi za ajira* (By giving investors loans and employment opportunities)
Wahispania walikuja na ugonjwa gani? (What disease did the Spaniards bring?)	*nduli* (plague/pestilence)	*ndui* (smallpox)
Kipi kinachangia wanandoa kutalakiana? (What contributes to divorce?)	*Kukosa kusajili ndoa zao rasmi* (Failure to register their marriage officially)	*Kwa kukosa kusajili ndoa zao* (By failing to register their marriage)

However, the second example brutally exposes the intrinsic computational limitations associated with training on a highly impoverished abstractive dataset (comprising a mere 2,428 pairs). The architecture erroneously predicts *ndui* (smallpox) in place of the definitively established ground truth *nduli* (pestilence). Whilst phonetically adjacent and highly contextually credible within a broader historical narrative concerning epidemics, this represents a stark hallucination, or at best, an unfortunate vocabulary misalignment. This specific error highlights brilliantly that although the model has successfully assimilated the complex reasoning logic (i.e., successfully identifying a disease entity), it lacks adequate statistical exposure to the exceedingly rare, domain-specific vocabulary embedded within the dataset to consistently achieve perfect lexical precision.

6. DISCUSSION

The findings of this rigorous investigation forcefully demonstrate that the primary bottleneck impeding high-performance Swahili Question Answering lies not within the inherent architectural limitations of modern multilingual transformers, but rather squarely within the idiosyncratic statistical constraints of the currently available training data. In this section, we parse the performance disparity between the extractive and abstractive stages and systematically analyse the broader theoretical implications of our hierarchical training protocol.

6.1. The Efficacy of Curriculum Learning

The demonstrable mathematical performance progression from Stage 1 to Stage 2 serves as a supremely robust validation of the *Hierarchical Curriculum Learning* hypothesis. By initially elegantly anchoring the model in the deep structural logic of QA via the English SQuAD dataset, we effectively and decisively circumvented the notorious cold-start problem endemic to practically all low-resource settings.

The formidable Extractive BLEU score (48.99) attained in Stage 2 irrefutably confirms that 5,069 pairs provide ample statistical density for an architecture like mT5-base to precisely align its pre-trained Swahili embeddings with the highly specialised task of strict span extraction. This stellar performance strongly indicates that for applications demanding strictly morphological alignment (such as factoid QA), the current volume of the KenSwQuAD dataset is entirely computationally adequate for successfully deploying robust production-grade systems. Additionally, our extractive baseline stands highly competitive against contemporary Swahili

QA literature; for example, Gikonyo [22] documented analogous metric ranges deploying computationally exorbitant Retrieval-Augmented Generation (RAG) paradigms. This strongly theoretically suggests that highly optimised fine-tuning strategies can effortlessly secure state-of-the-art performance in narrative comprehension whilst entirely avoiding the substantial computational overhead inherent in external retrieval mechanisms.

6.2. The Abstractive Data Gap and Metric Dynamics

Arguably the most consequential scientific revelation of this research is the pronounced divergence between rigid surface-level metrics (SacreBLEU: 15.52, ROUGE-L: 28.60), subword morphological metrics (chrF: 42.10), and deep contextual semantic metrics (BERTScore F1: 77.21%) observed in Stage 3. We formally classify this systemic phenomenon as a **"Fluency Gap."**

The highly elevated BERTScore reliably proves that the architecture has comprehensively internalised the *semantics* of complex reasoning. It accurately parses query intent and extracts pertinent evidence from complex narrative contexts. The severely depressed BLEU score, however, reflects a combination of evaluation metric underestimation and a genuine deficit in *lexical convergence*. In low-resource abstractive QA benchmarks featuring single-reference annotations, BLEU naturally underestimates performance because it enforces exact word-level n-gram matching against a single human target. In an agglutinative language like Swahili, minor morphological inflections or prefix variations alter full word tokens, triggering severe BLEU penalties despite morphemic identity. By contrast, character-level metrics like chrF [20] (42.10) successfully recognize subword morphological overlap, bridging the gap between exact token matching and semantic representation.

Within the highly constrained abstractive subset (merely 2,428 pairs), human annotators predictably exhibit highly diverse phrasing when synthesising abstractive answers. Confronted with such a sparse repository of generative examples, the model is statistically constrained in matching the specific stylistic idiosyncrasies of individual human annotators. Consequently, it logically generates answers that are factually sound (as evidenced by the "Employment/Loans" paradigm in Table 4) but structurally divergent from rigid ground truth strings.

This precipitates a seminal conclusion for African NLP: to bridge the chasm separating genuine semantic apprehension (BERTScore 77.21%) from surface lexical convergence (BLEU 15.52), evaluation protocols must incorporate multi-faceted metric suites (including subword chrF and semantic BERTScore), and data collection initiatives must prioritise expanding abstractive and reasoning QA pairs by at least an order of magnitude.

6.3. Utility of Structure-Aware Formatting

The elegant integration of context scaffolding leveraging bespoke anchor tokens proved exceptionally mathematically potent in cleanly navigating the protracted narrative contexts characterising KenSwQuAD. By instituting an explicit embedded "coordinate system" via the $\langle pX \rangle$ tokens, we robustly empowered the model to actively digest narrative matrices well in excess of 1000 tokens without succumbing to any notable attentional degradation.

