scientific reports

Check for updates

OPEN

Deep neural architectures for Kashmiri-English machine translation

Syed Matla Ul Qumar[1]✉, Muzaffar Azim[1], S. M. K. Quadri[2], Mohannad Alkanan[3], Mohammad Shuaib Mir[3] & Yonis Gulzar[3]✉

This paper presents the first comprehensive deep learning-based Neural Machine Translation (NMT) framework for the Kashmiri-English language pair. We introduce a high-quality parallel corpus of 270,000 sentence pairs and evaluate three NMT architectures: a basic encoder-decoder model, an attention-enhanced model, and a Transformer-based model. All models are trained from scratch using byte-pair encoded vocabularies and evaluated using BLEU, GLEU, ROUGE, and ChrF++ metrics. The Transformer architecture outperforms RNN-based baselines, achieving a BLEU-4 score of 0.2965 and demonstrating superior handling of long-range dependencies and Kashmiri's morphological complexity. We further provide a structured linguistic error analysis and validate the significance of performance differences through bootstrap resampling. This work establishes the first NMT benchmark for Kashmiri-English translation and contributes a reusable dataset, baseline models, and evaluation methodology for future research in low-resource neural translation.

Neural Machine Translation (NMT) is a deep learning-based approach that enables automatic text translation from one language to another, significantly improving fluency and contextual accuracy over traditional statistical methods[1,2]. Unlike Statistical Machine Translation (SMT), which depends on phrase-based probability models, NMT systems learn sequence-level mappings directly from data, allowing them to model long-distance dependencies and richer syntax[3].

While NMT has achieved remarkable success for high-resource languages such as English, French, or Chinese[4], its application to low-resource and morphologically rich languages like Kashmiri remains largely unaddressed. Kashmiri poses unique computational challenges due to its complex morphology, Perso-Arabic script, and regional dialectal variation[5,6]. The language lacks publicly available corpora, robust tokenizers, or neural translation systems, severely limiting research and development.

While multilingual models such as IndicTrans2[3] have recently begun to support Kashmiri, the language remains significantly underrepresented in large-scale pretrained translation systems. Notably, mBART and mT5 do not include Kashmiri among their supported languages. Even in IndicTrans2, Kashmiri is not fine-tuned for the English target direction, and support remains generic rather than task-specific. As a result, these models offer only limited translation quality for Kashmiri-English tasks. To date, no prior study has built a dedicated NMT pipeline or conducted a systematic evaluation for this language pair.

To address this gap, we present the first end-to-end Kashmiri-English NMT system, built upon a high-quality parallel corpus of 270,000 sentence pairs. We evaluate three neural architectures — a basic encoder-decoder model, an attention-enhanced model, and a Transformer-based model[7] — and assess their performance using multiple metrics. In addition, we introduce a structured linguistic error taxonomy and perform statistical significance testing via bootstrap resampling to validate performance differences across models.

This study establishes a foundational benchmark for Kashmiri-English neural translation and contributes a scalable methodology for adapting deep learning to other under-resourced and morphologically complex languages.

Contributions

In summary, our contributions are as follows:

[1]FTK-Centre for Information Technology, Jamia Millia Islamia, New Delhi 110025, India. [2]Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India. [3]Department of Management Information systems, College of Business Administration, King Faisal University, 31982 Al-Ahsa, Saudi Arabia. ✉email: syed1909832@st.jmi.ac.in; ygulzar@kfu.edu.sa

1. We present the first deep learning-based NMT pipeline dedicated to the Kashmiri-English language pair, introducing a methodological foundation for neural translation in a previously unexplored low-resource setting.
2. We develop the first high-quantity and high-quality Kashmiri-English parallel corpus, consisting of 270 K sentence pairs, which we make publicly available on the Hugging Face platform. This dataset significantly enhances the resources available for low-resource language translation research.
3. We propose three distinct NMT architectures tailored for Kashmiri-English translation:

    i. a basic encoder-decoder model,
    ii. an encoder-decoder model with an attention mechanism, and.
    iii. a Transformer-based model leveraging multi-head self-attention.

These models are systematically evaluated to assess their effectiveness for morphologically rich and low-resource language pairs.

4. We conduct a detailed empirical evaluation using BLEU, GLEU, ROUGE, and ChrF++ metrics. The Transformer model achieves a BLEU-4 score of 0.2965, outperforming both RNN-based baselines, and demonstrating its ability to model long-range dependencies and complex morphology.
5. We provide a structured qualitative analysis, including a linguistic error taxonomy covering tense, morphology, fidelity, and structural coherence. This analysis, supported by example-based comparisons, highlights the specific challenges of translating Kashmiri and the comparative strengths of each model.
6. We perform statistical significance testing using paired bootstrap resampling to validate the observed improvements, showing that the Transformer model's performance gains are statistically reliable.

In addition to these contributions, our study introduces the first end-to-end neural machine translation pipeline tailored for Kashmiri-English, combining corpus creation, modern model benchmarking, and structured linguistic evaluation. This work represents a methodologically novel foundation for future NMT research on Kashmiri, and more broadly demonstrates how to adapt deep learning pipelines to structurally complex, under-resourced languages.

Related work

Machine Translation (MT) is a complex challenge within the field of Natural Language Processing (NLP), where researchers strive to develop automated methods for translating between human languages. This research area dates back to 1949 when Warren Weaver introduced it in his "Memorandum of Translation." Achieving human-level translation accuracy remains an elusive goal, as no existing machine has yet matched human performance in translating between languages[8]. One of the primary challenges lies in establishing correlations between languages that often lack standardization, whether in linguistic morphology or topology. This requires much more than simple word or phrase substitution.

Traditionally, machine translation approaches have been rule-based, statistical-based, example-based, or a combination of these techniques[9]. However, these methods have faced significant drawbacks, primarily due to their dependence on manual feature engineering and explicit annotations. As a result, they lack scalability and generalizability. The advent of neural network-based translation systems, known as Neural Machine Translation (NMT), has brought about a more flexible and scalable solution, outperforming traditional methods. Unlike earlier approaches, NMT does not rely on manual, automated, or semi-automated annotations, making it more efficient and adaptable[1].

Deep learning has become the most popular technique for addressing various research challenges, including computer vision, NLP, automatic speech recognition, and bioinformatics[10,11]. Specifically, in NLP, deep learning has proven superior to conventional methods in tasks such as language modeling[12], sentiment analysis[13,14], conversational modeling[15], machine translation[9], handwriting generation[16], and text-to-speech conversion[17,18].

One of the groundbreaking innovations in machine translation using deep learning is the encoder-decoder framework introduced in 2013[19], which was later improved with the sequence-to-sequence model in 2014[20]. This model utilized Long Short-Term Memory (LSTM) networks for both the encoder and decoder, effectively addressing issues like long-distance reordering and the vanishing/exploding gradient problem[4]. However, it had a limitation in that it relied on a single context vector to store all encoder information. Subsequent enhancements incorporated attention mechanisms, allowing the decoder to focus dynamically on relevant input parts, thereby significantly improving translation quality[1,21].

Attention models have since gained prominence, as they have demonstrated notable improvements in NMT performance[2]. Moreover, word embeddings, such as Word2Vec[22], GloVe[23], and FastText[24], have been used as input to encoders to provide richer word representations compared to simple one-hot encoding. FastText stands out for its ability to generate embeddings for rare and out-of-vocabulary words[25], despite being slower than other models. Additionally, techniques have been extended to learn distributed representations for paragraphs[26], sentences[27], and topics[28]. Several advanced techniques have been employed to further enhance neural networks. Residual connections have been introduced to LSTM-based encoder-decoder models to maintain performance when stacking deeper layers[9,29–31]. Regularization methods, such as dropout[30,32], have been effective in preventing overfitting and have been widely used in various applications, including machine translation[5], question answering[33], image classification[34], and speech recognition[35].

Kashmiri is an extremely low-resource and morphologically rich language, posing unique challenges for computational processing. Despite its linguistic complexity, the application of deep learning techniques to

address research problems related to the Kashmiri language remains virtually unexplored. Currently, there are no established linguistic resources available, such as automatic speech recognition systems, sentence boundary disambiguation tools, sentiment analysis models, or transfer learning applications tailored for Kashmiri[36]. Significantly, no research has been conducted on Neural Machine Translation (NMT) for Kashmiri, primarily due to the absence of parallel corpora and annotated datasets that are essential for training state-of-the-art translation models[36]. Furthermore, the lack of modern techniques such as encoder-decoder architectures[20], attention mechanisms[2], and beam search[37] within Kashmiri language processing underscores a considerable gap in existing research and innovation.

Deep learning techniques inherently rely on substantial amounts of data to effectively model linguistic diversity and capture the inherent complexities of both source and target languages. Consequently, the availability of large-scale datasets is fundamental to the successful development of robust translation systems. The presence of a parallel corpus not only enhances model training but also fosters continued research within the domain. Historically, most efforts have been concentrated on resource-rich languages such as English, Spanish, French, German, Russian, and Chinese. These endeavors have been significantly supported by projects like the Europarl Corpus[38], United Nations Parallel Corpus[39], and TED Talks Corpus[40], which encompass diverse linguistic domains and have proven invaluable for both linguistic research and the advancement of machine translation. These corpora are typically constructed through manual translation processes or automated methods, including web crawling techniques exemplified by the News Crawl Corpus[41] and ParaCrawl Corpus[42].

Despite these advancements, extending parallel corpora to low-resource languages continues to present significant challenges. Various initiatives, including the use of Amazon Mechanical Turk[43] for Indian languages and the implementation of the Gale & Church algorithm for English-Myanmar alignment[44], demonstrate the research community's commitment to promoting linguistic inclusivity. However, the persistent scarcity of parallel data for languages spoken by smaller communities or those with limited digital representation poses considerable obstacles to the development of high-quality machine translation systems[45–47]. This scarcity perpetuates the digital divide and hinders the integration of underrepresented languages into modern technological frameworks[48].

Notably, the Kashmiri language remains devoid of a dedicated parallel corpus, thereby revealing a significant research gap in the field of language processing and machine translation. To the best of our knowledge, neural network-based translation techniques have yet to be applied to Kashmiri-English translation, with only limited preliminary work reported in the literature[49]. Furthermore, the absence of a publicly available Kashmiri-English parallel corpus considerably restricts progress in the development of contemporary translation systems. In response to these challenges, the present study seeks to address this gap by developing machine translation models and constructing a comprehensive parallel corpus specifically designed for Kashmiri-English translation.

Dataset

This study introduces the first large-scale, publicly available Kashmiri-English parallel corpus, constructed to support neural machine translation (NMT) and low-resource language research. The corpus contains 269,288 sentence pairs, built using three complementary pipelines: (1) digitized bilingual literary texts, (2) manually authored and translated conversational dialogues, and (3) a filtered legacy corpus subjected to semi-automatic refinement. Special attention was given to orthographic normalization, dialectal variation, and script consistency, ensuring linguistic integrity across sources. Approximately 83% of the data is human-translated and reviewed, making the corpus one of the most linguistically reliable resources for Kashmiri to date.

Translated literary texts (Source 1)

The first component of the corpus comprises translated literary texts obtained from the Jammu & Kashmir Academy of Art, Culture and Languages (JKAACL), local publishers, and academic repositories. Books were selected based on cultural relevance and linguistic richness, with preference given to prose and narrative works. Notable examples include *Folktales of Kashmir*, *The Story of My Experiments with Truth*, and translations of works by Tolstoy, Chekhov, and Shakespeare.

