

Article

MythoBiLLM: BiLSTM-Guided Parameter-Efficient Fine-Tuning of Large Language Models for Coherent Summarization and Generation of Indian Mythological Texts

Shweta Bansal ^{1,2} , Sumendra Yogarayan ^{1,3,*}  and Siti Fatimah Abdul Razak ^{1,3} 

- ¹ Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia; shweta.bansal@krmangalam.edu.in (S.B.); fatimah.razak@mmu.edu.my (S.F.A.R.)
² School of Engineering & Technology, K.R. Mangalam University, Sohna 122103, India
³ Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia
 * Correspondence: sumendra@mmu.edu.my

Abstract

Indian mythological narratives contain long event chains, recurring characters, moral conflicts, interactions between human and divine agents, and source-specific narrative styles. General-purpose large language models can generate fluent text while losing character continuity, thematic relations, or source-supported events. This study presents MythoBiLLM, a parameter-efficient framework for summarization and continuation generation from Indian mythological texts. The framework combines a frozen Llama 3.2 3B-Instruct backbone, LoRA-based adaptation, and a gated BiLSTM narrative-memory adapter. A corpus of public-domain English translations from the Ramayana, Mahabharata, Bhagavad-Gita, Vishnupuranam, Harivamsha, Hindu Tales, and Indian Myth and Legend contains 3,684,838 word-level tokens and 6057 segmented passages. Evaluation covers language modeling, summarization, continuation generation, entity consistency, theme retention, component ablation, robustness, human assessment, and statistical testing. Relative to LLM+LoRA, the complete framework reduces average perplexity from 23.4 to 19.8. In controlled comparisons, the BiLSTM adapter achieves an MCS of 0.713 on both tasks, compared with 0.699 for the parameter-matched MLP adapter, 0.704 for independently trained long-context LoRA, and 0.708 for retrieval augmentation. Full MythoBiLLM reaches MCS values of 0.762 for summarization and 0.744 for continuation generation. After entity consistency and style alignment are excluded from MCS, the complete configuration retains the highest scores of 0.751 and 0.731. These findings support the complete framework on the evaluated corpus, while the controlled comparisons indicate a modest complementary contribution from the BiLSTM and do not identify it as the sole source of the performance gains.

Keywords: Indian mythology; large language models; BiLSTM; LoRA; parameter-efficient fine-tuning; mythological summarization; continuation generation; narrative coherence



Academic Editor: Sven Groppe

Received: 15 June 2026

Revised: 11 July 2026

Accepted: 21 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Indian mythological literature contains extended narrative chains, recurring characters, interactions between human and divine agents, moral dilemmas, philosophical dialogues, genealogies, exile, war, sacrifice, devotion, curses, vows, and boons. These works function not only as literary resources but also as culturally situated knowledge sources in which character identity, moral obligation, event order, and narrative style carry substantial

meaning. Computational modeling of these texts therefore requires more than surface-level fluency. A suitable model should preserve entities, event dependencies, themes, and stylistic continuity across extended passages.

Large language models (LLMs) have substantially improved text generation, summarization, dialogue modeling, and instruction following. Transformer architectures support modern language modeling by learning contextual relations through self-attention [1]. General-purpose LLMs can nevertheless generate unsupported or unfaithful content when a task requires strong factual or narrative fidelity [2]. In a mythological narrative, a fluent continuation may assign an action to the wrong character, alter a moral conflict, weaken an earlier vow, or fail to connect a divine promise with its later consequence. Such errors reduce the value of generated material for literary, cultural, educational, and digital-humanities applications.

Domain adaptation provides a straightforward approach for improving performance when the target corpus differs from the broad pretraining distribution [3]. Parameter-efficient fine-tuning methods adapt a small set of parameters while keeping most pre-trained weights frozen. Low-Rank Adaptation (LoRA) adds trainable low-rank updates to selected transformer layers and reduces the cost of domain adaptation [4]. Transformer self-attention already models contextual relationships among tokens within the available input window. The motivation for the BiLSTM adapter is therefore not that the transformer lacks contextual modeling. Instead, the adapter introduces a complementary recurrent sequence bias that explicitly propagates ordered forward and backward states across the available mythological passage.

A BiLSTM processes a sequence in forward and backward directions and represents each position using sequentially accumulated context from both directions [5,6]. This recurrent structure differs from self-attention, which learns flexible pairwise relationships, and from a position-wise MLP, which adds nonlinear capacity without sequential state propagation. The BiLSTM adapter is designed to emphasize ordered narrative dependencies involving characters, vows, curses, promises, moral decisions, and their later consequences. Its contribution is evaluated empirically rather than assumed from the architecture.

Controlled comparisons separate this contribution from three alternative explanations. A parameter-matched MLP tests whether additional adapter capacity alone produces the improvement. An independently trained 4096-token LoRA baseline tests whether a larger context window provides the same benefit. A retrieval-augmented baseline tests whether external narrative evidence can replace the recurrent source-memory mechanism. The BiLSTM adapter achieves an MCS of 0.713, compared with 0.699 for the MLP adapter, 0.704 for longer-context LoRA, and 0.708 for retrieval augmentation. These differences indicate a modest residual advantage for ordered recurrent memory, although retrieval remains stronger on selected entity- and theme-level measures.

This study presents MythoBiLLM, a parameter-efficient framework for summarizing and generating continuations from Indian mythological texts. The framework integrates a base LLM, a BiLSTM narrative-memory adapter, and LoRA-based fine-tuning. The base LLM provides contextual language generation, the BiLSTM adapter introduces forward-backward recurrent sequence modeling, and LoRA supports efficient domain adaptation. The corpus was constructed from public-domain English translations of the Ramayana, Mahabharata, Bhagavad-Gita, Vishnupuranam, Harivamsha, Hindu Tales, and Indian Myth and Legend obtained through Project Gutenberg. The resulting corpus contains 3,684,838 word-level tokens and 6057 segmented passages and supports both source-controlled training and cross-dataset evaluation.

The study evaluates MythoBiLLM on two major tasks: mythological summarization and continuation generation. The validation protocol compares the proposed model with classical, neural, transformer, Indic, and recent open LLM baselines, including Llama 3.2,

Gemma 3, and Qwen3 [7–9]. The evaluation uses language modeling, summarization, generation, entity consistency, theme retention, human evaluation, ablation testing, robustness analysis, and statistical significance testing. Standard metrics such as ROUGE, BERTScore, and MAUVE are used with the proposed mythological coherence score (MCS) to capture both general text quality and mythology-specific consistency [10–12]. The reported results show that MythoBiLLM reduces average perplexity from 23.4 to 19.8 compared with the LLM+LoRA baseline, improves intra-dataset MCS from 0.713 for Qwen3+LoRA to 0.762, and improves continuation-generation MCS from 0.681 to 0.744 over LLM+LoRA.

The main contributions of this study are summarized as follows:

- (1) A BiLSTM-guided narrative-memory adapter is integrated with LoRA-based LLM adaptation to provide an explicit forward–backward sequence bias for mythological passages.
- (2) A controlled comparison isolates the BiLSTM contribution from additional adapter capacity, longer input context, and retrieval through a parameter-matched MLP adapter, a 4096-token LoRA baseline, and a retrieval-augmented baseline.
- (3) A mythology-aware evaluation design measures entity consistency, wrong-entity introduction, theme retention, style alignment, narrative coherence, and the combined mythological coherence score.
- (4) The complete framework is evaluated through summarization, continuation generation, intra-dataset and cross-dataset testing, component ablation, robustness analysis, human evaluation, and statistical testing.

The remaining manuscript is organized as follows. Section 2 reviews related work on literary NLP, domain-specific LLM fine-tuning, BiLSTM-based sequence modeling, and parameter-efficient adaptation. Section 3 presents the proposed MythoBiLLM framework, including corpus construction, preprocessing, segmentation, model architecture, loss functions, and training algorithm. Section 4 describes the datasets, experimental settings, baselines, evaluation metrics, and validation protocol. Section 5 presents the results and discussion. Section 6 concludes the study and outlines future research directions.

2. Related Work

2.1. Natural Language Processing for Literary and Mythological Texts

Existing literary NLP research can be grouped into feature-based literary analysis and structure-based narrative representation. Morphosyntactic annotation and stylometric methods provide interpretable linguistic features for analysing literary style and authorship [13,14]. Their main strength is the ability to relate model decisions to grammatical, lexical, and stylistic properties. Machine-learning studies of poetry extend this direction by combining representations such as TF-IDF and Doc2Vec with classifiers to identify literary categories and thematic patterns [15]. These methods support comparative literary analysis but generally treat the text as an object of classification rather than as a source for controlled generation.

Graph-based narrative approaches provide a complementary strength by representing characters, events, and relationships explicitly [16]. Such representations make narrative structure queryable and interpretable, but they do not directly provide fluent summarization or continuation generation. Across these approaches, the main limitation is that they do not jointly address long-form generation, recurring-character consistency, thematic preservation, and unsupported event introduction in Indian mythological narratives. MythoBiLLM extends literary analysis toward generation while retaining explicit evaluation of entities, themes, style, and narrative coherence.

2.2. Large Language Models for Domain-Specific Text Generation

Large language models provide strong contextual representations and fluent generation across many NLP tasks, but their general pretraining does not ensure reliable

behaviour in specialised narrative domains [17,18]. Domain-adaptive and task-adaptive training address this limitation by exposing the model to the vocabulary, discourse patterns, and reasoning structures of the target domain [3]. The principal strength of this approach is improved domain relevance without redesigning the complete model architecture.

Parameter-efficient adaptation and retrieval provide two widely used extensions. LoRA enables domain adaptation while updating only a small proportion of the model parameters, whereas retrieval-augmented generation supplies external evidence at inference time. He et al. [19] combined parameter-sensitive LoRA with retrieval-augmented reasoning for specialised-domain question answering, while Repede and Brad [20] demonstrated the practical value of LoRA-based LLaMA adaptation for fake-news detection. These studies establish the efficiency of targeted adaptation and the value of retrieved evidence, but their evaluation focuses mainly on question answering or classification.

Indian mythological generation introduces different requirements. A continuation may be fluent while assigning an action to the wrong character, weakening an earlier vow, changing a moral conflict, or introducing an unsupported event. MythoBiLLM therefore combines LoRA-based adaptation with a recurrent narrative-memory path and evaluates it against both longer-context and retrieval-based alternatives. This comparison distinguishes efficient domain adaptation from the additional effects of ordered sequence modelling and external narrative evidence.

2.3. BiLSTM and Sequential Memory in Text Modeling

LSTM networks model ordered dependencies through recurrent state propagation and reduce the vanishing-gradient problem encountered by conventional recurrent networks [5]. BiLSTM extends this mechanism by processing the available sequence in both directions, allowing each position to incorporate preceding and following context [6]. Its main strength is an explicit sequential inductive bias that differs from the flexible pairwise interactions learned by transformer self-attention.

Transformer attention remains more effective for broad contextual modelling and parallel computation, whereas recurrent models encode order through a persistent state. The two mechanisms are therefore not interchangeable. In mythological narratives, recurrent propagation may help connect an earlier vow, boon, curse, or moral decision with a later consequence. However, an observed improvement may also arise from additional trainable parameters or a larger adapter capacity.