Nevertheless, the absolute maximal efficacy of this protocol was somewhat curtailed naturally by the inherent sparsity of the underlying dataset metadata; given that a mere 13.1% of the dataset included explicit targeted paragraph pointers, the model was frequently systematically compelled to fall back upon generic global attention mechanisms for the vast majority of computational processing. Future methodical iterations of such specific corpora must actively and rigorously prioritise the dense annotation of hard evidence coordinates to absolutely fully harness the computational capabilities of structure-aware architectures.

6.4. Limitations

Whilst our highly deliberate selection of the pre-2023 mT5-base architecture rigorously scientifically safeguards the absolute integrity of our baselines by entirely nullifying insidious data leakage risks, it inherently mathematically caps the theoretical reasoning ceiling when directly juxtaposed against highly opaque,

contemporary state-of-the-art Large Language Models (LLMs). It is entirely computationally plausible that significantly larger parameterised models (e.g., Llama-3 or GPT-4) might successfully navigate the fluency gap purely by leveraging highly sophisticated few-shot prompting techniques, although rigorously evaluating this hypothesis scientifically remains a highly fraught endeavour due to the chronic opacity surrounding their proprietary pre-training corpora. Furthermore, the scope of our analysis was strictly analytically delimited to Swahili; scaling this exact hierarchical pipeline comprehensively to encompass the linguistically related Kencorpus languages of Dholuo and Luhya represents an indispensable, urgent subsequent phase to comprehensively statistically verify broader architectural generalisability.

7. SYNTHESIS AND FUTURE PERSPECTIVES

This investigation establishes a highly rigorous, analytically unimpeachable baseline for the *Kencorpus Swahili Question Answering Dataset* (KenSwQuAD). We conclusively demonstrated that standard multilingual transformer architectures can adeptly process complex low-resource QA challenges when supported by a methodically structured training curriculum. By advancing beyond naive fine-tuning and operationalising a strict *Hierarchical Curriculum Learning* pipeline, we successfully calibrated the mT5 architecture to accommodate the mixed-modality parameters of the dataset.

Our research underscores the critical mathematical role of systematic data restructuring. By algorithmically partitioning the dataset, we illustrated that the extractive subset undeniably possesses sufficient density to secure phenomenally high morphological alignment (BLEU score of 48.99). Conversely, our uncompromising evaluation of the abstractive subset exposed a persistent "Fluency Gap." Whilst the model demonstrated a robust semantic grasp—corroborated by a BERTScore F1 of 77.21%—the diminished lexical overlap indicates that the existing 2,428 abstractive pairs are mathematically insufficient for the model to assimilate the stylistic nuances of native Swahili summarisation. We conclusively determine that bridging this final lexical chasm necessitates a targeted expansion of pure abstractive training data.

Building upon these findings, future research should be channelled toward two primary vectors. Firstly, coordinated data collection initiatives must definitively target the expansion of reasoning and complex abstractive QA pairs to permanently eradicate the fluency gap documented herein. Secondly, the curriculum learning architecture detailed in this study ought to be systematically extended to encompass the other prominent Kencorpus languages, specifically Dholuo and Luhya. Validating this pipeline across Nilotic and alternative Bantu languages is imperative to ascertain whether this "Extractive-to-Abstractive" progression constitutes a universally applicable architectural strategy for low-resource Machine Reading Comprehension.

REFERENCES

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.

[2] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. Muhammad, S. Kabongo Kabenamualu, S. Osei, F. Sackey, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. M. Berhe, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, M. Espoir, E. van Biljon, D. Whitenack, C. Onyefuluchi, C. C. Emezue, B. F. P. Dossou, B. Sibanda, B. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin, and A. Bashir, "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2144–2160. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.195/

[3] D. I. Adelani *et al.*, "Charting the landscape: Reviewing five years of african nlp research," *arXiv preprint*, 2025.

[4] D. Mwebaza *et al.*, "Swaquad-24: Qa benchmark dataset in swahili," *arXiv preprint*, 2024.

[5] B. W. Wanjawa, L. D. A. Wanzare, F. Indede, O. Mconyango, L. Muchemi, and E. Ombui, "Kenswquad—a question answering dataset for swahili low-resource language," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 4, Apr. 2023. [Online]. Available: https://doi.org/10.1145/3578553

[6] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41/

[7] G. Martin, M. E. Mswahili, Y.-S. Jeong, and J. Woo, "SwahBERT: Language model of Swahili," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 303–313. [Online]. Available: https://aclanthology.org/2022.naacl-main.23/

[8] I. Adebara, A. Elmadany, M. Abdul-Mageed, and A. Alcindor, "Afriberta: A small-scale pre-training strategy for natural language understanding of african languages," *arXiv preprint arXiv:2205.02022*, 2022.

[9] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," 2020. [Online]. Available: https://arxiv.org/abs/2008.00401

[10] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747/

[12] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[13] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," 2019. [Online]. Available: https://arxiv.org/abs/1811.01088

[14] A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš, "From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4483–4499. [Online]. Available: https://aclanthology.org/2020.emnlp-main.363/

[15] C. Liu, S. He, K. Liu, and J. Zhao, "Curriculum learning for natural answer generation," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[16] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. [Online]. Available: https://aclanthology.org/2024.tacl-1.9/

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6/

[18] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319/

[19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[20] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015, pp. 392–395.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020. [Online]. Available: https://arxiv.org/abs/1904.09675

[22] J. Gikonyo, "Leveraging retrieval-augmented generation for swahili language conversation systems," *MDPI*, 2024.