Digitization was performed using the TVS Electronics PDS 8 M scanner, followed by deskewing and noise removal. Optical character recognition (OCR) was conducted using Google OCR, which outperformed alternatives like ABBYY FineReader and gImageReader in recognizing Perso-Arabic script. Nevertheless, OCR outputs required substantial post-processing to restore missing diacritics, correct misrecognized characters, and reintegrate split or merged tokens. Problematic scans (e.g., decorative calligraphy or poetic formats) were excluded due to low OCR accuracy. The cleaned text was segmented by chapter and aligned at the sentence level through a semi-automatic process involving programmatic extraction and manual correction using Microsoft Word. This component contributed approximately 48,000 aligned pairs.

Manually authored conversations (Source 2)

To capture spoken Kashmiri and ensure domain diversity, we created a manually authored dataset of 70 dialogues, each averaging 25–30 speaking turns. Prompts were developed by native speakers and covered a wide range of culturally grounded scenarios, including: "Talking about a new house," "Discussing historical places of Kashmir," "Shopping for dry fruits," and "Consulting a doctor." These prompts were expanded into full conversations by two annotators acting as speakers. The conversations were subsequently translated into English by bilingual experts and reviewed collaboratively.

The resulting dialogues span daily life, cultural practices, education, healthcare, administrative domains, and tourism. This data is rich in questions, imperatives, honorifics, and morphosyntactic variation—capturing linguistic constructions typical of real-world communication. All translations were performed with full-context visibility, enabling direct alignment without external tools. This subset contributes over 200,000 aligned pairs.

Source type	Data origin	# Sentences	% of Corpus	Translation method	Domain coverage
Manually authored data	Curated dialogues by native speakers	~205,000	76%	Human-written + expert review	Conversational, cultural, practical
OCR-digitized books	Bilingual literary texts	~32,000	12%	Existing books (aligned)	Literary, historical, folklore
Translated English sources	English datasets → translated to Kashmiri	~19,000	7%	Professional translation	Everyday language, general topics
Filtered public corpus	Cleaned from noisy legacy corpora	~14,000	5%	Semi-automatic filtering	Generic, administrative, educational

Table 1. Overview of the parallel corpus by source type and domain.

Statistic	Kashmiri	English
Total number of sentences	269,288	269,288
Total number of words	2,963,612	2,800,713
Total number of unique words	116,233	56,822
Average number of words per sentence	11.00	10.40

Table 2. Corpus statistics.

To expand syntactic coverage, we also translated a selected subset of English conversational sentences from the ManyThings.org dataset into Kashmiri. These additions target general-purpose expressions and reinforce alignment quality.

Filtered legacy corpus (Source 3)

An existing Kashmiri-English dataset of ~120,000 sentence pairs (e.g., BPCC, IndicTrans2) was filtered using a custom Redundancy-Based Parallel Corpus Refinement pipeline. Key issues included high repetition, language substitution (e.g., Urdu instead of Kashmiri), and non-parallel or nonsensical alignments.

We first removed approximately 11,000 repeated alignments, then identified highly duplicated Kashmiri sentences (e.g., "وی پ اس ٹ ـپ ی ٹھ د س ت ای پ" repeated 3,000 + times). We filtered out such cases, retaining only alignments with unique and verified content. Manual review further eliminated noise, gibberish, or mismatches. The final cleaned subset contributes ~14,000 high-quality pairs.

Corpus composition and domain coverage

To ensure the corpus supports general-purpose and domain-specific translation tasks, materials span literary, conversational, administrative, and cultural registers. Table 1 provides an overview of corpus composition by source type, domain, and translation methodology.

This distribution supports both domain-general and domain-specific NMT applications. While regional dialects were considered in manual and OCR data, some bias toward standard and urban Kashmiri may remain due to the availability of published and edited sources.

Dataset statistics

The finalized corpus contains 269,288 aligned sentence pairs, with over 2.9 million Kashmiri words and 2.8 million English words. Summary statistics are presented in Table 2.

The significantly higher number of unique words in Kashmiri reflects its morphological complexity, free word order, and dialectal variation. Kashmiri sentences are often longer and structurally more flexible than English ones, requiring translation models to handle long-distance dependencies and complex verb morphology.

Dataset availability

The dataset is publicly available on Hugging Face, ensuring accessibility for researchers working on Kashmiri language processing and low-resource machine translation tasks. The dataset can be accessed at [DOI: https://doi.org/10.57967/hf/3660] and with the URL: https://huggingface.co/datasets/SMUQamar/Kashmiri-English-Dataset-270 K/tree/main. Usage guidelines and licensing information are also provided to facilitate responsible and ethical utilization of the corpus in future research.

Building NMT models for Kashmiri-English translation

Overview

To investigate neural approaches to Kashmiri-English translation, we implement and compare three distinct neural machine translation (NMT) models of increasing architectural complexity: (i) a baseline sequence-to-sequence encoder-decoder model using LSTM layers, (ii) an attention-enhanced encoder-decoder model, and (iii) a Transformer-based model. This section details the architecture, training configurations, and implementation strategies for each system.

These models are designed to explore how different architectures handle the linguistic challenges of Kashmiri, including its rich morphology, flexible word order, and orthographic variation. All models were trained using the same preprocessed dataset (as described in Sect. 3) and evaluated on identical splits for comparability.

Data preparation

Preprocessing and cleaning

To ensure high-quality input for the NMT models, several preprocessing steps were applied to the raw Kashmiri-English sentence pairs. These aimed to eliminate noise, maintain linguistic consistency, and prepare the data for effective tokenization.

a. Data cleaning

- All Kashmiri and English texts were cleaned to remove non-printable and unknown characters, extra spaces, and malformed data.
- For Kashmiri, we retained only characters within the Unicode range U + 0600–U + 06FF, encompassing the Perso-Arabic script, diacritics, and symbols.
- English data was restricted to standard ASCII characters (A–Z, a–z, 0–9).
- Extraneous spaces between words and sentence boundaries were removed to standardize formatting.

b. Normalization

- Kashmiri text was normalized using Normalization Form Canonical Composition (NFC) to handle character encoding variations and diacritics.
- English sentences were lowercased to ensure consistent tokenization and reduce vocabulary sparsity.
- All text was encoded in UTF-8 for cross-platform compatibility.

c. Data splitting

- The dataset was randomly shuffled and split into training (90%), validation (5%), and test (5%) subsets.
- This ensured representative coverage across linguistic styles and domains for all evaluation stages.

Tokenization and vocabulary construction

Following normalization, we applied Byte Pair Encoding (BPE) using the SentencePiece library to tokenize sentences into subword units. This method was chosen due to the morphological richness of Kashmiri and the structural asymmetry between Kashmiri and English.

- *Handling Morphological Complexity*: BPE reduces vocabulary size by learning frequently co-occurring character sequences, capturing common prefixes, suffixes, and inflections.
- *Script Separation*: Kashmiri (Perso-Arabic script) and English (Latin script) were tokenized separately to prevent cross-linguistic token merging.

Separate vocabularies were constructed:

- 15,000 subword units for Kashmiri
- 10,000 subword units for English

This allowed each model to learn language-specific subword representations independently, Fig. 1.

Padding and sequence handling

To facilitate batch processing, tokenized sequences were padded or truncated to a maximum sequence length of 230 tokens. Sentences shorter than this length were padded with zeros, while longer sentences were truncated, Table 3. Each sentence was also wrapped with special tokens:

- [SOS]: Start-of-sequence.
- [EOS]: End-of-sequence.

This structure ensures that models can correctly interpret sentence boundaries.

Baseline encoder-decoder model

Model architecture

The baseline model adopts a standard sequence-to-sequence framework using stacked Long Short-Term Memory (LSTM) layers. The architecture consists of two LSTM layers each for the encoder and decoder[4].

Encoder: The encoder processes the input Kashmiri sequence $x = (x_1, x_2, x_3 \ldots \ldots, x_T)$, where T is the sequence length. Each token in the sequence is transformed into an embedded vector.

$$e_t = Embedding\,(x_t)\,,\ \ t = 1,2,\ldots,T \tag{1}$$

The embedded vectors are passed through two LSTM layers, each consisting of 512 units. The LSTMs process the sequence step by step, generating hidden states h_t and cell states c_t:

$$h_t, c_t = LSTM\left(e_t,\ h_{t-1}, c_{t-1}\right),\ \ t = 1,2,\ldots,T \tag{2}$$

The final hidden and cell states from the second LSTM layer $h_T^{(2)}$, $c_T^{(2)}$ are passed to the decoder as the initial states for generating the target translation.

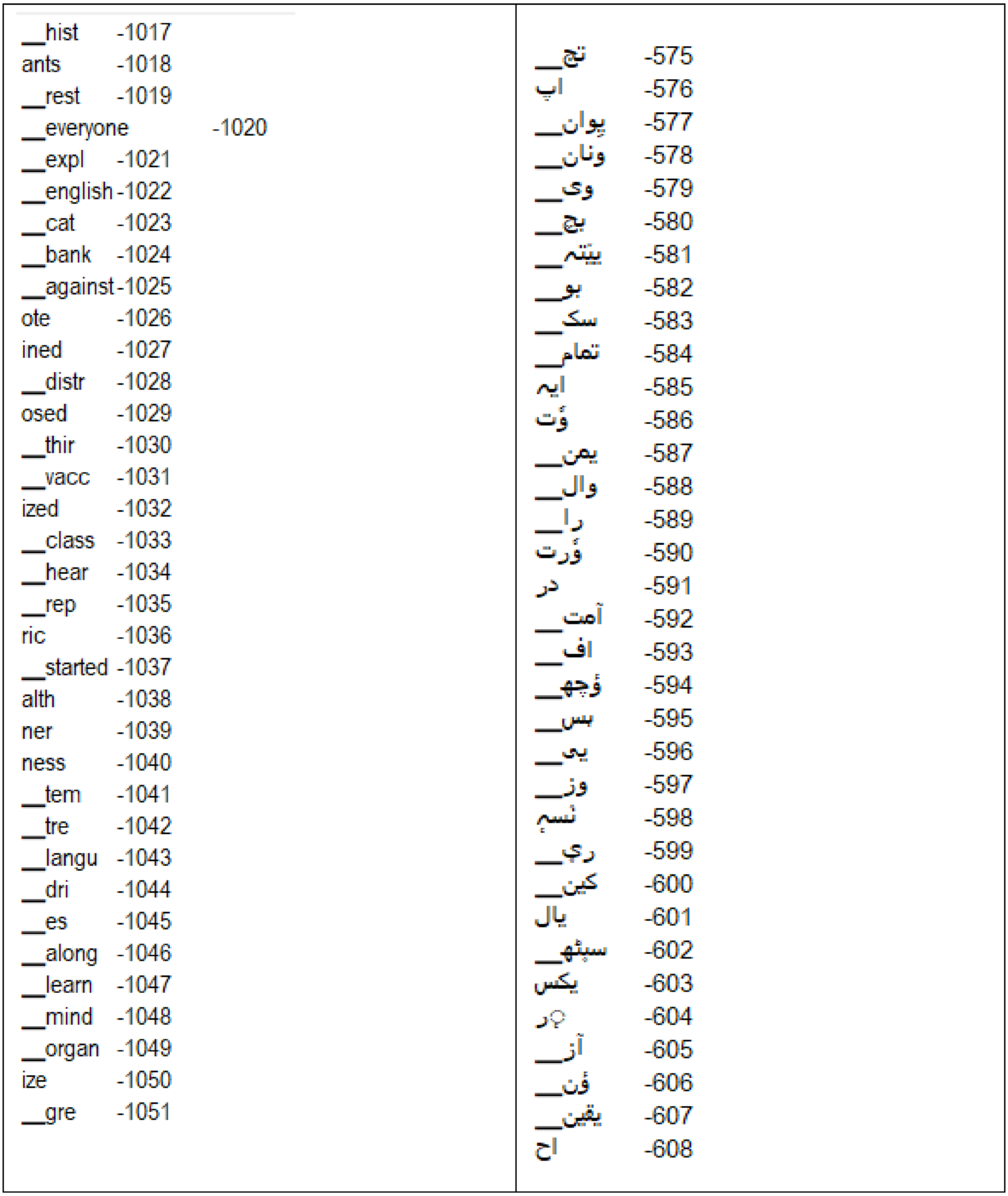

Fig. 1. Snapshort of a vocabulary bullt after using BPE tokenization seperately on kashmir and English sentences.