MythoBiLLM addresses this ambiguity through a parameter-matched MLP adapter, an independently trained longer-context LoRA model, and a retrieval-augmented baseline. The comparison tests whether recurrent state propagation provides information beyond feed-forward capacity, a wider transformer context, or retrieved passages. The resulting interpretation is limited to a complementary sequence-modelling contribution rather than a claim that recurrent modelling replaces or universally outperforms transformer attention.

2.4. Parameter-Efficient Fine-Tuning of LLMs

Parameter-efficient fine-tuning reduces adaptation cost by updating a limited set of parameters while preserving the pretrained backbone. LoRA represents model updates through low-rank matrices and provides a favourable balance between computational efficiency and task adaptation [4]. Its main strength is that domain-specific behaviour can be learned without the memory and storage requirements of full-model fine-tuning. Comparative PEFT research also shows that adapters and low-rank updates can retain much of the benefit of full adaptation with substantially fewer trainable parameters [19,21].

Most PEFT evaluations focus on classification, question answering, fact checking, or computer-vision tasks. These settings do not directly test whether efficient adaptation

preserves recurring characters, ordered narrative dependencies, moral themes, and stylistic continuity during long-form generation. LoRA also adapts the existing transformer representations but does not itself introduce an explicit recurrent state.

MythoBiLLM uses LoRA for efficient domain adaptation and adds a gated BiLSTM adapter as a separate narrative-memory path. The design retains the efficiency advantage of PEFT while enabling a controlled comparison between feed-forward adaptation, recurrent adaptation, longer context, and retrieval. The evaluation further separates general text quality from domain-specific properties through entity, theme, style, narrative, robustness, and human-assessment measures.

2.5. Research Gap

Existing research leaves four connected gaps. First, literary NLP studies mainly address stylometry, authorship attribution, classification, or narrative knowledge extraction rather than summarization and continuation generation for Indian mythological narratives. Second, domain-adapted LLM studies rarely evaluate character continuity, moral themes, ordered event dependency, and unsupported entity introduction across distinct mythological texts. Third, the contribution of a recurrent narrative-memory path has not been sufficiently separated from additional adapter capacity, a larger transformer context window, or retrieved external evidence. Fourth, mythology-specific evaluation requires component-level measures in addition to general lexical and semantic metrics.

MythoBiLLM addresses these gaps through a public-domain mythological corpus, LoRA-based domain adaptation, a BiLSTM-guided narrative-memory adapter, and controlled comparisons with a parameter-matched MLP adapter, longer-context LoRA, and retrieval augmentation. The evaluation reports standard generation metrics together with entity consistency, theme retention, style alignment, wrong-entity introduction, MCS, cross-dataset performance, robustness, and human assessment.

Table 1 compares the principal strengths, validation settings, limitations, and relationship of the reviewed studies to MythoBiLLM.

Table 1. Comparative strengths, limitations, and relationship of existing studies to MythoBiLLM.

Ref.	Main Strength	Dataset/Domain	Method	Task	Validation	Main Limitation	Relationship to MythoBiLLM
[13]	Provides interpretable grammatical features for literary analysis	Literary texts	Morphosyntactic annotation and stylometry	Style analysis	Stylometric evaluation	Does not perform narrative generation or mythology-specific validation	MythoBiLLM extends interpretable literary properties toward summarization, continuation generation, and coherence assessment
[14]	Synthesises feature, dataset, and model choices for authorship analysis	Authorship datasets	Review of stylometric and machine-learning methods	Author identification	Comparative literature review	Focuses on author identity rather than narrative continuity or generation	MythoBiLLM uses style as one evaluation dimension within a broader generation framework
[15]	Demonstrates the value of multiple text representations for literary classification	Modernist French poetry	TF-IDF, Doc2Vec, and machine-learning classifiers	Literary classification	Classification metrics	Does not address long-form generation or Indian mythological narratives	MythoBiLLM shifts literary NLP from classification toward controlled mythological generation
[16]	Represents narrative entities and relations explicitly	Textual narratives	Graph modelling and structural queries	Narrative representation	Structural querying	Does not integrate parameter-efficient LLM adaptation or generation evaluation	MythoBiLLM models fluent generation while evaluating entities, themes, and narrative relations
[17]	Provides broad coverage of LLM architectures and applications	General LLM research	Literature review	General NLP	Research synthesis	Does not provide a domain-specific narrative model	MythoBiLLM applies LLM adaptation to a specialised public-domain mythology corpus
[18]	Organises LLM challenges and solutions into a structured taxonomy	General LLM research	Taxonomic review	General LLM applications	Structured synthesis	Does not examine cultural narrative continuity or mythology-specific errors	MythoBiLLM evaluates unsupported entities, themes, style, and narrative coherence directly
[19]	Combines LoRA efficiency with retrieval-based domain evidence	Medical and legal domains	Parameter-sensitive LoRA and RAG	Domain question answering	Domain-specific metric comparison	Does not model long-form mythological dependency or continuation generation	MythoBiLLM uses LoRA and compares recurrent memory directly with a retrieval-augmented baseline
[20]	Demonstrates practical LLaMA 3 adaptation using PEFT	Fake-news datasets	LLaMA 3 with LoRA	Classification	Classification metrics	Evaluates classification rather than narrative generation	MythoBiLLM applies PEFT to summarization and continuation generation
[21]	Provides comparative evidence on the efficiency of adapters and LoRA	Instance-segmentation datasets	Adapter and LoRA comparison	Efficient fine-tuning	Accuracy and parameter analysis	Focuses on computer vision rather than language generation	MythoBiLLM evaluates PEFT within a text-generation and narrative-memory setting
[22]	Provides a broad taxonomy of hallucination causes and mitigation methods	LLM-generated text	Survey and taxonomy	Reliability analysis	Literature synthesis	Does not provide mythology-specific entity or event validation	MythoBiLLM measures unsupported characters and narrative inconsistencies directly
[23]	Supports multilingual and Indic-oriented sequence-to-sequence generation	Indic languages and English	Multilingual sequence-to-sequence pretraining	NLG and summarization	Benchmark evaluation	Does not target Indian mythological narrative consistency	MythoBiLLM includes IndicBART as a baseline and adds mythology-specific evaluation
Proposed	Combines efficient LLM adaptation, recurrent narrative memory, and mythology-aware evaluation	Public-domain Indian mythological texts	Llama 3.2 + LoRA + gated BiLSTM adapter	Summarization and continuation generation	Intra-dataset, cross-dataset, controlled baselines, robustness, human evaluation, and statistical testing	Limited to one compact backbone, English translations, and the selected corpus	Integrates the complementary strengths of PEFT and sequential modelling while testing them against matched alternatives

3. Proposed Methodology

The proposed MythoBiLLM framework is designed for Indian mythological text summarization and continuation generation. The methodology has five main stages: public corpus acquisition, text cleaning and normalization, passage-level segmentation, BiLSTM-guided LLM adaptation, and task-specific fine-tuning. The framework uses a base transformer LLM for contextual language generation, a BiLSTM adapter for forward–backward narrative memory, and LoRA for parameter-efficient model adaptation. Transformer models provide strong contextual representation through self-attention [1], while LSTM and BiLSTM models support ordered sequence-memory modeling [5,6]. LoRA reduces fine-tuning cost by updating low-rank trainable matrices instead of the full model weights [4].

3.1. Problem Formulation

Let the Indian mythological corpus be represented as:

$$D = \{(x_i, m_i)\}_{i=1}^N, \quad (1)$$

where x_i denotes a text unit and m_i denotes its metadata, including source text, Gutenberg ID, corpus group, unit type, and task split. Equation (1) defines the complete corpus as a set of text–metadata pairs, which supports reproducible tracking from raw public text to final training samples.

The study addresses two generation tasks. In continuation generation, the model receives a mythological passage $X = \{x_1, x_2, \dots, x_n\}$ and generates a continuation $Y = \{y_1, y_2, \dots, y_m\}$. The conditional generation objective is defined as:

$$P(Y | X) = \prod_{t=1}^m P(y_t | X, y_{<t}). \quad (2)$$

Equation (2) estimates each output token using the input passage and the previously generated output tokens. This task evaluates whether the model can continue a mythological passage while preserving character continuity, event dependency, and narrative tone.

In summarization, the model maps a chapter, canto, story, or passage X into a concise summary S :

$$S = f_{\theta}(X), \quad (3)$$

where f_{θ} denotes the fine-tuned model. Equation (3) formalizes the summarization task, where the generated summary should retain major entities, events, themes, and moral context.

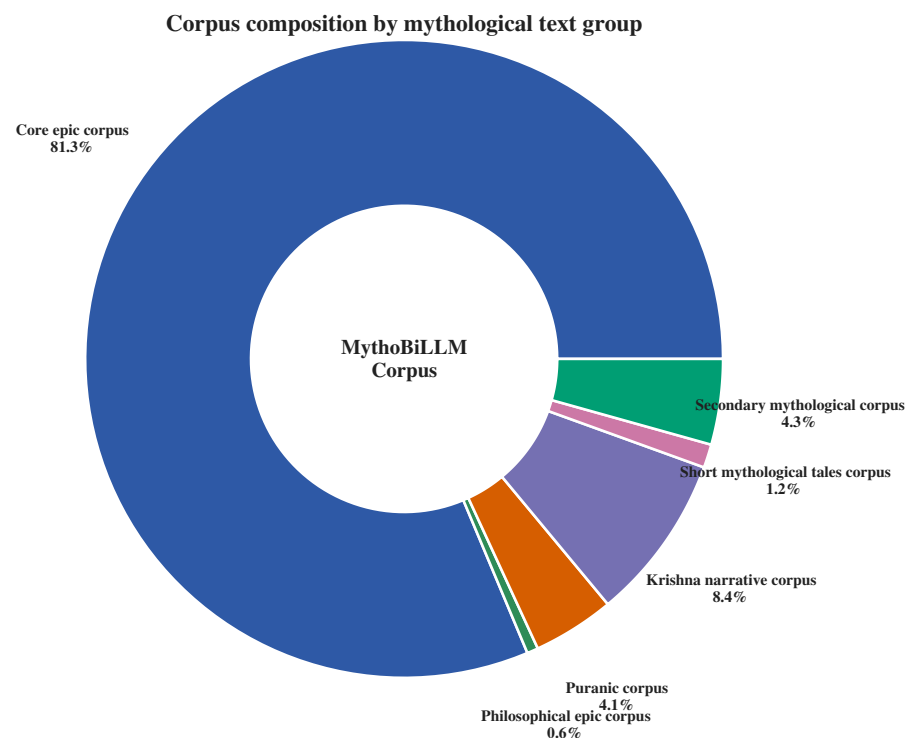
3.2. Corpus Acquisition

The corpus was constructed from ten public-domain English translations obtained through Project Gutenberg. As summarized in Table 2, the collection includes epic narratives, philosophical dialogue, Puranic material, Krishna-centered narratives, short mythological tales, and secondary mythological narration. This combination provides long-form narrative structure together with variation in source style, thematic emphasis, and passage length.