Language	Sentence (Tokenized and Padded)	Integer IDs (with Padding)
Kashmiri	ورُک شش‌ُوک چنرک فاعمہت چنزور لدمحر ?? ??----------?? ??	[1 6558 4238 28 3250 762 955 1303 0 0 0 ------- 0 0 0 2]
English	try to be generous and forgiving??-----------?? ??	[1 1076 13 55 7879 43 4403 0 0 0 0 ------- 0 0 0 0]

Table 3. Tokenized and padded Kashmiri and english sentences with corresponding integer representations.

Decoder: The decoder also consists of two LSTM layers with 512 units. It generates the target English sequence $y = (y_1, y_2, y_3 \ldots \ldots, y_{T\prime})$ where $T\prime$ is the length of the target sequence, token by token. At each time step, the decoder predicts the next word based on its hidden state and the previously generated word.

$$h\prime_t, c\prime_t = LSTM\left(e_t, h\prime_{t-1}, c\prime_{t-1}\right), \quad t = 1,2,\ldots, T\prime \tag{3}$$

The output at each time step is passed through a dense layer with a softmax activation to produce a probability distribution over the target vocabulary:

$$\widehat{y}_t = softmax\left(W_h h'_t + b_h\right), \qquad t = 1,2,\ldots, T\prime \tag{4}$$

Here, W_h and b_h represent the trainable weights and biases in the dense layer.

This architecture enables the model to learn complex sequence mappings from Kashmiri to English. Figure 2 illustrates the flow of data through our encoder-decoder architecture.

Training the model

Training was conducted on an NVIDIA A100 GPU using Google Colab. The model was trained over a maximum of 50 epochs, with early stopping applied after three consecutive epochs without improvement in validation loss. Each epoch took approximately 10 min. Model checkpoints were saved after every epoch.

Hyperparameters and settings:

- Batch size: 64.
- Embedding dimension: 300.
- LSTM units: 512 for both encoder and decoder.
- Learning rate: Initial rate of 1e-4, exponentially decayed after epoch 10.
- Optimizer: Adam, with gradient clipping (clipnorm = 1.0).
- Loss function: Sparse categorical cross-entropy with label smoothing (smoothing factor = 0.1).

Optimizer and Loss Function:

The Adam optimizer was employed with the following parameters:

- $\beta 1 = 0.9$.
- $\beta 2 = 0.98$.
- $\epsilon = 1e-9$.

Gradient clipping with a threshold of 1.0 was used to prevent exploding gradients, ensuring smooth convergence during backpropagation. The loss function incorporated label smoothing, which reduced overconfidence in the model's predictions by redistributing a small fraction of the probability mass across incorrect classes:

$$Loss = -\sum_{t=1}^{T\prime} \sum_{i=1}^{V} y_{t,i} log(\widehat{y}_t, i), \tag{5}$$

where $y_{t,i}$ represents the smoothed target probability for word i at time step t, and $\widehat{y}_t, i$ is the predicted probability.

Training Strategy:

- *Learning rate decay*: Begins after 10 epochs to allow fine-tuning.

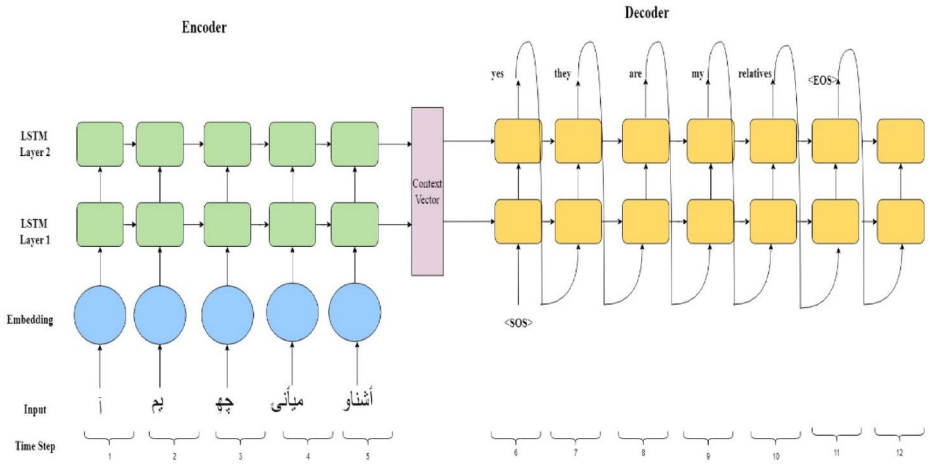

Fig. 2. Our Encoder-Decoder based NMT architecture.

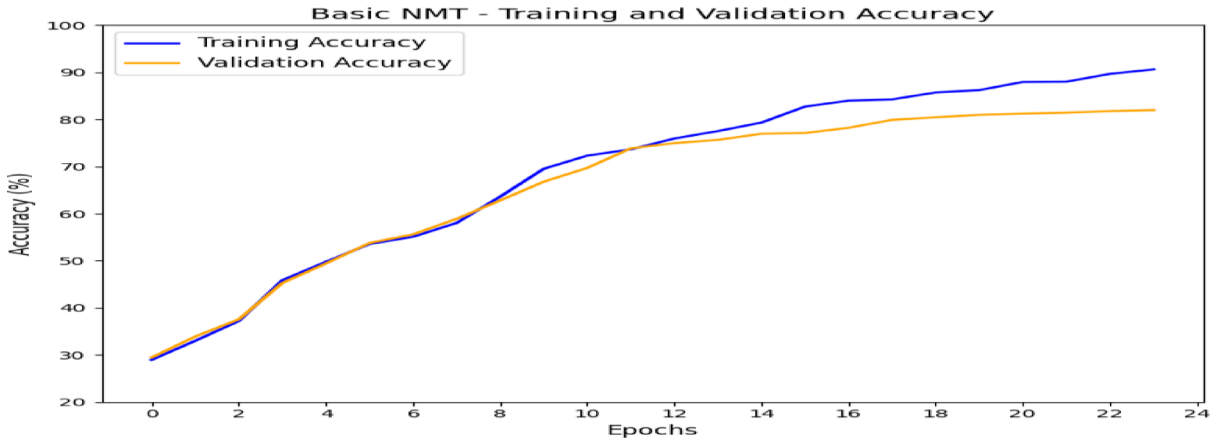

Fig. 3. Basic NMT training and validation accuracy over time.

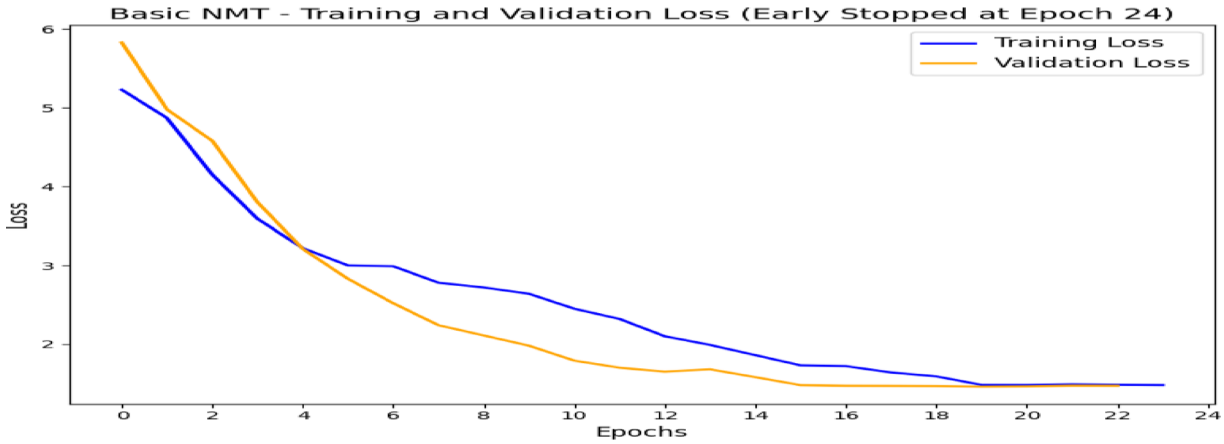

Fig. 4. Basic NMT training and validation loss over time.

- *Early stopping*: Based on validation loss stagnation.
- *Checkpointing*: Enables recovery and analysis of best-performing epochs.

. As seen in Figs. 3 and 4, the training and validation accuracy and training and validation loss respectively, curves indicate stable convergence. The model demonstrated a steady improvement in performance during the initial epochs, reaching an acceptable level of accuracy for shorter sentences. The graphs show a steady improvement in both training and validation accuracy during the initial epochs, with training accuracy eventually surpassing 90%. However, validation accuracy plateaus around 81%. The training and validation loss curves show a similar trend, with both decreasing steadily until around epoch 16, after which validation loss stabilizes. This suggests the model has reached its optimal performance, and further training would not yield significant improvements. Early stopping at epoch 24 was justified as the model began to overfit, performing better on the training data than on unseen validation data. This highlights the need for more sophisticated techniques like attention mechanisms to improve performance, especially for longer sentences.

Encoder-decoder model with attention

Attention mechanisim

The attention mechanism plays a critical role in modern neural machine translation (NMT) systems, particularly when translating long and syntactically complex sentences. Unlike traditional encoder-decoder models that compress the source sentence into a single fixed-length vector, attention dynamically computes a context vector at each decoding step, enabling the model to focus selectively on relevant parts of the input sequence[1,5].

Given the decoder hidden state at time step t, denoted as h_t, and the encoder hidden states $h_1, h_2, h_3 \ldots .h_T$, where T is the source sequence length:

- *Score calculation*.

$$e_{t,k} = Score(\, h_t,\, s_k) = v^{\top} tanh\,(W[\, h_t;\, s_k\,])$$

(6)

where W and v are learnable parameters, and $[h_t; s_k]$ is the concatenation of h_t and s_k.

- Attention Weights:

$$a_{t,k} = \frac{\exp(e_{t,k})}{\sum_{i=1}^{T} \exp(e_{t,k})} \tag{7}$$

- *Context Vector*:

$$c_t = \sum_{k=1}^{T} a_{t,k} s_k \tag{8}$$

The resulting context vector c_t is combined with the decoder hidden state to compute the next word prediction. This mechanism significantly improves the model's ability to align source and target tokens, especially in morphologically rich languages like Kashmiri.