The corpus composition according to text group is displayed in Figure 1. The bulk of the epic corpus consists of the core epics, which are the Ramayana and Mahabharata, which offer long-form narrative structure, repeated character arcs and multiple moral conflicts. The rest of the groups add philosophical, Puranic, Krishna based, short form and external narration styles which helps to enhance the variety of validation styles.

Table 2. Indian mythological datasets used in the study.

ID	Text	Gutenberg ID	Corpus Group	Use in Study
D1	Ramayana	24869	Core epic corpus	Main training and intra-text validation
D2	Mahabharata Vol. 1	15474	Core epic corpus	Main training
D3	Mahabharata Vol. 2	15475	Core epic corpus	Main training
D4	Mahabharata Vol. 3	15476	Core epic corpus	Main training
D5	Mahabharata Vol. 4	15477	Core epic corpus	Training and cross-parva testing
D6	Bhagavad-Gita	2388	Philosophical epic corpus	Philosophical dialogue testing
D7	Vishnupuranam	66208	Puranic corpus	Cross-dataset unseen testing
D8	Harivamsha	61937	Krishna narrative corpus	Cross-dataset unseen testing
D9	Hindu Tales	11310	Short mythological tales corpus	Intra-dataset and annotation analysis
D10	Indian Myth and Legend	47228	Secondary mythological corpus	Corpus-diversity and secondary-source analysis

**Figure 1.** Corpus composition according to mythological text group. Percentages may not total 100% because of rounding.

3.3. Text Cleaning and Normalization

The dataset texts include license blocks, editorial notes, footnotes, page markers, scanner notes, inconsistent spacing and variant spellings. These elements can add noise to the model training. This cleaning stage thus generates a version ready for the model, but keeps the raw version for repeatability. The normalized text is defined as:

$$x_i^{\text{norm}} = f_{\text{norm}}(x_i), \quad (4)$$

where f_{norm} applies Unicode normalization, Gutenberg header/footer removal, spacing correction, scanner-note removal, footnote filtering, and mythological name normalization.

Equation (4) ensures that each raw text unit is converted into a consistent training unit. Table 3 summarizes all preprocessing steps and their purpose.

Table 3. Text preprocessing operations applied to the corpus.

Step	Operation	Purpose	Output
P1	Header/footer removal	Removes Project Gutenberg license blocks from training text	Clean narrative text
P2	Unicode normalization	Standardizes encoding and special characters	Normalized text
P3	Footnote removal	Removes editorial notes and non-narrative blocks	Narrative-only text
P4	Page marker removal	Removes page numbers and scanning artifacts	Continuous text
P5	Name normalization	Maps accented variants to consistent spellings	Model-ready corpus
P6	Sentence boundary preservation	Keeps sentence-level structure for evaluation	Sentence-aware text

3.4. Chapter-Level Segmentation

Related story fragments can be split into the training and testing subsets, due to random sentence splitting. This may overestimate the performance of the model as the same chain of events may be present in different splits. To minimize this risk, the framework first maintains the higher-level boundaries of narrative, like chapter, canto, parva or story unit, wherever possible. Each larger unit is then broken down into fixed-length training passages. A passage is represented as:

$$p_j = [t_j, t_{j+1}, \dots, t_{j+k}], \quad (5)$$

where t_j denotes a token and k denotes the maximum passage length. Equation (5) defines each passage as an ordered token sequence.

The sliding-window passage construction is defined as:

$$p_j = x[j : j + k], \quad p_{j+1} = x[j + s : j + s + k], \quad (6)$$

where s denotes the stride. Equation (6) preserves partial continuity between adjacent passages while keeping the passage length suitable for fine-tuning. Table 4 reports the segment-level word-length distribution. Most passages have a median length of 700 words, which creates comparable context sizes across the corpus. The lower minimum values occur mainly at the final segment of a text or story.

Table 4. Segment-level word-length distribution after passage construction.

Text	Min	Q1	Median	Mean	Q3	Max	Std.
Ramayana	163	700.0	700.0	699.2	700.0	700	20.8
Mahabharata Vol. 1	698	700.0	700.0	700.0	700.0	700	0.1
Mahabharata Vol. 2	537	700.0	700.0	699.9	700.0	700	4.9
Mahabharata Vol. 3	470	700.0	700.0	699.8	700.0	700	6.3
Mahabharata Vol. 4	333	700.0	700.0	699.5	700.0	700	13.5
Bhagavad-Gita	349	700.0	700.0	689.7	700.0	700	60.2
Vishnupuranam	409	700.0	700.0	698.9	700.0	700	18.2
Harivamsha	373	700.0	700.0	699.4	700.0	700	14.4
Hindu Tales	107	700.0	700.0	691.9	700.0	700	69.4
Indian Myth and Legend	381	700.0	700.0	698.8	700.0	700	19.7

Figure 2 visualizes the passage-length distribution after segmentation. The ridgeline plot shows that most datasets concentrate near the target passage length, while shorter values mainly appear in final residual segments. This indicates that the segmentation process produces controlled and comparable passage lengths across texts.

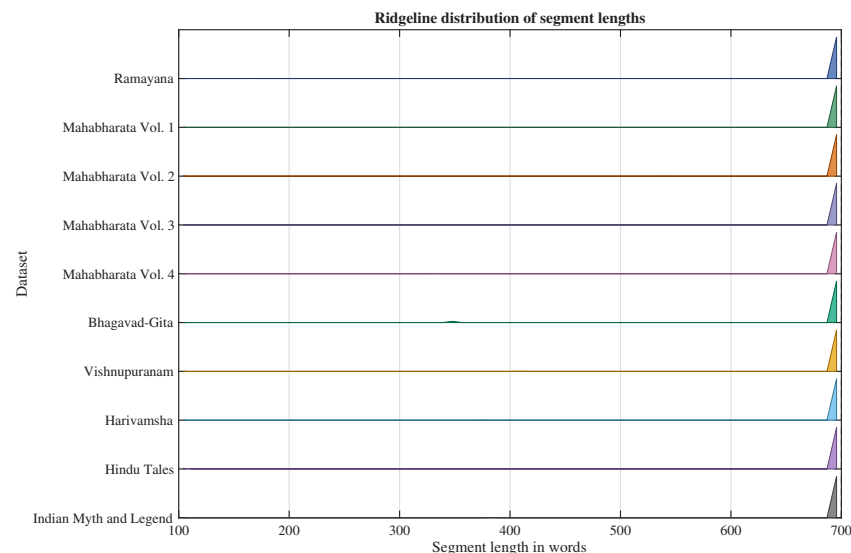


Figure 2. Ridgeline distribution of passage lengths after segmentation.

3.5. Annotation Scheme

A manually checked validation subset was created to evaluate mythology-specific quality beyond generic text metrics. Table 5 defines the annotation scheme. The scheme covers characters, entity roles, narrative events, themes, and tone. These labels support later evaluation of entity consistency, theme retention, and human-perceived narrative quality.

Table 5. Annotation scheme for mythological narrative validation.

Annotation Type	Description	Example Labels
Character/entity	Important named entities in the passage	Rama, Sita, Ravana, Krishna, Arjuna, Vishnu
Entity role	Mythological or narrative role	Deity, king, sage, warrior, demon, devotee
Narrative event	Major action or event type	Exile, battle, boon, curse, marriage, journey
Theme	Moral or philosophical concept	Dharma, karma, devotion, sacrifice, fate
Narrative tone	Dominant stylistic or emotional tone	Heroic, devotional, tragic, philosophical, moral

The entity set for the i th passage is defined as:

$$E_i = \{e_1, e_2, \dots, e_m\}, \quad (7)$$

and the theme set is defined as:

$$T_i = \{t_1, t_2, \dots, t_r\}. \quad (8)$$

Equations (7) and (8) are later used to compute entity consistency and theme retention. Annotator agreement is measured using Cohen's kappa [24]:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (9)$$

where P_o is the observed agreement and P_e is the expected agreement by chance. Equation (9) provides a chance-corrected agreement measure for the manually checked validation subset.

3.6. Overall MythoBiLLM Framework

Figure 3 presents the MythoBiLLM architecture. A mythological passage is first processed by the LoRA-adapted Llama 3.2 3B-Instruct backbone. The final transformer layer produces a 3072-dimensional contextual representation for each token. These hidden states are projected to a 256-dimensional adapter space and processed by a two-layer

BiLSTM with 128 hidden units in each direction. A vector-valued fusion gate combines the projected transformer representation with the BiLSTM memory representation. The fused representation is projected back to 3072 dimensions and added to the final transformer hidden state before the language-modeling or task head.

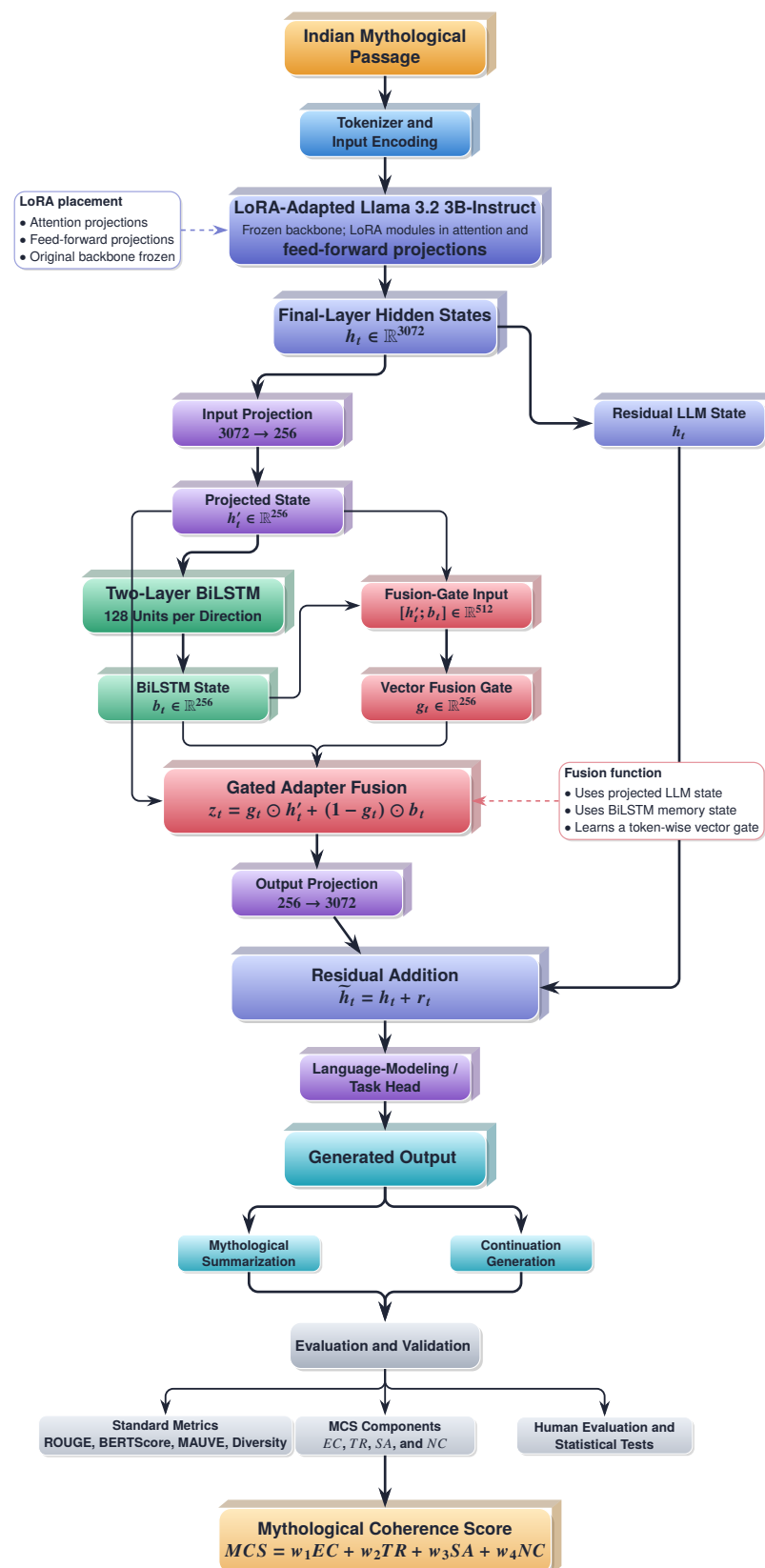


Figure 3. Overall architecture of the proposed MythoBiLLM framework.