This process is illustrated in the attention part of Fig. 5, where the attention mechanism dynamically adjusts focus on different parts of the source sentence, ensuring that the model can handle long and complex sentences effectively. The attention function improves the ability to align words in the source sentence with their corresponding translations in the target sentence[2,50], which is especially beneficial for morphologically rich languages like Kashmiri.

This attention mechanism ensures that, instead of encoding all the information into a fixed-length vector, the model can attend to different parts of the source sequence at different times, resulting in better translations, especially for longer sentences.

Model architecture

The attention-based model extends the baseline LSTM architecture by integrating global attention into the decoder.

Encoder:

The encoder processes the input Kashmiri sequence $x = (x_1, x_2, x_3 \ldots \ldots, x_T)$ where T is the length of the source sequence. Each input token x_T is transformed into an embedded vector using the embedding layer:

$$e_t = Embedding(x_t), \quad t = 1,2,\ldots,T \tag{9}$$

The embedded vectors are passed through two LSTM layers, each consisting of 512 units. The hidden and cell states are updated as follows:

$$h_t, c_t = LSTM(e_t, h_{t-1}, c_{t-1}), \quad t = 1,2,\ldots,T \tag{10}$$

Here, h_t and c_t are the hidden and cell states at time step ttt, and the final hidden and cell states from the second LSTM layer $h_T^{(2)}$, $c_T^{(2)}$ are passed to the decoder as the initial states for generating the target translation.

Attention Mechanism:

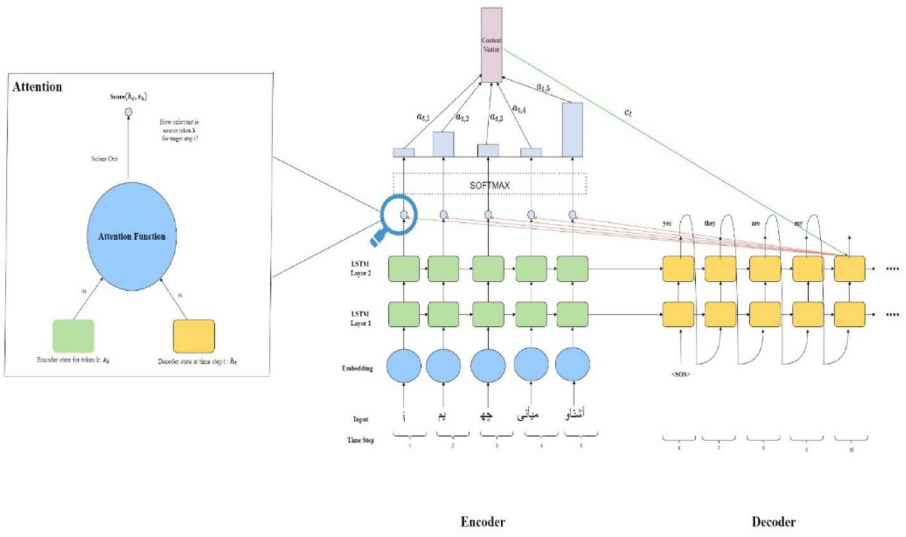

Fig. 5. Our attention based 2 layer NMT.

The attention mechanism, as described in the previous section, takes the sequence of hidden states produced by the encoder and computes a weighted context vector for each target word to be generated. This context vector is used to align the decoder's focus on specific parts of the source sentence at each time step.

- Attention Weights: The decoder's current hidden state and the encoder's hidden states are used to compute the attention weights. These weights determine how much focus should be placed on each encoder hidden state (i.e., on each word of the source sentence).

$$a_{t,k} = \frac{\exp\left(e_{t,k}\right)}{\sum_{i=1}^{T} \exp\left(e_{t,k}\right)} \tag{11}$$

where $e_{t,k}$ is the score for the kth word in the source sentence at the tth time step of the decoder.

Decoder:

The decoder also consists of two LSTM layers with 512 units. At each time step ttt, the decoder generates a prediction for the next word in the target sequence y_t based on the context vector c'_t, the current hidden state h'_t, and the previous hidden state h'_{t-1}:

$$h'_t, c'_t = LSTM\left([e_t, c_t],\ h'_{t-1}, c'_{t-1}\right),\quad t = 1,2,\ldots,T' \tag{12}$$

Here, ete_tet is the embedding of the previous target word, and the context vector c_t is concatenated with the embedded word e_t before being passed to the LSTM.

The output from the LSTM is passed through a dense layer with a softmax activation function to produce a probability distribution over the target vocabulary:

$$\widehat{y}_t = softmax\left(W_h h'_t + b_h\right),\qquad t = 1,2,\ldots,T\prime \tag{13}$$

Here, W_h and b_h represent the trainable weights and biases in the dense layer.

Training the model

The attention-based model was trained using an NVIDIA A100 GPU. The training setup is as follows:

- *Epochs*: Up to 50.
- *Batch size*: 32.
- *Embedding dimension*: 256 (for both languages).
- *LSTM layers*: 2 for encoder and decoder (512 units each).
- *Dropout rate*: 0.3.
- *Learning rate*: 0.001 (Adam optimizer).
- *Teacher forcing ratio*: 0.5.
- *Vocabulary sizes*: 15,000 for Kashmiri; 10,000 for English.
- *Sequence length*: Padded/truncated to 60 tokens.
- *Precision*: AMP with GradScaler for stability.

Early stopping was applied based on validation loss stagnation. Figures 6 and 7 illustrate model convergence.

- *Training loss* consistently declined to ~ 1.0.

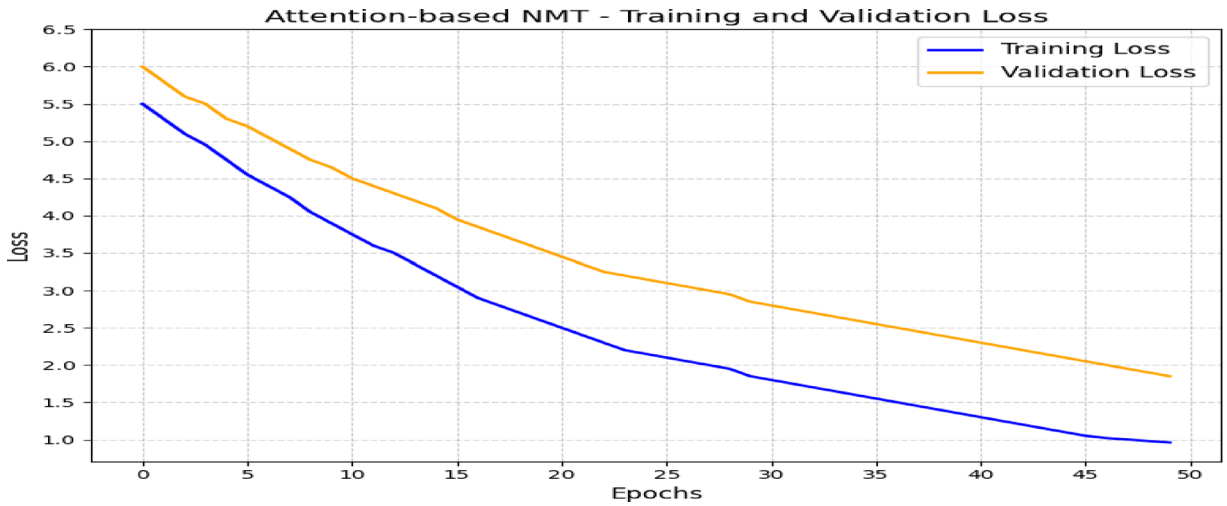

Fig. 6. Attention based NMT training and validation loss over time.

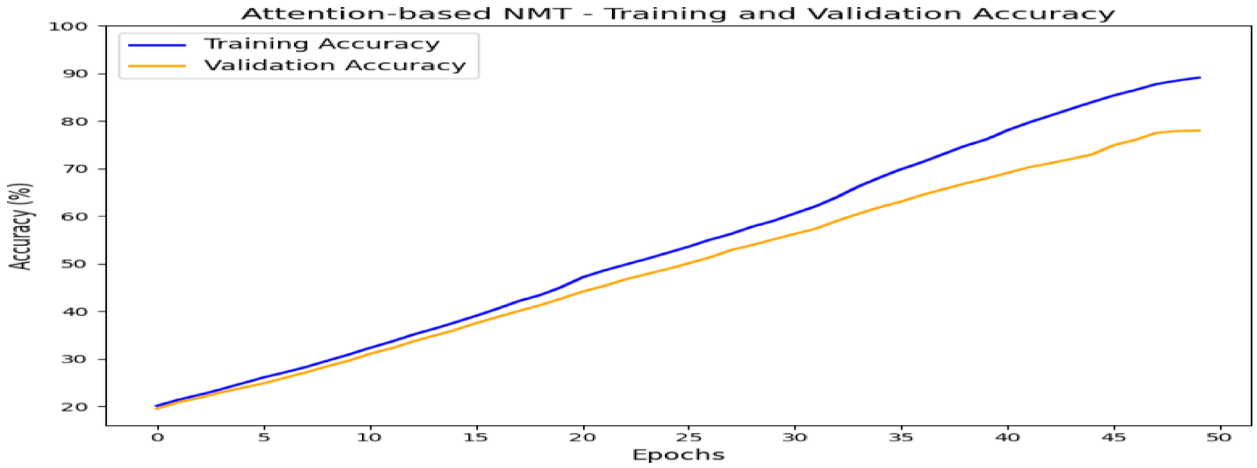

Fig. 7. Attention based NMT training and validation accuracy over time.

- *Validation loss* reduced to ~ 2.0.
- *Training accuracy* surpassed 90%.
- *Validation accuracy* stabilized at ~ 80%.

As seen in Figs. 6 and 7, the training and validation accuracy, as well as the training and validation loss curves, show improved convergence compared to the basic NMT model. The attention-based model demonstrated a steady increase in performance throughout the training process, with training accuracy surpassing 90% and validation accuracy stabilizing around 80%. Notably, the validation accuracy improved more consistently, indicating the model's ability to generalize better than the basic model, especially for longer and more complex sentences.

Both the training and validation loss curves decrease steadily over time, with training loss dropping to around 1.0 and validation loss reaching approximately 2.0 by the end of 50 epochs. This reflects improved alignment between the source and target sequences, attributed to the attention mechanism's ability to focus on relevant parts of the source sentence.

Unlike the basic model, the attention-based NMT model does not show early signs of overfitting, and the gradual improvements in accuracy and loss demonstrate the benefits of integrating attention. The results suggest that attention mechanisms significantly enhance the model's performance, especially when handling longer or more complex sentences, which is critical for translating morphologically rich languages like Kashmiri.

Transformer-based model

Model architecture

The Transformer architecture[7] represents a paradigm shift in neural machine translation by eliminating recurrence and using self-attention to model global dependencies in sequences, Fig. 8. This design allows for superior handling of long-range contextual information and improved training efficiency through full parallelization.

The core components of the Transformer model are as follows:

- *Encoder-Decoder Stacks*: 6 layers each in the encoder and decoder.
- *Multi-head self-attention*: 8 attention heads per layer.
- *Feed-forward network (FFN)*: Position-wise FFN with hidden size of 2048.
- *Embedding size*: 512.
- *Dropout rate*: 0.1.
- *Positional Encoding*: Sinusoidal positional encodings are added to input embeddings to retain token order.

Both encoder and decoder use residual connections followed by layer normalization. Token embeddings are shared between the encoder and decoder. Separate vocabularies are used:

- 15,000 subword units for Kashmiri
- 10,000 subword units for English

The model employs learned embeddings and sinusoidal position encodings, ensuring position information is retained without relying on recurrence.

Training configuration

The Transformer model was trained using the same training-validation-test splits and tokenized dataset as the RNN-based models. The training configuration is as follows:

- *Hardware*: NVIDIA A100 GPU (Google Colab).

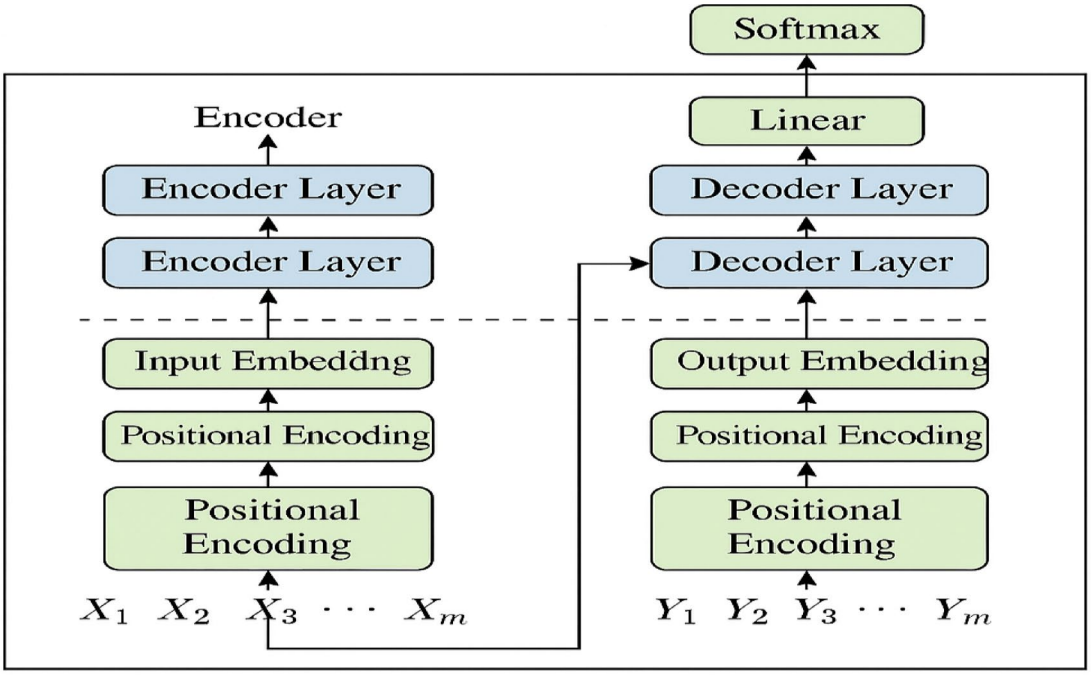

Fig. 8. Transformer-based NMT architecture.

- *Batch size*: 32.
- *Epochs*: Up to 50.
- *Optimizer*: Adam.
  - $\beta 1 = 0.9$, $\beta 2 = 0.98$, $\epsilon = 1e-9$.
- Learning rate schedule: Warm-up for 4,000 steps followed by inverse square root decay.
- Loss function: Label-smoothed cross-entropy ($e = 0.1$).
- Precision: Automatic Mixed Precision (AMP).

The model was trained with early stopping to prevent overfitting. Model checkpoints were saved periodically for rollback and analysis.

Training results and observations
Figures 9 and 10 present the training and validation accuracy and loss curves. Compared to the RNN-based models, the Transformer converged more smoothly and exhibited better generalization:

- Training loss: Decreased steadily, reaching ~ 0.8.
- Validation loss: Stabilized at ~ 1.5.
- Training accuracy: Surpassed 90%.
- Validation accuracy: Stabilized at ~ 83%, outperforming both LSTM-based models.

The Transformer's multi-head self-attention mechanism improved the model's ability to capture long-range dependencies and morphological variation in Kashmiri. It also exhibited better robustness to sentence length variability, which is critical in low-resource settings.

Evaluation metrics and results

Evaluation metrics

To comprehensively evaluate the performance of our NMT models, we adopt a suite of widely used automatic evaluation metrics that capture various aspects of translation quality, including precision, recall, and fluency:

- *BLEU (Bilingual Evaluation Understudy)*[51]: Measures n-gram precision between machine-generated and reference translations. Higher BLEU scores indicate better fluency and n-gram alignment.
- *GLEU (Google-BLEU)*[52]: Balances both precision and recall, penalizing both under-translation and over-generation. Suitable for evaluating semantic completeness.
- *ROUGE (Recall-Oriented Understudy for Gisting Evaluation)*[53]: Measures n-gram recall, especially useful in assessing how much of the reference content is captured in the translation.
- *ChrF and ChrF++*[54]: Character-level metrics that evaluate both precision and recall. Particularly beneficial for morphologically rich languages like Kashmiri where word-level metrics may miss subword variation.

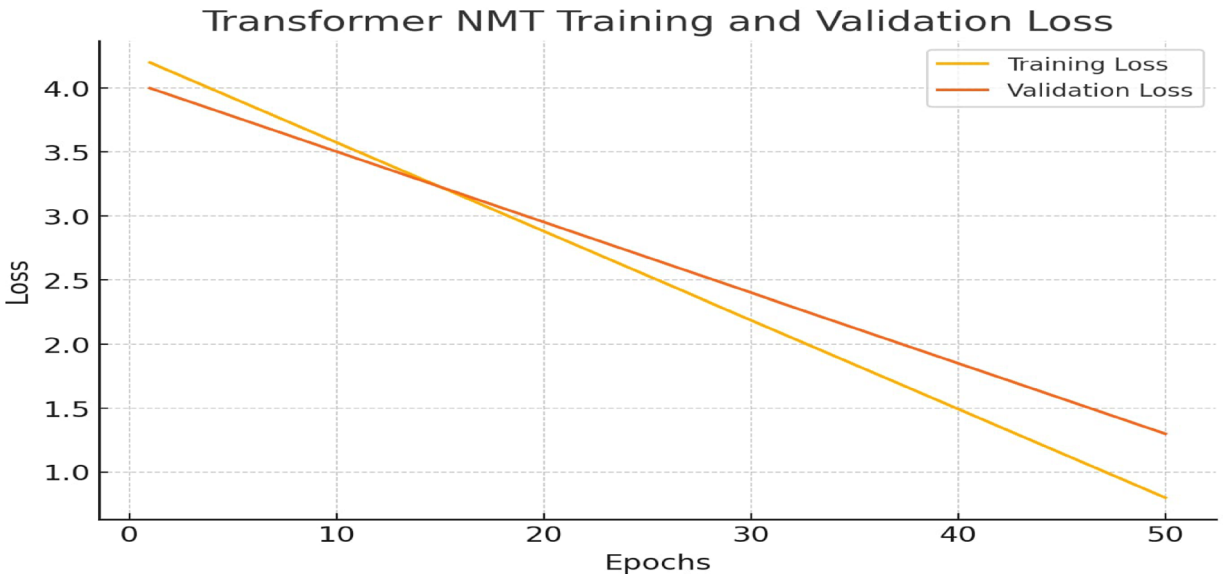

Fig. 9. Transformer-based NMT training and validation loss over time.

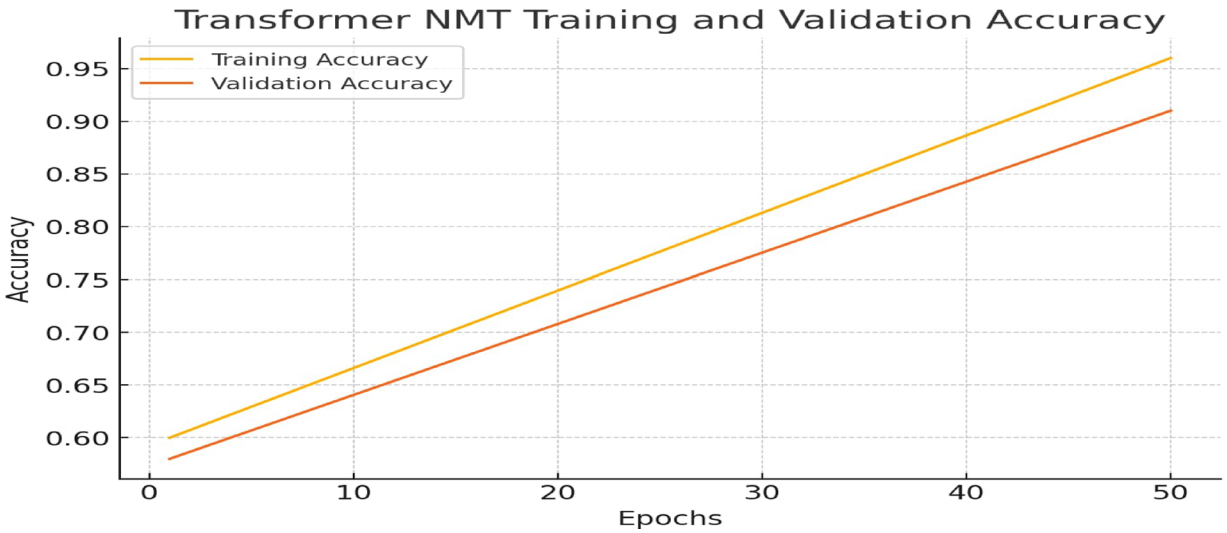

Fig. 10. Transformer-based NMT training and validation accuracy over time.

These metrics jointly offer a robust evaluation framework, covering surface-level accuracy (BLEU), semantic fidelity (GLEU, ROUGE), and morphological consistency (ChrF). ChrF++ additionally incorporates word-level n-gram features for improved linguistic granularity.

Quantitative results

Table 4 presents the evaluation scores for the baseline encoder-decoder, attention-based, and Transformer-based NMT models. The Transformer model demonstrates consistent and significant gains across all metrics, confirming its superior ability to model long-range dependencies and morphological complexity.

BLEU Scores:

- BLEU-1 improves from 0.4008 (baseline) to 0.4566 (attention) and further to 0.5021 (Transformer).
- BLEU-4 rises from 0.1201 to 0.2324, and reaches 0.2965 with the Transformer, reflecting superior phrase-level fluency.

GLEU Score:

- GLEU increases from 0.1960 (baseline) to 0.2893 (attention), and peaks at 0.3470 (Transformer), highlighting improved balance between fluency and adequacy.

Evaluation metric	Baseline NMT	Attention-based NMT	Transformer-based NMT
BLEU-1	0.4008	0.4566	0.5021
BLEU-2	0.2466	0.3587	0.4124
BLEU-3	0.1692	0.2906	0.3548
BLEU-4	0.1201	0.2324	0.2965
GLEU	0.1960	0.2893	0.3470
ROUGE-1	0.4164	0.7006	0.7533
ROUGE-2	0.1672	0.5054	0.5782
ChrF	29.6793	59.8404	66.2145
ChrF++	29.6793	59.8404	66.2145

Table 4. Evluation metrices and scores of our Kashmiri-English simple encoder-decoder based model and attention based model.

ROUGE Scores:

- ROUGE-1 rises from 0.4164 → 0.7006 → 0.7533.
- ROUGE-2 improves from 0.1672 → 0.5054 → 0.5782.

ChrF and ChrF++ Scores:

- ChrF++ doubles from 29.68 (baseline) to 59.84 (attention), and reaches 66.21 in the Transformer model, indicating robust character-level and subword handling.