The LoRA-adapted backbone produces the final-layer token representations

$$H = \{h_1, h_2, \dots, h_n\}, \quad h_t \in \mathbb{R}^{3072}, \quad (10)$$

where n is the number of tokens in the available input passage. A trainable input projection maps each transformer hidden state to the adapter dimension:

$$h'_t = W_{\text{in}} h_t + c_{\text{in}}, \quad W_{\text{in}} \in \mathbb{R}^{256 \times 3072}, \quad h'_t \in \mathbb{R}^{256}. \quad (11)$$

The projected sequence $H' = \{h'_1, \dots, h'_n\}$ is processed by two stacked bidirectional LSTM layers. Each direction contains 128 hidden units. The first layer is defined as

$$B^{(1)} = \text{BiLSTM}^{(1)}(H'), \quad B^{(1)} \in \mathbb{R}^{n \times 256}, \quad (12)$$

and the second layer is

$$B = \text{BiLSTM}^{(2)}\left(\text{Dropout}_{0.10}\left(B^{(1)}\right)\right), \quad B \in \mathbb{R}^{n \times 256}. \quad (13)$$

For token position t , the output of the second BiLSTM layer is

$$b_t = [\vec{h}_t; \overleftarrow{h}_t], \quad \vec{h}_t, \overleftarrow{h}_t \in \mathbb{R}^{128}, \quad b_t \in \mathbb{R}^{256}. \quad (14)$$

The projected transformer state and the BiLSTM state are concatenated to obtain a 512-dimensional gate input. A vector-valued fusion gate is then computed as

$$g_t = \sigma(W_g[h'_t; b_t] + c_g), \quad W_g \in \mathbb{R}^{256 \times 512}, \quad g_t \in \mathbb{R}^{256}, \quad (15)$$

where $\sigma(\cdot)$ is the sigmoid function. The fused adapter representation is

$$z_t = g_t \odot h'_t + (1 - g_t) \odot b_t, \quad z_t \in \mathbb{R}^{256}. \quad (16)$$

An output projection returns the fused representation to the LLM hidden-state dimension:

$$r_t = W_{\text{out}} z_t + c_{\text{out}}, \quad W_{\text{out}} \in \mathbb{R}^{3072 \times 256}, \quad r_t \in \mathbb{R}^{3072}. \quad (17)$$

The adapter output is included through a residual update before the language-modeling or task head:

$$\tilde{h}_t = h_t + \text{Dropout}_{0.10}(r_t). \quad (18)$$

The fusion gate is not trained through a separate objective. Its parameters W_g and c_g are initialized with the adapter and optimized jointly with the BiLSTM, input and output projections, and LoRA parameters through $\mathcal{L}_{\text{total}}$. The original backbone parameters remain frozen. Gradients from the task losses therefore determine, for each token and each adapter dimension, the relative contribution of the projected transformer state and the recurrent memory state.

3.7. LoRA-Based Parameter-Efficient Fine-Tuning

Full LLM fine-tuning requires large memory and computation because all model weights are updated. LoRA reduces this cost by freezing the pretrained weights and learning low-rank update matrices [4]. The adapted weight matrix is defined as:

$$W' = W + \Delta W, \quad (19)$$

where W is the frozen pretrained weight and ΔW is the trainable update. The low-rank update is defined as:

$$\Delta W = \frac{\alpha}{r} BA, \quad (20)$$

where A and B are trainable low-rank matrices, r is the LoRA rank, and α is the scaling factor. Equations (19) and (20) define the parameter-efficient adaptation mechanism used in MythoBiLLM.

LoRA was applied to the attention projections q_proj , k_proj , v_proj , and o_proj , and to the feed-forward projections $gate_proj$, up_proj , and $down_proj$. These modules were inserted into all 28 decoder blocks, indexed from layer 0 to layer 27. No other backbone module received a LoRA update. The final configuration used rank $r = 16$, scaling factor $\alpha = 32$, and LoRA dropout of 0.05. The original transformer weights remained frozen, while the LoRA matrices were jointly optimized with the BiLSTM adapter, fusion gate, and projection layers. The complete implementation configuration is reported in Table 6.

Table 6. Architecture, fusion, LoRA, and auxiliary-loss configuration used in the MythoBiLLM experiments.

Component	Configuration	Implementation Detail
Base LLM	Llama 3.2 3B-Instruct	Frozen pretrained transformer used for contextual hidden-state extraction. The original backbone weights remain frozen, while LoRA, adapter, projection, and fusion parameters are trainable.
LLM hidden-state dimension	3072	Dimension of each final-layer transformer hidden state h_t .
BiLSTM input	Projected LLM hidden states	The adapter receives final-layer token-level LLM hidden states after linear projection rather than raw token embeddings.
Input projection	$3072 \rightarrow 256$	A trainable linear layer maps each LLM hidden state to the adapter space before BiLSTM encoding.
BiLSTM architecture	Two stacked bidirectional LSTM layers	Two BiLSTM layers model forward and backward narrative dependencies within the available passage.
BiLSTM hidden size	128 per direction	Each layer contains 128 hidden units in the forward direction and 128 hidden units in the backward direction.
BiLSTM output dimension	256	Forward and backward states are concatenated as $b_t = [\vec{h}_t; \overleftarrow{h}_t]$.
BiLSTM dropout	0.10	Dropout is applied between the stacked BiLSTM layers and before residual fusion.
Fusion-gate input	512	The gate receives the concatenation of the projected LLM state $h'_t \in \mathbb{R}^{256}$ and the BiLSTM state $b_t \in \mathbb{R}^{256}$.
Fusion-gate output	256-dimensional vector gate	A vector-valued gate is learned at each token position to balance transformer context and BiLSTM memory.
Fusion-gate activation	Sigmoid	The gate is computed as $g_t = \sigma(W_g[h'_t; b_t] + c_g)$.
Fusion training	Joint end-to-end optimization	The gate, BiLSTM, input and output projections, and LoRA parameters are updated jointly through $\mathcal{L}_{total}$, while the original LLM weights remain frozen.
Fusion insertion point	Before the LM or task head	The fused 256-dimensional state is projected to 3072 dimensions and added as a residual update to the final decoder hidden state.
LoRA target modules	$q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj$	LoRA is applied to the attention and feed-forward projection modules.
Transformer layers receiving LoRA	All decoder layers, 0–27	LoRA modules are inserted into all 28 transformer decoder blocks.
LoRA rank and scaling	$r = 16, \alpha = 32$	Final values used in the reported experiments after validation-based tuning.
LoRA dropout	0.05	Dropout applied to the LoRA updates.
Style-loss weight	$\lambda_1 = 0.10$	Weight assigned to $\mathcal{L}_{style}$.
Entity-loss weight	$\lambda_2 = 0.20$	Weight assigned to $\mathcal{L}_{entity}$.
Summarization-loss weight	$\lambda_3 = 0.10$	Weight assigned to $\mathcal{L}_{summary}$.
Loss-weight candidates	$\{0.05, 0.10, 0.20\}$ for each auxiliary loss	A validation-only grid search evaluated $\lambda_1, \lambda_2, \lambda_3 \in \{0.05, 0.10, 0.20\}$, producing 27 candidate combinations.
Weight-selection criterion	Validation ROUGE-L and MAUVE	Validation ROUGE-L was used for summarization and validation MAUVE for continuation generation. MCS and test-set results were not used for loss-weight or checkpoint selection.

3.8. Training Objective

The training objective combines language modeling, style alignment, entity consistency, and task-specific summarization. The final weighted objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + 0.10\mathcal{L}_{\text{style}} + 0.20\mathcal{L}_{\text{entity}} + 0.10\mathcal{L}_{\text{summary}}. \quad (21)$$

The entity-consistency term receives the largest auxiliary weight because preservation of character identity and reduction of unsupported entities are central to mythological faithfulness. The lower style and summarization weights allow these objectives to influence training without dominating the language-modeling objective.

The language modeling loss is:

$$\mathcal{L}_{\text{LM}} = - \sum_{t=1}^n \log P(x_t | x_{<t}). \quad (22)$$

Equation (22) trains the model to predict the next token in the mythological corpus. The entity consistency loss is:

$$\mathcal{L}_{\text{entity}} = 1 - \frac{|E_{\text{ref}} \cap E_{\text{gen}}|}{|E_{\text{ref}} \cup E_{\text{gen}}|}. \quad (23)$$

Equation (23) penalizes missing or unsupported entities by comparing the reference entity set E_{ref} with the generated entity set E_{gen} . This term directly supports character consistency in mythological summarization and continuation generation.

The summarization loss is computed using teacher-forced cross-entropy over the reference summary:

$$\mathcal{L}_{\text{summary}} = - \sum_{t=1}^{|S|} \log P(s_t | X, s_{<t}), \quad (24)$$

where $S = \{s_1, s_2, \dots, s_{|S|}\}$ denotes the reference summary.

The style loss is computed as the distributional difference between source-style and generated-style representations:

$$\mathcal{L}_{\text{style}} = \text{JSD}(P_{\text{source}} \parallel P_{\text{generated}}), \quad (25)$$

where P_{source} and $P_{\text{generated}}$ denote normalized style-feature distributions of the source and generated text, respectively. Equations (24) and (25) complete the auxiliary loss definitions used in the total training objective in Equation (21).

The auxiliary-loss weights were selected through a validation-only grid search. Each of λ_1 , λ_2 , and λ_3 was evaluated at values in $\{0.05, 0.10, 0.20\}$, resulting in 27 candidate combinations. Candidate configurations were trained using the same training split and optimization settings. Validation ROUGE-L was used for the summarization task, while validation MAUVE was used for continuation generation. The final setting was $\lambda_1 = 0.10$, $\lambda_2 = 0.20$, and $\lambda_3 = 0.10$. Neither test-set performance nor MCS was used for loss-weight selection or checkpoint selection, which prevents the mythology-specific evaluation score from directly influencing model selection.