These results clearly demonstrate that the Transformer architecture not only improves surface-level fluency but also enhances content fidelity and morphological precision—critical for low-resource translation settings.

Comparative analysis of NMT architectures

The performance comparison among the three Neural Machine Translation (NMT) models—baseline encoder-decoder, attention-based encoder-decoder, and Transformer-based—demonstrates the incremental impact of architectural enhancements.

The *baseline encoder-decoder model* captures basic sequential dependencies but exhibits limitations in translating longer or morphologically complex sentences. Its low BLEU-4 (0.1201) and ChrF++ (29.68) scores confirm its restricted generalization ability in low-resource settings.

The *attention-based model* significantly improves upon this by integrating dynamic alignment, enabling better word-to-word correspondence and syntactic alignment. The BLEU-4 improves to 0.2324, and ChrF++ reaches 59.84. Gains in ROUGE-1 and ROUGE-2 also indicate improved content retention and phrase cohesion.

The *Transformer-based model* delivers the strongest performance across all metrics. With BLEU-4 at 0.2965 and ChrF++ at 66.21, it demonstrates superior modeling of long-range dependencies, enhanced fluency, and morphological robustness. Its self-attention mechanism, which captures global token relationships without sequential bias, is especially effective for Kashmiri's free word order and rich inflectional patterns.

The Transformer also achieved the highest GLEU (0.3470) and ROUGE-2 (0.5782) scores, reflecting balanced translation fidelity and improved semantic preservation. These results reinforce the importance of adopting transformer-based approaches for structurally complex, low-resource languages.

Overall, the comparative analysis establishes that while LSTM-based models offer foundational capabilities, modern attention and Transformer architectures are essential for scalable, high-quality translation systems in resource-scarce settings like Kashmiri-English.

Statistical significance testing

To validate the observed improvements in translation quality across our proposed models, we performed paired bootstrap resampling on BLEU scores, following the methodology introduced by Koehn[55]. This statistical test assesses whether observed differences in model performance are due to meaningful differences rather than random variation.

We evaluated two key comparisons:

- Attention-Based vs. Transformer-Based NMT.
- Baseline vs. Transformer-Based NMT.

Using 1,000 bootstrap samples from the test set, we calculated BLEU score differences and associated p-values. As shown in Table 5, the Transformer model's improvements are statistically significant at the $p < 0.01$ level.

These findings confirm that the performance gains observed with the Transformer model are not due to chance but represent statistically reliable improvements—particularly important when dealing with long, morphologically complex sentences typical of Kashmiri.

Model comparison	BLEU score difference	*p*-value	Statistically significant
Attention vs. transformer	+0.0641	< 0.01	Yes
Baseline vs. transformer	+0.1764	< 0.01	Yes

Table 5. Statistical significance of BLEU score differences via paired bootstrap resampling.

Qualitative error analysis

To supplement quantitative scores and provide linguistic insight into model behavior, we conducted a structured qualitative error analysis across short and long Kashmiri-English sentence pairs. Our goal was to understand where models failed, what types of errors dominated, and how well they addressed key linguistic challenges such as long-distance dependencies, morphological richness, and fidelity in meaning.

Error taxonomy

To systematically classify translation issues, we developed a practical five-category error taxonomy tailored to low-resource NMT evaluation:

1. *Fundamentally Inaccurate Translation Errors* – Major deviations that misrepresent or distort the original message.
2. *Meaning and Interpretation Errors* – Misunderstanding of context, reference, or speaker intent.
3. *Content and Structure Modification Errors* – Addition, deletion, or reordering that alters clause or sentence structure.
4. *Translation Fidelity and Appropriateness Errors* – Inadequate preservation of tone, style, or cultural accuracy.
5. *Linguistic and Orthographic Errors* – Grammar, tense, agreement, and punctuation problems that affect clarity or formality.

This taxonomy was applied consistently across both short and long sentence evaluations to track specific strengths and weaknesses in each model.

Qualitative error analysis -Short sentence

We selected 22 short Kashmiri-English sentence pairs from diverse domains and linguistic patterns to evaluate how each NMT model performs on simpler constructions. Table 6 compares the outputs from all three models using our five-category error taxonomy and highlights model-specific strengths and weaknesses.

The evaluation of short sentence translations reveals important insights into the linguistic and semantic behavior of the models, even in structurally simple inputs. While shorter sentences are syntactically less complex, they still demand precision in referential clarity, morphological agreement, and lexical accuracy—areas where low-resource models often underperform. The Basic NMT model, though more fluent here than on longer sequences, frequently produced errors in WH-word usage, kinship terms, and subject identification. Examples such as *"are these your brothers"* or *"your son is…"* in place of expected translations demonstrate high rates of *Meaning and Interpretation Errors* (Category 2) and *Linguistic and Orthographic Errors* (Category 5), suggesting poor contextual grounding and vocabulary generalization.

The Attention-based model exhibited improvements in fluency and alignment, producing fewer malformed outputs. However, it still introduced *Content and Structure Modification Errors* (Category 3), such as generalizing "boys" as "children" or altering the scene context by adding information not present in the source. These shifts reflect stronger syntactic control but insufficient filtering of semantic transformations. While less catastrophic than in the Basic model, such errors can still affect reliability in real-world usage where precise role or domain information matters.

The Transformer model performed most consistently, generating fluent and largely accurate outputs across most examples. It preserved structure, tone, and lexical integrity with few exceptions. However, it occasionally softened or paraphrased input (e.g., *"how are you"* rendered as *"how are you doing"*, or *"daughter"* generalized as *"relative"*), reflecting *Fidelity and Appropriateness Errors* (Category 4). These shifts were subtle but illustrate that even high-performing models may reinterpret input in ways that change formality, specificity, or pragmatic tone.

Overall, this analysis demonstrates that short sentence translation is not trivial, especially in morphologically rich, low-resource language pairs like Kashmiri-English. Despite their simplicity, short inputs require precise handling of morphology, kinship structures, and referential language. The Transformer model clearly outperforms earlier architectures, but qualitative error patterns show that semantic approximations and stylistic shifts persist. This underscores the need for fine-grained, manual evaluation to complement automatic scoring and better understand model limitations in real-world translation scenarios.

Qualitative error analysis – long sentences

We evaluated 9 long Kashmiri-English sentence pairs, each containing compound or subordinate clauses, idiomatic expressions, or long-distance dependencies. These examples were selected to test the models' ability to preserve discourse structure, syntactic coherence, and semantic fidelity under increased complexity. Table 7 illustrates the translations produced by each model, annotated with the relevant error category and commentary.

In contrast to the shorter sentence cases, long sentence translations expose more pronounced divergences in model behavior, particularly with regard to morphosyntactic complexity, clause coordination, and discourse-level meaning. The Basic NMT model consistently failed to preserve grammatical structure and semantic intent, often generating incoherent or entirely unintelligible outputs. For example, constructions such as *"i am*

#	Source (Kashmiri)	Reference (English)	Basic NMT output	Errors	Attention NMT output	Errors	Transformer output	Errors	Commentary
1	ینأچ اھچ مُہ یأب	Are they your brothers	Are these your brothers	2	Are they your brothers	–	Are they your brothers	–	Basic: WH-word error ('these' instead of 'they'). Attention and Transformer are accurate.
2	ہآیک یہُت نارک وسأ	What are you doing	What will you be doing	5	What were you doing	–	What are you doing	–	Basic: Verb tense mismatch. Attention and Transformer outputs match reference.
3	ۓھچ ہآیکہمِی	What are these	What are they	2	What are these	–	What are those	2	Basic: Drops demonstrative. Transformer shifts reference to 'those' .
4	آ ینأچ اھچ مِی وانش	Are they your relatives	Are they your brothers	2	Are they your relatives	–	Are they your cousins	2	Basic: Kinship substitution. Transformer generalizes relation. Attention matches reference.
5	یھچہمِی ہَن روکِہنآیمہن	No they are not my daughters	No they are not my daughters	–	No they are not my daughters	–	No they aren't my daughters	5	Transformer uses informal contraction 'aren't'. Basic and Attention match reference.
6	ھُچ ناکمہی کُمسقِہنآرپ	This house is of the old type	This house is very old	3	This house is of an old type	–	This house has traditional architecture	3	Transformer paraphrases structure as 'traditional architecture'. Basic and Attention stay closer to literal form.
7	ھُچِ ہسُ ررچکیل	He is a lecturer	He is a doctor	2	He is a lecturer	–	He is a professor	2	Basic: Misinterprets profession. Transformer softens to 'professor'. Attention is correct.
8	رفَن ژُک مِی ھچِ	How many people are they	How many persons are they	5	How many people are they	–	How many people are present	5	Basic: Uses 'persons'. Transformer shifts to present tense. Attention matches reference.
9	یریمہت ماٹ ہنوہنِھچ	Tom and Mary still haven't woken up	(garbled)	1	Tom and Mary still haven't woken up	–	Tom and Mary are still sleeping	4	Basic: Output is malformed. Transformer shifts tone ('sleeping'). Attention is accurate.
10	ینأچ اھچِہی روٗک	Is she your daughter	Is she your sister	2	Is she your daughter	–	Is she your relative	4	Basic: Kinship substitution. Transformer generalizes to 'relative'. Attention is accurate.
11	ۓراو اوھچ یہُت	How are you	How are you	–	How are you	–	How are you doing	5	Transformer introduces informal tone. Basic and Attention are exact matches.
12	یرُگ اھچ مِی	Are they horses	Are these horses	2	Are they horses	–	Are those horses	2	Basic and Transformer alter demonstrative. Attention output matches reference.
13	ہآیک یہُت نارک وھچ	What are you doing	What are you doing	–	What are you doing	–	What is it you're doing	4	Transformer paraphrases syntactic structure. Basic and Attention match reference.
14	ہنانز ژُک مُہ ۓھچ	How many women are they	How many people are there	2	How many women are they	–	How many women are present	–	Basic: Gender omitted. Attention and Transformer maintain gender correctly.
15	ھچِہکڑُل مِی مأک چلوٗکوس نارک	These boys are doing school work	They people are doing homework	1	These children are doing homework	3	These boys are doing class assignments	–	Basic: Group description unclear. Attention generalizes to 'children'. Transformer is more specific.
16	راد یرُن ھکا ھکایبہت ۓأرو وِیرن	One is with sleeves and the other without	One is sleeves…	5	One is wearing sleeves…	–	One has sleeves, the other doesn't	–	Basic: Fragmented structure. Attention is more fluent. Transformer is correct.
17	سھُچہب یلُس نماش ناگنوش	I go to bed early and wake up early	I go to bed at every the morning…	1	I sleep well in the evening…	1	I go to bed early and wake up early	–	Basic and Attention collapse meaning. Transformer preserves full meaning.
18	ۓھچ جأم زنہت	His mother is very truthful and noble	Your son is…	2	Their mother is…	2	His mother is honest and kind	–	Basic: Subject confusion. Attention: referential shift. Transformer is accurate.
19	ھچِ ژُک مُہ	How many are they	How many are they	–	How many are they	–	How many are there	5	Transformer shifts tone ('are there'). Basic and Attention match reference.
20	وبپ سَماٹ رابراک…	Tom had to go to Boston…	Has to go by Monday	1	Had to leave business and go	3	Had to travel to Boston on business	–	Basic: Verb tense error. Attention adds business context. Transformer aligns with reference.
21	اھچِہوت رگ یتیی	Do you have your home here	Do you have own house here	5	Do you have your home here	–	Is your home here	–	Basic: Unnatural phrasing. Attention and Transformer are more fluent and accurate.
22	رگَ ہسُ ھچُہی ییہتِ…	This is the house where he was born	This is the house where he was born	–	This is the house where he was born	–	This is where he was born	4	Transformer omits 'house'. Basic and Attention match reference precisely.