Algorithm 1 presents the complete training procedure. The original Llama 3.2 3B-Instruct parameters remain frozen, while LoRA, projection, BiLSTM, and fusion-gate parameters are jointly optimized. The algorithm obtains the 3072-dimensional final-layer hidden states, projects them to 256 dimensions, applies the two-layer BiLSTM, performs gated residual fusion, and optimizes the weighted objective in Equation (21). Model selection uses validation ROUGE-L for summarization and validation MAUVE for continuation generation.

Algorithm 1 Training procedure of the MythoBiLLM framework

Require: Training corpus D_{train} , validation corpus D_{val} , Llama 3.2 3B-Instruct, auxiliary-weight candidate set $\{0.05, 0.10, 0.20\}$

Ensure: Trained MythoBiLLM parameters

- 1: Clean and normalize the public-domain mythological texts
- 2: Split the corpus at chapter, canto, parva, or story level
- 3: Construct summarization and continuation-generation samples
- 4: Initialize Llama 3.2 3B-Instruct and freeze the original backbone parameters
- 5: Insert LoRA modules into q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj in decoder layers 0–27
- 6: Set LoRA rank $r = 16$, scaling factor $\alpha = 32$, and dropout to 0.05
- 7: Initialize the $3072 \rightarrow 256$ input projection
- 8: Initialize two BiLSTM layers with 128 hidden units per direction and dropout 0.10
- 9: Initialize the 512-dimensional input and 256-dimensional output fusion gate
- 10: Initialize the $256 \rightarrow 3072$ output projection
- 11: **for** each auxiliary-weight combination in the 27-point validation grid **do**
- 12: **for** each training batch **do**
- 13: Obtain final-layer hidden states $H \in \mathbb{R}^{n \times 3072}$
- 14: Project H to $H' \in \mathbb{R}^{n \times 256}$
- 15: Compute $B = \text{BiLSTM}^{(2)}(\text{BiLSTM}^{(1)}(H'))$
- 16: Compute $g_t = \sigma(W_g[h'_t; b_t] + c_g)$
- 17: Compute $z_t = g_t \odot h'_t + (1 - g_t) \odot b_t$
- 18: Compute $\tilde{h}_t = h_t + \text{Dropout}_{0.10}(W_{\text{out}}z_t + c_{\text{out}})$
- 19: Compute $\mathcal{L}_{\text{total}}$ using the current auxiliary weights
- 20: Update LoRA, input/output projection, BiLSTM, and fusion-gate parameters using AdamW
- 21: **end for**
- 22: Evaluate validation ROUGE-L for summarization and validation MAUVE for continuation generation
- 23: **end for**
- 24: Select $\lambda_1 = 0.10$, $\lambda_2 = 0.20$, and $\lambda_3 = 0.10$ using validation-only results
- 25: Select task checkpoints using validation ROUGE-L and MAUVE
- 26: **return** Trained MythoBiLLM parameters

4. Experimental Setup

This section describes the experimental design used to evaluate MythoBiLLM. The setup covers corpus statistics, annotation statistics, data splitting, baseline and recent open LLMs, training configuration, and evaluation metrics. The design aims to test whether the proposed BiLSTM-guided LLM adaptation improves summarization, continuation generation, entity consistency, theme retention, style alignment, and cross-dataset generalization.

4.1. Dataset Description and Corpus Statistics

The experimental corpus was built from ten public-domain Indian mythological texts downloaded from Project Gutenberg. Table 7 reports the cleaned corpus size after preprocessing and segmentation. The corpus contains 194,021 sentences, 3,684,838 word-level tokens, and 6057 passages. The Mahabharata volumes provide the largest portion of the corpus, while the Ramayana, Harivamsha, Vishnupuranam, Hindu Tales, Bhagavad-Gita, and Indian Myth and Legend add complementary narrative styles for validation and cross-dataset testing.

Table 7. Corpus statistics for the unique cleaned source texts and constructed passages.

Text	Sentences	Word-Level Tokens	Passages	Avg. Words/Sent.	Use
Ramayana	21,329	409,734	667	19.2	Main training and intra-text validation
Mahabharata Vol. 1	31,319	639,216	1053	20.4	Main training
Mahabharata Vol. 2	34,074	678,810	1111	19.9	Main training
Mahabharata Vol. 3	46,180	821,145	1350	17.8	Main training
Mahabharata Vol. 4	25,017	447,609	737	17.9	Training and cross-parva testing
Bhagavad-Gita	1276	20,835	34	16.3	Philosophical-dialogue evaluation
Vishnupuranam	7341	152,764	255	20.8	Cross-dataset evaluation
Harivamsha	16,425	311,183	514	18.9	Cross-dataset evaluation
Hindu Tales	2057	43,818	73	21.3	Short-form narrative evaluation
Indian Myth and Legend	9003	159,724	263	17.7	Secondary-source evaluation
Total	194,021	3,684,838	6057	19.0	Complete corpus

Note: Bold values indicate the aggregate statistics for the complete corpus. The sentence and token totals describe the unique cleaned source texts. Passage construction used overlapping windows; therefore, the source-token totals divided by the passage counts do not represent actual passage length. Passage-length statistics are reported separately in Table 4.

Figure 4 complements Table 7 by showing cumulative corpus concentration rather than repeating the individual counts. The curve shows that the four Mahabharata volumes and the Ramayana contribute most of the available text, while the smaller works add stylistic and narrative diversity.

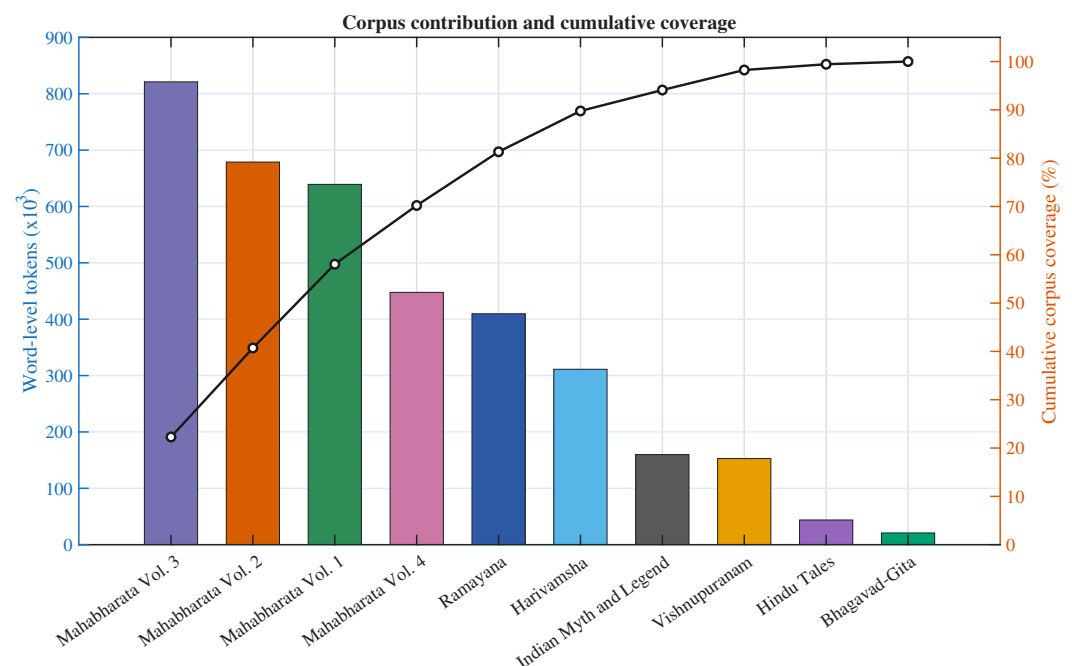
**Figure 4.** Corpus contribution and cumulative coverage across datasets.

Figure 5 provides a normalized comparison across word-level tokens, sentences, passages, average tokens per passage, and average words per sentence. This representation highlights relative corpus structure that is not immediately visible from the absolute values in Table 7.

A manually checked validation subset was prepared to evaluate mythology-specific quality. Table 8 reports 500 annotated passages containing 2474 entity mentions, 903 event labels, and 769 theme labels. Annotation agreement was measured using Cohen's kappa [24]. The overall agreement score of 0.77 indicates acceptable agreement for the manually checked labels. Figure 6 shows the distribution of key mythological themes used for theme-retention evaluation.

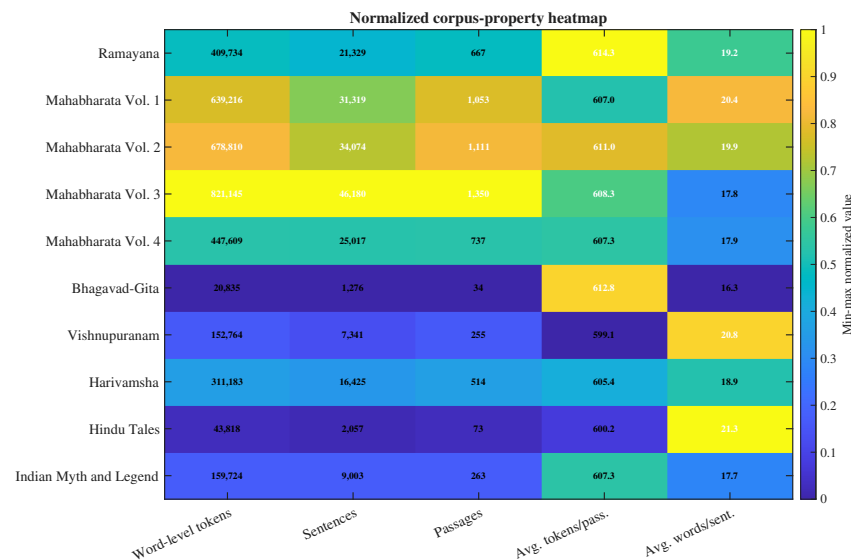


Figure 5. Normalized corpus-property heatmap showing relative differences in corpus scale, passage count, passage density, and sentence-level structure across the collected mythological texts.

Table 8. Annotation statistics for the mythological validation set.

Text	Annotated Passages	Entities	Events	Themes	Avg. Entities/Pass.
Ramayana	125	604	218	188	4.83
Mahabharata	175	986	326	274	5.63
Vishnupuranam	75	342	132	119	4.56
Harivamsha	75	371	143	116	4.95
Hindu Tales	50	171	84	72	3.42
Total	500	2474	903	769	4.95

Annotation agreement was computed on 120 double-annotated passages. The overall Cohen's kappa was 0.77, with 0.82 for entity labels, 0.76 for event labels, and 0.74 for theme labels.