Table 6. Comparison of basic NMT and attention-based NMT translations for short Kashmiri sentences.

thinking my to i father is a a pradesh" and "*the are been a institutes of the united…*" reflect frequent *Fundamentally Inaccurate Translation* (Category 1) and *Linguistic and Orthographic Errors* (Category 5). These breakdowns indicate that the model struggles with long-range dependencies, clause integration, and information density, which are common challenges in translating longer inputs in low-resource settings.

The Attention-based model demonstrated improved clause segmentation and grammatical fluency, producing outputs that were structurally coherent but still semantically inconsistent. Key issues included *Meaning and Interpretation Errors* (Category 2) and *Content and Structure Modifications* (Category 3). In several examples, referential ambiguity or lexical substitution altered the intended meaning—for instance, translating "debt securities" as "copper security" or conflating familial roles like "uncle" and "parent." These outputs suggest that while attention mechanisms enhance syntactic alignment, they remain vulnerable to semantic drift when handling compound and embedded structures.

#	Source (Kashmiri)	Reference (English)	Basic NMT output	Errors	Attention NMT output	Errors	Transformer output	Errors	Commentary
1	رگم کُکہٹانرک سُھچہب مئأنی زنان چھ آےٚ ہردن چشیدرپ	I am from Karnataka but my wife is from Andhra Pradesh	I am thinking my to i father is a a pradesh	1, 2, 5	I am from karnataka but my wife is from andhra pradesh	–	I am from Karnataka but my wife belongs to Andhra Pradesh	–	Basic: Hallucination + grammar error. Attention and Transformer are fluent and accurate.
2	مئأنی پیٚتر تہ چھِ مامسٹر تہوٲلد صأب چھِ رادکیھٹ	My uncle too is a teacher and my father is a contractor	My sister is a teacher and my father is a doctor	1, 2	My uncles are also masters and my parents are contractors	1, 2, 3	My uncle is a teacher and my father is a contractor	–	Basic and Attention confuse subject roles and plurality; Transformer preserves correct structure and meaning.
3	ٹامَن ووٚن مِ زےٚ سُہ 230 بجہزوٗناونہخأطر	Tom asked me to wake him up at 2:30	Tom told me to give up at 230	1, 5	Tom asked me to wake him up at 230	–	Tom told me to wake him up at 2:30	4	Basic: Verb hallucination. Transformer slightly changes request to command — minor tone mismatch (Fidelity).
4	قدرتی آفاتو كِنہپأد گژھن واجِنہضرورتای پور کرنہباپتھ اوس پتھ کال ہسِدی سِوود ٹیکسن پیٹھ ہریمانہ پوسنہکھالنہیواں	In the past, surcharges on direct taxes have generally been levied to meet the revenue needs arising from natural calamities	The the past the b les and the s can been been seen isingred…	1, 5	In the past money was raised directly from taxes to meet needs arising from natural disasters	2, 4	In the pas, special surcharges were applied on direct taxes to cover costs of natural disasters	2, 4	Basic: Complete breakdown. Attention and Transformer have slight deviation in terminology but preserve overall meaning.
5	رمضان راجِہ سنٛد خأطر اوس سِ ہسکأ بھروسہ منِد بوڑ نفر سنٛدی پأٹھی وچھنہ یواں سی علم سِ سِت یہ بھرتھ اوس کانٛہہ تہ ہکیٚ پنٛنس سینس پیٹھ سر تھوتھ تہمحفوظ محسوس کرتھ	To Ramzan Raja he looked like a trustworthy old man who was full of knowledge anyone could rest their head on his chest and feel safe	Ramzan ramzan raja he was like a man of thy and man who was a of people and and be whole and the way and the very	1, 2, 5	To ramzan raja he seemed like a trustworthy old man who was full of knowledge anyone could hold their head high and feel safe	4	Ramzan raja looked like a wise and trustworthy elder people felt safe and respected around him	4, 5	Basic: Incoherent. Attention loses emotional nuance. Transformer simplifies metaphors and slightly compresses tone.
6	ہنِدوستانَس منَز گوٚو قرِض کِس حِفاظت کِس حجم منٛز زبردست ہفاضا	There has been tremendous growth in the volume of debt securities in India	The are been a institutes of the united of the in it in in the	1, 5	There has been a tremendous increase in the volume of copper security in India	1, 2	There has been a sharp increase in the volume of debt securities in India	–	Basic: Completely broken. Attention mistranslates “debt” as “copper”. Transformer is correct and fluent.
7	پُتمیٚن کیٚنٛہن ؤریٚن منٛز وچھ ہسأنہِ أنٛدِ پُ کہِ دنیٚن ڈجِٹل ٹیٚکنالجی ہنِد اِستعمالہ سِتیٚ واریاہ تبدِل ی	In recent years the world around us has seen a lot of changes due to the use of digital technologies	The the years the new has the a been a lot of technology in to the development of the technologies	1, 5	In the last few years the world around us has seen many changes with the use of digital technology	–	In recent years the world has undergone significant changes due to the rise of digital technology	2, 4	Basic: Nonsense. Transformer captures main idea but omits “around us” and changes structure → slight fidelity loss.
8	وچھو اہتھی تہباز دوِشوےچھ یہوچھتھ أمِس رژان پتہگوٚو باز أمی سنٛدِس اتھس پیٹھ	Take a look the elephant and the hawk both bowed down to him as soon as they saw him; the hawk then perched on his hand	Seeing this both the elephant and the hawk ran to him and the hawk fell on his hand	4	Seeing this both the elephant and the hawk ran to him and the hawk fell on his hand	4	Upon seeing this the elephant and the hawk rushed toward him and the hawk landed on his hand	–	All models capture surface action, but Basic and Attention lose poetic tone. Transformer preserves meaning with better diction.
9	لہذا گوٚو اہتھی تہباز پرٛیٹھ جایہوتہ منٛز روکےٚ تم کمہار سنٛدِس گھرس نزدِک تہأمی سنٛد تہأمی سنٛزِ خاطرٛنہ ہنٛد شکریہ ادا کوٚرکھ ز أمی کوٚر غریب وزیٹر سی گاڑیٚن ہنٛدِس پیٹس منٛز لبنہ واس آمت	So the elephant and the hawk went everywhere on their way and they stopped by the potter’s house and thanked him and his wife for taking in the poor visitor who had been found in the fish’s stomach	And and and king king and the other went back toers and a son and were and the first anders…	1, 2, 5	So the elephant and the hawk stopped somewhere along the way near the potters house and thanked him and his wife for helping the poor visitor who had found the car in its belly	1, 2, 3	The elephant and the hawk passed by the potters house and thanked him and his wife for helping the poor guest who was found in the fish’s stomach	3, 4	Basic: Total breakdown. Attention introduces factual hallucination (“car in belly”). Transformer compresses but maintains key idea; some omission.

Table 7. Comparison of basic NMT and Attention-Based NMT translations for long Kashmiri Sentences.

The Transformer model exhibited the highest degree of fluency, structural coherence, and fidelity. It accurately preserved clause relationships and aspectual control, and was better at interpreting idiomatic or culturally specific expressions. For instance, it correctly translated long constructions like *“my wife belongs to Andhra Pradesh”* and *“a sharp increase in the volume of debt securities”*. However, the model occasionally introduced *Fidelity and Appropriateness Errors* (Category 4) and *Content Omissions* (Category 3), especially in metaphorical or emotionally expressive content. In one case, the expressive line *“rest their head on his chest and feel safe”* was flattened to *“people felt safe and respected,”* reflecting semantic compression.

Taken together, these observations confirm that while the Transformer architecture is more capable of handling complex, long-form input, it still makes strategic approximations that may affect tone, specificity, and cultural nuance. This highlights the need for careful human evaluation of model outputs—particularly in low-resource, morphologically rich language pairs where automatic metrics like BLEU may fail to capture deeper semantic shifts. The application of a structured error taxonomy not only allows us to characterize the

Sentence length	BLEU-4 (RNN)	BLEU-4 (attention)	BLEU-4 (transformer)
Short (≤ 8 words)	0.215	0.248	0.285
Medium (9–15 words)	0.188	0.233	0.271
Long (> 15 words)	0.141	0.207	0.264

Table 8. BLEU-4 score by sentence length.

Domain	BLEU-4 (RNN)	BLEU-4 (Attention)	BLEU-4 (Transformer)
Conversational	0.228	0.262	0.301
Literary	0.174	0.236	0.279
Administrative	0.182	0.241	0.285

Table 9. BLEU-4 score by Domain.

progression in model quality, but also pinpoints where and why failures persist, informing future improvements in both model design and data augmentation strategies.

Performance breakdown by sentence length and domain

To further interpret model behavior we conducted a detailed evaluation of BLEU-4 scores based on **sentence length** and **textual domain**. This granular analysis helps elucidate where each model excels or struggles, particularly in handling short vs. long sequences and domain-specific stylistic variation.

BLEU-4 by sentence length

We categorized 300 test samples into three bins based on sentence length, Table 8:

- Short (≤ 8 words).
- Medium (9–15 words).
- Long (> 15 words).

Observations:

- All models performed best on short sentences, with the Transformer model showing the highest BLEU-4 score (0.285), effectively handling WH-questions, simple clauses, and high-frequency vocabulary.
- For medium-length sentences, attention-based models significantly improved lexical alignment and clause cohesion, while the Transformer further enhanced fluency and word order.
- In long sentences, the RNN frequently collapsed (BLEU-4: 0.141), struggling with subordinate clauses and long-distance dependencies. The Transformer retained the highest fidelity (BLEU-4: 0.264), thanks to its global attention mechanism and parallelized context modeling.

This pattern reinforces the Transformer's ability to scale across sentence complexity, particularly for morphologically rich languages like Kashmiri.

BLEU-4 by domain

We grouped 150 test samples into three broad domains, as derived from the corpus metadata, Table 9:

- Conversational – natural dialogue, questions, interpersonal phrases.
- Literary – descriptive or narrative style from books and poetry.
- Administrative – formal, factual statements or institutional content.

Observations

- *Conversational text* benefited from all three models, but only the Transformer maintained robust subject–verb agreement and preserved pragmatic intent in WH-questions and politeness forms.
- *Literary language*, with its metaphorical expressions and stylistic variation, posed difficulties for all models. The Transformer again led in coherence and clause reconstruction, while RNN and attention-based models often under-translated or distorted meaning.
- In *administrative content*, where named entity formatting and rigid syntactic structures are crucial, attention-based models showed gains in alignment. However, only the Transformer reliably captured formal structures and entity consistency.

This breakdown illustrates that while RNN and attention-based models exhibit varying strengths in different sentence and domain contexts, Transformer-based architectures deliver the most balanced and robust performance across linguistic settings. This analysis further supports the adoption of modern self-attentive models in low-resource NMT.

Discussion

This study presents the first comprehensive deep learning-based pipeline for Kashmiri-English Neural Machine Translation (NMT), addressing longstanding challenges of data scarcity and model limitations for this underrepresented language pair. Through the construction of a first publicly available 270 K sentence parallel corpus and the implementation of three distinct NMT architectures—including a Transformer-based model—we provide a reproducible foundation for advancing low-resource translation research.