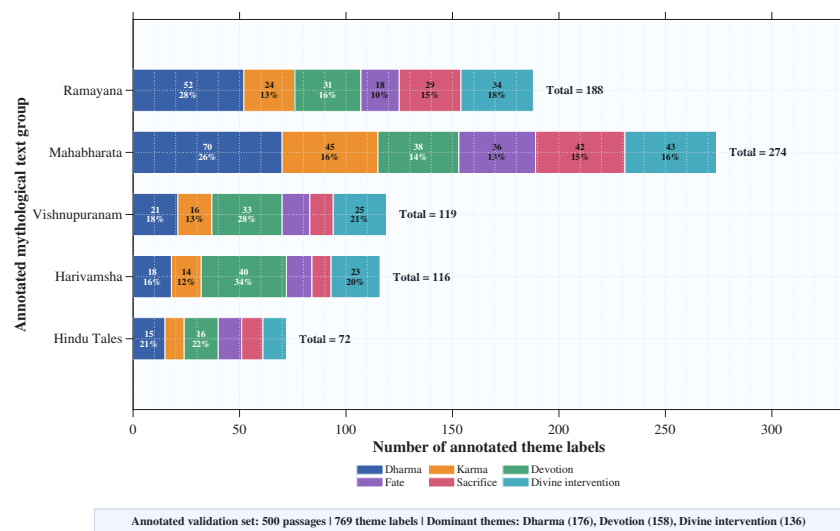


Figure 6. Theme distribution in the annotated mythological validation set.

4.2. Train–Validation–Test Splitting

The corpus was split at chapter, canto, parva, or story level whenever these boundaries were available. This design reduces leakage because related story events are less likely to appear simultaneously in training and testing subsets. Table 9 shows the intra-dataset split used for main experiments. The 70:10:20 setting provides sufficient training data while preserving separate validation and test portions for model selection and final evaluation.

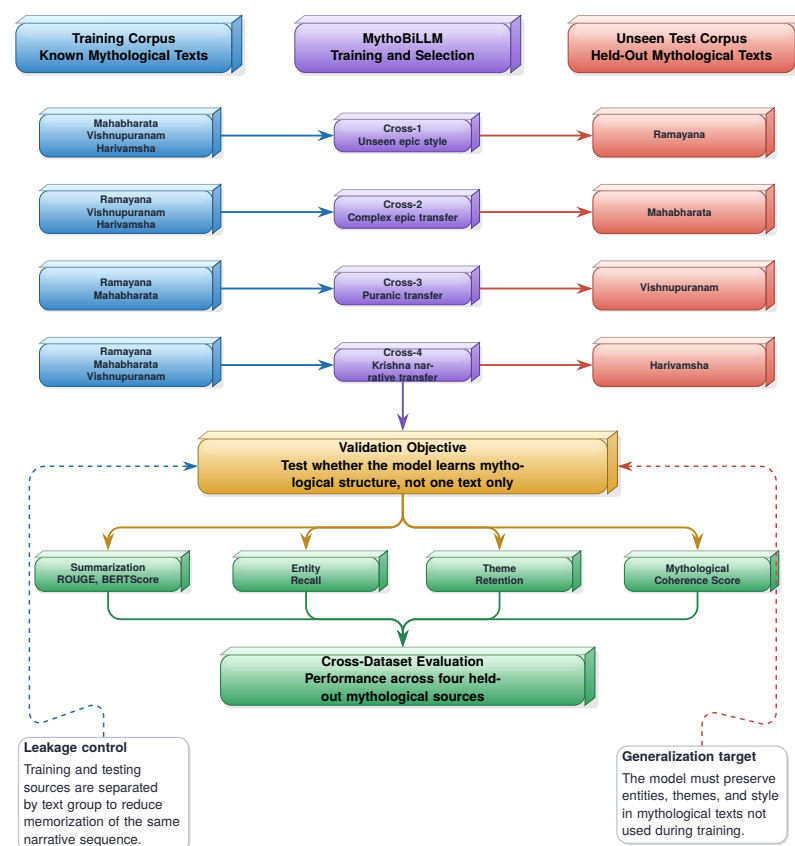
Table 9. Intra-dataset train–validation–test split.

Dataset	Train	Validation	Test	Split Level
Ramayana	70%	10%	20%	Chapter/canto
Mahabharata	70%	10%	20%	Parva/chapter
Vishnupuranam	70%	10%	20%	Chapter
Harivamsha	70%	10%	20%	Chapter
Hindu Tales	70%	10%	20%	Story

Cross-dataset validation was also used because strong performance on the same source text does not establish transfer to unseen sources. Table 10 lists four source-separated cross-dataset settings. Each experiment trains the model on one group of mythological texts and evaluates it on an unseen epic, Puranic, or Krishna-centered source. Figure 7 summarizes this validation design.

Table 10. Cross-dataset validation design.

Experiment	Training Texts	Testing Text	Purpose
Cross-1	Mahabharata + Vishnupuranam + Harivamsha	Ramayana	Tests unseen epic style
Cross-2	Ramayana + Vishnupuranam + Harivamsha	Mahabharata	Tests complex epic generalization
Cross-3	Ramayana + Mahabharata	Vishnupuranam	Tests Puranic generalization
Cross-4	Ramayana + Mahabharata + Vishnupuranam	Harivamsha	Tests Krishna-style transfer

**Figure 7.** Cross-dataset validation protocol for testing mythological generalization.

4.3. Baseline and Recent Open LLMs

The proposed model was compared with classical, neural, transformer, Indic, and recent open LLM baselines. Table 11 summarizes the role of each baseline. The 5-gram model provides a simple statistical baseline. LSTM and BiLSTM models test sequential learning without a large pretrained LLM. GPT-2 and GPT-Neo/GPT-NeoX represent autoregressive

transformer baselines [25,26]. T5 and BART provide strong sequence-to-sequence baselines for summarization [27,28]. IndicBART provides an Indic-oriented NLG baseline [23]. Llama 3.2, Gemma 3, and Qwen3 provide recent open-model comparisons [7–9]. This model pool avoids relying only on weak baselines.

Table 11. Baseline and Recent Open LLMs used for comparison.

Category	Model	Task Role	Reason for Inclusion
Classical baseline	5-gram model	language Continuation generation	Checks whether deep models are necessary
Neural baseline	LSTM model	language Continuation generation	Tests unidirectional sequential modeling
Neural baseline	BiLSTM-only model	Continuation generation	Tests BiLSTM without LLM
Transformer baseline	GPT-2 / GPT-Neo	Generation	Tests standard transformer fine-tuning
Seq2Seq baseline	T5 / BART	Summarization	Tests summarization-specific transformer baseline
Indic baseline	IndicBART	Summarization/generation	Tests Indic-oriented transformer baseline
Recent open LLM	Llama 3.2 1B/3B	Summarization/generation	Tests modern compact LLM baseline
Recent open LLM	Gemma 3	Summarization/generation	Tests modern lightweight LLM baseline
Recent open LLM	Qwen3	Summarization/generation	Tests modern multilingual LLM baseline
Proposed	MythoBiLLM	Summarization/generation	Tests BiLSTM-guided LLM adaptation

4.4. Controlled Adapter, Context-Length, and Retrieval Baselines

A controlled experiment was designed to determine whether the performance gain arises from the BiLSTM sequence-memory mechanism rather than from additional parameters, a larger context window, or external retrieval. All configurations use Llama 3.2 3B-Instruct, the same corpus splits, task instances, LoRA rank and scaling factor, optimizer, learning rate, batch setting, number of epochs, decoding settings, and five random seeds.

The LLM+LoRA configuration provides the 2048-token domain-adapted reference. The MLP configuration replaces the BiLSTM with a feed-forward adapter while retaining comparable input and output dimensions. It contains 28.64 million trainable parameters, compared with 28.65 million for the BiLSTM configuration, which controls for added parameter capacity. The long-context LoRA baseline is trained independently with a 4096-token context and contains the same 26.15 million trainable parameters as the standard LoRA model.

The retrieval-augmented configuration uses a 4096-token input and retrieves additional passages from an index constructed only from the training corpus. Retrieval provides non-trainable external context, and no validation or test passages are included in the index. The LLM+LoRA+BiLSTM configuration uses the 2048-token input and excludes the entity- and style-aware auxiliary objectives. Full MythoBiLLM combines the BiLSTM adapter with the complete training objective. This separation distinguishes the contribution of recurrent narrative memory from the additional gains produced by the auxiliary objectives.

4.5. Experimental Configuration

The experiments were implemented in Python using PyTorch and Hugging Face Transformers [29]. Table 12 reports the software, hardware, optimization, sequence-length, and controlled-baseline settings. Table 6 provides the exact adapter dimensions, fusion-gate training procedure, LoRA insertion locations, trainable modules, auxiliary-loss weights, and validation-only tuning protocol. The main model uses Llama 3.2 3B-Instruct with LoRA rank $r = 16$, scaling factor $\alpha = 32$, AdamW optimization, four training epochs, and five random seeds.

Table 12. Experimental and hardware configuration.

Item	Configuration
Programming language	Python 3.10
Deep learning framework	PyTorch 2.2 and Hugging Face Transformers
GPU	NVIDIA A100 GPU with 40 GB memory
CPU	32-core AMD EPYC processor
RAM	128 GB
CUDA version	CUDA 12.1
Base LLM	Llama 3.2 3B-Instruct
Tokenizer	Model-native tokenizer
Maximum sequence length	2048 tokens for the main and adapter-controlled runs; 4096 tokens for the independently trained long-context and retrieval-augmented baselines
LoRA rank and alpha	$r = 16, \alpha = 32$
LoRA dropout	0.05
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj
LoRA insertion layers	All decoder blocks, layers 0–27
Adapter input/output	3072 $\rightarrow$ 256 input projection and 256 $\rightarrow$ 3072 output projection
BiLSTM configuration	Two bidirectional layers with 128 hidden units per direction and dropout 0.10
Fusion gate	512-dimensional input and 256-dimensional vector output; jointly trained through $\mathcal{L}_{\text{total}}$
Auxiliary-loss weights	$\lambda_1 = 0.10, \lambda_2 = 0.20$, and $\lambda_3 = 0.10$
Loss-weight search	27 validation-only combinations from $\{0.05, 0.10, 0.20\}^3$
Model-selection metrics	Validation ROUGE-L for summarization and validation MAUVE for continuation generation
Optimizer	AdamW [30]
Learning rate	3×10^{-5}
Batch setting	Batch size 4 with gradient accumulation of 8 steps
Epochs	4
Random seeds	1, 2, 3, 4, 5
Controlled trainable parameters	26.15M for LoRA, long-context LoRA, and RAG; 28.64M for the MLP adapter; 28.65M for the BiLSTM adapter and full MythoBiLLM
Parameter-matching control	MLP and BiLSTM configurations differ by 0.01M trainable parameters
Retrieval corpus	Training passages only; retrieval adds no trainable parameters
Controlled comparison	Same base model, corpus splits, LoRA settings, optimization budget, decoding settings, and random seeds
Evaluation tasks	Summarization and continuation generation

4.6. Evaluation Metrics

The evaluation combines automatic metrics, mythology-aware metrics, human evaluation, and statistical testing. Perplexity measures language modeling quality and is computed as:

$$\text{PPL} = \exp \left(-\frac{1}{N} \sum_{t=1}^N \log P(x_t | x_{<t}) \right). \quad (26)$$

A lower value of Equation (26) indicates better next-token prediction.

For summarization, ROUGE-1, ROUGE-2, and ROUGE-L measure lexical overlap with reference summaries [10], while BERTScore measures semantic similarity using contextual embeddings [11]. Entity recall is computed as:

$$ER = \frac{|E_{\text{summary}} \cap E_{\text{reference}}|}{|E_{\text{reference}}|}. \quad (27)$$

Equation (27) evaluates whether the generated summary retains the important mythological entities from the reference.

For continuation generation, MAUVE measures the distributional gap between model-generated text and reference text [12]. Distinct-n measures lexical diversity by counting unique n-grams [31], while Self-BLEU measures repetitive similarity among generated samples [32]. Entity consistency is computed as:

$$EC = \frac{|E_{\text{input}} \cap E_{\text{generated}}|}{|E_{\text{input}} \cup E_{\text{generated}}|}. \quad (28)$$

Equation (28) checks whether generated text preserves entities from the input without adding unrelated characters.

Diversity is measured as:

$$\text{Distinct-}n = \frac{\text{Number of unique n-grams}}{\text{Total number of generated n-grams}}. \quad (29)$$

Equation (29) measures lexical diversity as the ratio of unique generated n-grams to the total number of generated n-grams.

To evaluate mythological quality more directly, the study defines the mythological coherence score:

$$MCS = w_1 EC + w_2 TR + w_3 SA + w_4 NC, \quad (30)$$

where EC denotes entity consistency, TR denotes theme retention, SA denotes style alignment, and NC denotes narrative coherence. Equal weights were used for the primary MCS because no validated empirical or expert-derived weighting standard is available for these four mythology-specific dimensions. Assigning $w_1 = w_2 = w_3 = w_4 = 0.25$ provides a transparent and reproducible default and prevents the score from being designed to favour a particular model strength. The weighting rule was applied consistently to every model and was not tuned using the test set. Alternative component emphasis was examined through the sensitivity analysis in Section 5.5, where each component was omitted separately and the remaining weights were renormalized. A further variant removed both entity consistency and style alignment, the components with the clearest overlap with the auxiliary training objectives. MCS is therefore treated as a composite diagnostic rather than an independent primary outcome and is interpreted together with perplexity, ROUGE, BERTScore, MAUVE, diversity, wrong-entity rate, and blinded human evaluation. MCS was not used for auxiliary-weight selection or checkpoint selection.

$$TR = \frac{|T_{\text{input}} \cap T_{\text{generated}}|}{|T_{\text{input}} \cup T_{\text{generated}}|}. \quad (31)$$

Equation (31) measures thematic overlap between the input passage and the generated text. Style alignment is computed as:

$$SA = 1 - \text{JSD}(P_{\text{source}} \parallel P_{\text{generated}}). \quad (32)$$

Equations (30)–(32) provide a compact diagnostic of mythology-specific output properties. The composite score does not independently establish model superiority because some of its components are related to the auxiliary training objectives. Independent task metrics, human ratings, controlled baselines, and the component-removal analysis are therefore considered alongside MCS.

Narrative coherence was computed using a deterministic sentence-transition coherence procedure. Each generated output was first split into sentences, and sentence-level contextual embeddings were extracted using the same pretrained encoder used for BERTScore evaluation. For an output containing m sentences, adjacent-sentence continuity was calculated as the mean cosine similarity between consecutive sentence embeddings:

$$C_{\text{adj}} = \frac{1}{m-1} \sum_{i=1}^{m-1} \max(0, \cos(v_i, v_{i+1})),$$

where v_i and v_{i+1} denote the contextual embeddings of two adjacent generated sentences. To ensure that the generated passage also remained connected to the source passage, a source-anchoring score was computed as the mean maximum cosine similarity between each generated sentence and the sentence embeddings of the input passage:

$$C_{\text{src}} = \frac{1}{m} \sum_{i=1}^m \max_j \max(0, \cos(v_i, u_j)),$$

where u_j denotes the embedding of the j th source sentence. The final narrative coherence score was computed as

$$\text{NC} = 0.60C_{\text{adj}} + 0.40C_{\text{src}}.$$

This weighting gives slightly greater importance to local sentence-to-sentence continuity while still checking that the generated text remains anchored to the input passage. The procedure was not trained on the test set and was not used for auxiliary-loss tuning or checkpoint selection. For each generated output, the resulting score was clipped to the interval $[0, 1]$ before inclusion in MCS. The same calculation and normalization procedure were applied to every model and evaluation task.

Table 13 summarizes all metrics, their tasks, purposes, and desired direction. The final evaluation uses these metrics together because a single metric cannot fully capture summarization quality, generation diversity, entity preservation, theme consistency, style alignment, and expert-perceived mythological quality.

Table 13. Evaluation metrics used in the study.

Metric	Task	Purpose	Desired Direction
Perplexity	Language modeling	Measures next-token prediction quality	Lower
ROUGE-1	Summarization	Measures unigram overlap	Higher
ROUGE-2	Summarization	Measures bigram overlap	Higher
ROUGE-L	Summarization	Measures longest common subsequence	Higher
BERTScore	Summarization/generation	Measures semantic similarity	Higher
MAUVE	Generation	Measures distributional similarity	Higher
Distinct-n	Generation	Measures lexical diversity	Higher
Self-BLEU	Generation	Measures repetition across outputs	Lower
Entity consistency	Generation	Measures character/entity preservation	Higher
Theme retention	Generation	Measures mythological theme preservation	Higher
Style alignment	Generation	Measures stylistic closeness	Higher
MCS	Generation/summarization	Composite diagnostic of entity, theme, style, and narrative coherence	Higher
Human score	Both tasks	Measures expert-perceived quality	Higher

5. Results and Discussion

This section assesses MythoBiLLM through language modeling, summarization, cross-dataset generalization, continuation generation, controlled adapter comparisons, MCS

component-removal sensitivity, entity consistency, component ablation, robustness analysis, human evaluation, and statistical testing. Lower perplexity, Self-BLEU, and wrong-entity rates indicate better performance, while higher ROUGE, BERTScore, MAUVE, Distinct-n, entity consistency, theme retention, style alignment, MCS, and human scores indicate better performance. MCS is interpreted as a domain-specific diagnostic and not as the sole evidence of model quality.

5.1. Language Modeling Performance

The perplexity results are reported on four Indian mythological test sets in Table 14. The perplexity of the 5-gram language model is the highest with an average of 94.1, which means that a shallow statistical model can only capture the narrative and lexical complexity of the corpus to a limited extent. The baselines of LSTM and BiLSTM-only decrease average perplexity to 49.1 and 43.7, respectively, which shows that sequential memory helps to predict better than n-gram modeling. The perplexity of the transformer baseline is further reduced to 31.6 and perplexity of LLM+LoRA is 23.4. MythoBiLLM obtains the lowest average perplexity of 19.8.

Table 14. Language modeling performance on Indian mythological test sets.

Model	Ramayana PPL	Mahabharata PPL	Vishnupuranam PPL	Harivamsha PPL	Avg. PPL
5-gram LM	94.8	101.6	88.7	91.3	94.1
LSTM LM	49.2	53.7	45.4	47.9	49.1
BiLSTM-only	43.5	47.8	40.9	42.6	43.7
GPT-2/GPT-Neo + FT	31.4	34.7	29.8	30.6	31.6
LLM + LoRA	22.6	25.3	21.8	23.9	23.4
MythoBiLLM	18.9	21.7	18.2	20.3	19.8

The reduction from 23.4 to 19.8 corresponds to a 15.4% relative decrease in average perplexity. LoRA accounts for the largest initial improvement by adapting the backbone to the vocabulary and discourse distribution of the mythological corpus. The additional decrease obtained by the complete framework is consistent with a combination of recurrent sequence modelling, gated residual fusion, and mythology-aware supervision. The recurrent path can propagate ordered contextual information across the available passage, while the gate can retain the original transformer state when the recurrent update is less useful. Because this comparison contrasts the complete framework with LoRA alone, it does not assign the full reduction to the BiLSTM; the matched controls in Section 5.4 provide the narrower architectural comparison.

5.2. Summarization Results

Table 15 reports the intra-dataset summarization results. MythoBiLLM obtains the highest score for every reported measure. Relative to Qwen3+LoRA, ROUGE-L increases from 0.421 to 0.447, BERTScore from 0.896 to 0.904, and entity recall from 0.759 to 0.821. The comparatively small BERTScore difference suggests that the recent open LLMs already generate semantically fluent summaries. The larger improvement in entity recall shows that the main benefit is concentrated in retaining the characters and narrative participants required by the source passage.

Table 15. Intra-dataset summarization results.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Entity Recall	MCS
T5	0.416	0.174	0.334	0.862	0.641	0.566
BART	0.429	0.186	0.348	0.871	0.658	0.584
IndicBART	0.438	0.190	0.358	0.874	0.679	0.607
Llama 3.2 + LoRA	0.484	0.224	0.405	0.889	0.731	0.689
Gemma 3 + LoRA	0.491	0.229	0.414	0.893	0.748	0.704
Qwen3 + LoRA	0.497	0.235	0.421	0.896	0.759	0.713
MythoBiLLM	0.519	0.251	0.447	0.904	0.821	0.762

Entity retention has direct importance for mythological summarization because omitting a principal character or transferring an action to another character can change the meaning of the story even when the summary remains fluent. The improvement in entity recall is consistent with the entity-aware objective, while the simultaneous gains in ROUGE-L and BERTScore show that stronger entity preservation was not obtained by sacrificing general summary quality.

Cross-dataset performance is reported in Table 16. The complete framework improves ROUGE-L, BERTScore, entity recall, theme retention, and MCS on each unseen test source. The improvement is observed across epic, Puranic, and Krishna-centred narratives, which reduces the likelihood that the main result is limited to one source text. These experiments support transfer across the evaluated public-domain translations, although they do not establish generalisation to all mythological styles or languages.

Table 16. Cross-dataset summarization results.

Model	Test Dataset	ROUGE-L	BERTScore	Entity Recall	Theme Retention	MCS
LLM + LoRA	Ramayana	0.384	0.879	0.704	0.671	0.653
MythoBiLLM	Ramayana	0.413	0.891	0.766	0.724	0.712
LLM + LoRA	Mahabharata	0.371	0.873	0.682	0.648	0.631
MythoBiLLM	Mahabharata	0.402	0.886	0.747	0.706	0.694
LLM + LoRA	Vishnupuranam	0.392	0.884	0.727	0.693	0.676
MythoBiLLM	Vishnupuranam	0.421	0.895	0.781	0.741	0.725
LLM + LoRA	Harivamsha	0.366	0.871	0.675	0.641	0.625
MythoBiLLM	Harivamsha	0.397	0.884	0.739	0.698	0.686

The component-level pattern is similar across the unseen texts. BERTScore changes are smaller than the gains in entity recall and theme retention, suggesting that the framework mainly improves constrained narrative faithfulness rather than basic semantic fluency. A plausible explanation is that LoRA adapts the model to the mythological discourse, while the recurrent path preserves ordered context and the auxiliary objectives penalize character loss and stylistic drift. These mechanisms act jointly, so the cross-dataset improvement should be attributed to the complete configuration rather than to one component alone.

5.3. Continuation Generation Results

Table 17 reports continuation-generation quality and diversity. Compared with LLM+LoRA, MythoBiLLM increases MAUVE from 0.642 to 0.693 and Distinct-2 from 0.319 to 0.353, while reducing Self-BLEU from 0.642 to 0.613. The combination of higher MAUVE, higher lexical diversity, and lower repetition is important because an improvement in reference similarity alone could otherwise be obtained through conservative or repetitive generation. The complete framework produces continuations that are closer to the reference distribution without reducing output variety.