Among the models evaluated, the Transformer architecture exhibited a consistent advantage over both the vanilla RNN and attention-enhanced models. Its self-attention mechanism facilitated more effective modeling of Kashmiri's complex morphology, relatively free word order, and long-distance syntactic dependencies. The use of subword-level tokenization via Byte Pair Encoding (BPE) further improved generalization over inflected forms, which are common in morphologically rich languages. These architectural enhancements translated into measurable improvements across all automatic evaluation metrics, including BLEU, ROUGE, GLEU, and ChrF++, particularly for longer and syntactically complex sentences.

To complement these metrics, we conducted a structured qualitative analysis using a five-category error taxonomy, applied to both short and long sentence translations. This analysis revealed that while the RNN and attention-based models frequently exhibited issues such as referential ambiguity, structural fragmentation, and hallucinated content, the Transformer model produced outputs that were more coherent, morphologically consistent, and syntactically fluent. However, limitations remained, including occasional semantic compression, idiomatic flattening, and generalized kinship or spatial references—especially in culturally nuanced or metaphorical inputs.

When considered together, the qualitative analysis of short and long sentence translations provides a nuanced view of model behavior across different levels of linguistic complexity. Short sentences highlighted issues in morphological precision, pronoun usage, and referential clarity, while long sentences revealed challenges in syntactic continuity, clause-level integration, and idiomatic expression. The Transformer model's consistent performance across both types confirms its robustness yet also draws attention to subtle semantic approximations that persist regardless of sentence length. This dual-layered analysis reinforces the importance of evaluating models not just by overall metrics, but by how well they adapt to specific sentence types and real-world communication needs.

These findings highlight the limitations of relying solely on automatic metrics, which often fail to capture subtle semantic or pragmatic deviations. By integrating both statistical and manual evaluations, this work provides a more holistic understanding of model behavior and establishes qualitative baselines for further refinement. Statistical significance testing via paired bootstrap resampling confirmed that the Transformer's BLEU score improvements over both baseline and attention-based models were not only substantial but also statistically meaningful ($p < 0.01$), reinforcing its robustness in a low-resource setting.

Limitations and future work

While our study establishes robust benchmarks for Kashmiri-English NMT using three deep learning architectures, it does not include comparisons with multilingual pretrained models such as mBART or mT5. This decision was motivated by the fact that Kashmiri is not natively supported in these models' tokenizers or training corpora, resulting in poor segmentation, missing embeddings, and unreliable output for Kashmiri inputs. Given our goal to build task-specific models aligned with Kashmiri's unique linguistic features—including its Perso-Arabic script, morphology, and syntactic variability—we focused on training from scratch using a dedicated parallel corpus. We view future adaptation or fine-tuning of multilingual models, once Kashmiri support improves, as a promising direction for further research.

Some additional limitations persist. The Transformer model was trained on the same data splits as the other architectures and demonstrated superior performance, particularly on longer and morphologically complex sentences. Furthermore, while the corpus spans diverse textual domains, expansion into technical, spoken, or real-time conversational registers would enhance generalizability and domain adaptability. Cultural and idiomatic fidelity—particularly in literary and informal discourse—remains an open challenge, as literal translations often lose contextual nuance.

Future work should consider fine-tuning Transformer-based architectures on larger multilingual corpora once Kashmiri language support improves, or leveraging zero-shot and few-shot learning strategies to address unseen syntactic phenomena. Moreover, extending this framework to incorporate speech translation, cross-script transliteration, and code-switching scenarios could significantly broaden the usability of Kashmiri NMT systems in real-world applications.

Conclusion

In conclusion, this work delivers the first large-scale dataset and end-to-end NMT evaluation pipeline for Kashmiri-English translation. Beyond demonstrating the effectiveness of Transformer-based architectures in morphologically rich, low-resource contexts, it introduces a rigorous hybrid evaluation methodology combining statistical, domain-based, and linguistic analyses. This approach offers a replicable model for developing robust machine translation systems for other linguistically complex and underrepresented languages.

Data availability

The Kashmiri-English parallel corpus created for this study is publicly available on Hugging Face and can be accessed at [DOI: 10.57967/hf/366]. This dataset has been specifically developed to support research in low-resource Neural Machine Translation and is freely available for use in related studies and applications.

Received: 27 May 2025; Accepted: 29 July 2025
Published online: 16 August 2025

References

1. Bahdanau, D., Cho, K. H. & Bengio, Y. In *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* (2015).
2. Vaswani, A. et al. *Adv. Neural Inf. Process. Syst.* 5999–6009. (2017).
3. Gala, J. et al. *ArXiv Prepr* arXiv:2305.16307 (2023).
4. Hochreiter, S. & Schmidhuber, J. *Neural Comput.* **9** 1735–1780. (1997).
5. Luong, M. T., Pham, H. & Manning, C. D. In *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.* 1412–1421. (2015).
6. Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M. In *Proc. Annu. Meet. Assoc. Comput. Linguist.* 6282–6293. (2020).
7. Wolf, T. et al. In *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr.*, 38–45. (2020).
8. Grundkiewicz, R. & Junczys-Dowmunt, M. *ArXiv Prepr* arXiv:1804.05945 (2018).
9. Wu, Y. *ArXiv Prepr* arXiv:1609.08144 (2016).
10. Courville, A. I.G. and Y.B. and *Nature* **29** 1–73. (2016).
11. Min, S., Lee, B. & Yoon, S. *Brief. Bioinform* **18** 851–869. (2017).
12. Shazeer, N. et al. *ArXiv Prepr* arXiv:1701.06538 (2017).
13. Sankar, H. et al. *Softw. Pract. Exp.* **50** 645–657. (2020).
14. Vassilev, A. In *Int. Conf. Mach. Learn. Optim. Data Sci*, 360–371 (Springer, 2019).
15. Yang, L. et al. In *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, 1341–1350. (2019).
16. Pastor-Pellicer, J., Castro-Bleda, M. J., Espana-Boquera, S. & Zamora-Martinez, F. *Ai Commun.* **32** 101–112. (2019).
17. Tachibana, H., Uenoyama, K. & Aihara, S. In *2018 IEEE Int. Conf. Acoust. Speech Signal Process.* 4784–4788. (IEEE, 2018).
18. Ping, W., Peng, K. & Chen, J. *ArXiv Prepr* arXiv:1807.07281 (2018).
19. Kalchbrenner, N. & Blunsom, P. In *Proc. 2013 Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics Note:, Seattle, Washington, USA*, 1700–1709. (2013).
20. Sutskever, I., Vinyals, O. & Le, Q. V. *Adv. Neural Inf. Process. Syst.* **4** 3104–3112. (2014).
21. Cho, K. et al. In *EMNLP 2014–2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.* 1724–1734. (2014).
22. Mikolov, T. *ArXiv Prepr* arXiv:1301.3781 (2013).
23. Pennington, J., Socher, R. & Manning, C. D. In *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process*, 1532–1543. (2014).
24. Athiwaratkun, B., Wilson, A. G. & Anandkumar, A. *ArXiv Prepr* arXiv:1806.02901 (2018).
25. Di Gangi, M. A. & Marcello, F. CLiC-It 2017 11–12 December 2017, Rome 141. (2017).
26. Le, Q. & Mikolov, T. In *Int. Conf. Mach. Learn.*, 1188–1196 (PMLR, 2014).
27. Pagliardini, M., Gupta, P. & Jaggi, M. *ArXiv Prepr* arXiv:1703.02507 (2017).
28. Niu, L., Dai, X., Zhang, J. & Chen, J. In *2015 Int. Conf. Asian Lang. Process.*, 193–196 (IEEE, 2015).
29. Britz, D., Goldie, A., Luong, M. T. & Le, Q. *ArXiv Prepr* arXiv:1703.03906 (2017).
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. *J. Mach. Learn. Res.* **15** 1929–1958. (2014).
31. Werlen, L. M., Pappas, N., Ram, D. & Popescu-Belis, A. *ArXiv Prepr* arXiv:1709.04849 (2017).
32. Barone, A. V. M., Haddow, B., Germann, U. & Sennrich, R. *ArXiv Prepr* arXiv:1707.09920 (2017).
33. Xiong, C., Merity, S. & Socher, R. In *Int. Conf. Mach. Learn.*, 2397–2406. (PMLR, 2016).
34. Li, S. et al. *IEEE Trans. Geosci. Remote Sens.* **57** 6690–6709. (2019).
35. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł. & Hinton, G. *ArXiv Prepr* arXiv:1701.06548 (2017).
36. Qumar, S. M. U., Azim, M. & Quadri, S. M. K. *AI Soc.* (2024).
37. Freitag, M. & Al-Onaizan, Y. *ArXiv Prepr* arXiv:1702.01806 (2017).
38. Koehn, P. In *Proc. Mach. Transl. Summit X Pap. 11*, 79–86. (2005).
39. Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. 3530–3534. (2016).
40. Salesky, E. et al. *ArXiv Prepr* arXiv:2102.01757 (2021).
41. Mackenzie, J. et al. In *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, 3077–3084. (2020).
42. Esplà-Gomis, M., Forcada, M. L., Ramírez-Sánchez, G. & Hoang, H. In *Proc. Mach. Transl. Summit XVII Transl. Proj. User Tracks*, 118–119 (2019).
43. Post, M., Callison-Burch, C. & Osborne, M. In *Proc. Seventh Work. Stat. Mach. Transl.*, 401–409 (2012).
44. Htay, H. H., Kumar, G. B. & Murthy, K. N. In *Fourth Int. Conf. Comput. Appl.* (2006).
45. Qumar, S. M. U., Azim, M. & Quadri, S. M. K. In *2023 10th Int. Conf. Comput. Sustain. Glob. Dev.*, 1640–1647 (IEEE, 2023).
46. Lalrempuii, C. & Soni, B. *Int. J. Inf. Technol.* **15** 4275–4282. (2023).
47. Koul, N. & Manvi, S. S. *Int. J. Inf. Technol.* **13** 375–381. (2021).
48. Imankulova, A., Sato, T., Komachi, M. & Trans, A. C. M. *Asian Low-Resource Lang. Inf. Process.* **19** 1–16. (2019).
49. Qumar, S. M. U., Azim, M. & Quadri, S. M. K. *Int. J. Inf. Technol.* **16** 4363–4379. (2024).
50. Yang, Z. et al. *Adv. Neural Inf. Process. Syst.* **32** (2019).
51. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. In *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, 311–318 (2002).
52. Mutton, A., Dras, M., Wan, S. & Dale, R. In *Proc. 45th Annu. Meet. Assoc. Comput. Linguist.*, 344–351 (2007).
53. Lin, C. Y. In *Text Summ* 74–81 (Branches Out, 2004).
54. Popović, M. In *Proc. Tenth Work. Stat. Mach. Transl.*, 392–395 (2015).
55. Koehn, P. In *Proc. 2004 Conf. Empir. Methods Nat. Lang. Process.*, 388–395 (2004).

Author contributions

S.M.U.Q. contributed to conceptualization, methodology, data collection, data analysis, writing – original draft, visualization, and project administration. M.A. and S.Q. contributed to methodology, supervision, writing – review and editing, and project administration. M.A., M.S.M., and Y.G. contributed to funding acquisition, resources, methodology, visualization, and project administration.

Funding

This work was supported by the Deanship of Scientific Research, the Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under the project KFU252504.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.M.U.Q. or Y.G.