Table 17. Continuation generation results on Indian mythological texts.

Model	MAUVE	BERTScore	Distinct-1	Distinct-2	Self-BLEU	MCS
5-gram LM	0.331	0.771	0.049	0.162	0.803	0.331
LSTM LM	0.412	0.808	0.065	0.206	0.745	0.421
BiLSTM-only	0.438	0.819	0.071	0.224	0.724	0.452
GPT-2/GPT-Neo + FT	0.544	0.851	0.083	0.267	0.681	0.562
LLM + LoRA	0.642	0.884	0.098	0.319	0.642	0.681
MythoBiLLM	0.693	0.899	0.107	0.353	0.613	0.744

The generation pattern is consistent with the complementary functions of the model components. LoRA adapts the backbone to mythological vocabulary and discourse, the BiLSTM path propagates ordered information across the available passage, and the fusion gate controls the contribution of the recurrent update at each token. Entity- and style-aware objectives further discourage unsupported characters and stylistic drift. The resulting improvement spans distributional similarity, diversity, repetition, semantic similarity, and mythology-specific coherence, although the full-system comparison cannot uniquely assign each gain to an individual component.

5.4. Controlled Adapter, Context-Length, and Retrieval Comparison

The controlled comparison evaluates whether the BiLSTM contributes beyond LoRA adaptation, additional adapter parameters, a larger context window, and retrieval-based external evidence. All results are reported as mean $\pm$ standard deviation over five random seeds.

Table 18 shows that the MLP and BiLSTM configurations are closely parameter matched, with 28.64M and 28.65M trainable parameters, respectively. Relative to the MLP adapter, the BiLSTM adapter increases ROUGE-L from 0.411 to 0.417, entity F1 from 0.724 to 0.738, theme retention from 0.665 to 0.676, and MCS from 0.699 to 0.713. The BiLSTM configuration also exceeds the 4096-token LoRA and retrieval-augmented baselines in MCS, which reach 0.704 and 0.708, respectively.

Table 18. Controlled comparison of adapter, context-length, and retrieval alternatives for mythological text summarization.

Model Configuration	Context	Trainable Params	ROUGE-L	BERTScore	Entity F1	Theme Retention	Style Alignment	MCS
LLM + LoRA	2048	26.15M	0.405 $\pm$ 0.010	0.889 $\pm$ 0.004	0.711 $\pm$ 0.014	0.654 $\pm$ 0.015	0.691 $\pm$ 0.012	0.689 $\pm$ 0.013
LLM + LoRA + MLP adapter	2048	28.64M	0.411 $\pm$ 0.009	0.891 $\pm$ 0.004	0.724 $\pm$ 0.012	0.665 $\pm$ 0.014	0.697 $\pm$ 0.011	0.699 $\pm$ 0.012
LLM + LoRA (long context)	4096	26.15M	0.414 $\pm$ 0.010	0.892 $\pm$ 0.004	0.728 $\pm$ 0.013	0.670 $\pm$ 0.014	0.700 $\pm$ 0.012	0.704 $\pm$ 0.012
LLM + LoRA + RAG	4096	26.15M	0.416 $\pm$ 0.011	0.893 $\pm$ 0.005	0.736 $\pm$ 0.014	0.684 $\pm$ 0.016	0.703 $\pm$ 0.013	0.708 $\pm$ 0.014
LLM + LoRA + BiLSTM adapter	2048	28.65M	0.417 $\pm$ 0.009	0.892 $\pm$ 0.004	0.738 $\pm$ 0.012	0.676 $\pm$ 0.013	0.704 $\pm$ 0.012	0.713 $\pm$ 0.011
Full MythoBiLLM	2048	28.65M	0.447 $\pm$ 0.008	0.904 $\pm$ 0.003	0.812 $\pm$ 0.011	0.752 $\pm$ 0.012	0.776 $\pm$ 0.010	0.762 $\pm$ 0.010

The MLP adapter approximately matches the trainable parameter count of the BiLSTM adapter. The long-context model was independently trained with a 4096-token context. The RAG index contained only training passages, and retrieval was used as non-trainable external context.

The comparison with retrieval is more mixed at the component level. The BiLSTM produces slightly higher ROUGE-L, entity F1, style alignment, and MCS, whereas retrieval produces slightly higher BERTScore and theme retention. This pattern indicates that the recurrent adapter does not dominate every individual measure. Instead, it provides a small but consistent aggregate advantage while retaining a 2048-token context and avoiding dependence on retrieved passages.

Table 19 separates the strengths of the controlled alternatives. The BiLSTM adapter obtains lower perplexity, higher MAUVE, higher Distinct-2, lower Self-BLEU, and a slightly higher MCS than the MLP, longer-context, and RAG configurations. These differences are modest, which is consistent with the fact that the transformer backbone already provides strong contextual modelling.

Table 19. Controlled comparison of adapter, context-length, and retrieval alternatives for mythological continuation generation.

Model Configuration	PPL	MAUVE	BERTScore	Distinct-2	Self-BLEU	Entity F1	Wrong Entity Rate	Theme Retention	MCS
LLM + LoRA	23.4 ± 0.6	0.642 ± 0.012	0.884 ± 0.004	0.319 ± 0.010	0.642 ± 0.014	0.711 ± 0.014	0.122 ± 0.010	0.654 ± 0.015	0.681 ± 0.013
LLM + LoRA + MLP adapter	22.7 ± 0.5	0.653 ± 0.011	0.887 ± 0.004	0.326 ± 0.009	0.636 ± 0.013	0.724 ± 0.012	0.110 ± 0.009	0.665 ± 0.014	0.699 ± 0.012
LLM + LoRA (long context)	22.3 ± 0.6	0.661 ± 0.012	0.889 ± 0.004	0.332 ± 0.010	0.634 ± 0.013	0.728 ± 0.013	0.108 ± 0.009	0.670 ± 0.014	0.704 ± 0.012
LLM + LoRA + RAG	22.0 ± 0.7	0.665 ± 0.014	0.890 ± 0.005	0.330 ± 0.011	0.631 ± 0.015	0.742 ± 0.014	0.096 ± 0.010	0.688 ± 0.016	0.708 ± 0.014
LLM + LoRA + BiLSTM adapter	21.8 ± 0.5	0.667 ± 0.011	0.890 ± 0.004	0.334 ± 0.009	0.628 ± 0.012	0.738 ± 0.012	0.101 ± 0.009	0.676 ± 0.013	0.713 ± 0.011
Full MythoBiLLM	19.8 ± 0.5	0.693 ± 0.010	0.899 ± 0.003	0.353 ± 0.009	0.613 ± 0.012	0.812 ± 0.011	0.052 ± 0.007	0.752 ± 0.012	0.744 ± 0.010

Lower values are better for PPL, Self-BLEU, and wrong entity rate; higher values are better for all other metrics.

RAG performs better than the BiLSTM-only configuration for entity F1, theme retention, and wrong-entity rate. Retrieved passages provide explicit source evidence and are therefore particularly useful when the required character or theme is present in the retrieval index. The recurrent adapter instead learns a compact representation of sequence order and performs better on language modelling, distributional similarity, and diversity. The two mechanisms capture different information rather than offering interchangeable solutions.

Full MythoBiLLM obtains the strongest overall profile only after recurrent memory is combined with gated fusion and the auxiliary objectives. The difference between the BiLSTM-only and full configurations is therefore best interpreted as an interaction among model components. It does not establish the BiLSTM as the sole cause of the final improvement.

5.5. Sensitivity of MCS to Component Removal

Because entity consistency and style alignment overlap with auxiliary training objectives, the equal-weight MCS may partly reflect dimensions that the complete framework is trained to improve. A component-removal analysis was therefore performed using held-out per-sample values of entity consistency, theme retention, style alignment, and narrative coherence. After removing a component, the remaining weights were renormalized to sum to one. The final variant removes both entity consistency and style alignment and therefore evaluates only theme retention and narrative coherence.

Table 20 reports the sensitivity of MCS to the removal of individual components and to the joint removal of entity consistency and style alignment.

Table 20. Sensitivity of the mythological coherence score to component removal.

Task	MCS Variant	(w_{EC} , w_{TR} , w_{SA} , w_{NC})	LLM + LoRA	MLP Adapter	Long-Context LoRA	RAG	BiLSTM Adapter	Full MythoBiLLM
Summarization	Equal weighting	(0.25, 0.25, 0.25, 0.25)	0.689	0.699	0.704	0.708	0.713	0.762
	Entity consistency omitted	(0, $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	0.692	0.700	0.704	0.706	0.711	0.759
	Theme retention omitted	($\frac{1}{3}$, 0, $\frac{1}{3}$, $\frac{1}{3}$)	0.701	0.710	0.715	0.716	0.725	0.765
	Style alignment omitted	($\frac{1}{3}$, $\frac{1}{3}$, 0, $\frac{1}{3}$)	0.688	0.700	0.705	0.710	0.716	0.757
	Narrative coherence omitted	($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$, 0)	0.675	0.686	0.691	0.701	0.699	0.766
	Entity and style components omitted	(0, 0.50, 0, 0.50)	0.692	0.702	0.706	0.707	0.715	0.751
Continuation	Equal weighting	(0.25, 0.25, 0.25, 0.25)	0.681	0.699	0.704	0.708	0.713	0.744
	Entity consistency omitted	(0, $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	0.684	0.702	0.707	0.707	0.712	0.743
	Theme retention omitted	($\frac{1}{3}$, 0, $\frac{1}{3}$, $\frac{1}{3}$)	0.690	0.710	0.715	0.715	0.725	0.741
	Style alignment omitted	($\frac{1}{3}$, $\frac{1}{3}$, 0, $\frac{1}{3}$)	0.678	0.700	0.705	0.710	0.716	0.737
	Narrative coherence omitted	($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$, 0)	0.672	0.684	0.689	0.701	0.698	0.755
	Entity and style components omitted	(0, 0.50, 0, 0.50)	0.680	0.704	0.710	0.708	0.716	0.731

Each variant was recomputed from held-out per-sample component scores. Weights were renormalized after component removal. The final row excludes entity consistency and style alignment, which have the clearest overlap with the entity- and style-aware training objectives. The equal-weight rows match the controlled MCS values reported in Tables 18 and 19.

Full MythoBiLLM retains the highest score under all six weighting variants for both tasks. When entity consistency and style alignment are removed together, it reaches 0.751 for summarization and 0.731 for continuation generation, compared with 0.715 and 0.716 for the BiLSTM adapter. This result shows that the comparative advantage of the complete framework is not entirely produced by the two components with the clearest overlap with its auxiliary objectives.

The ordering among the controlled alternatives is not completely invariant. The BiLSTM adapter exceeds the parameter-matched MLP and long-context LoRA configurations under every tested variant. It also exceeds RAG in ten of the twelve task-variant compar-

isons. When narrative coherence is omitted, RAG slightly exceeds the BiLSTM adapter by 0.002 for summarization and 0.003 for continuation generation. This small reversal indicates that the individual MCS components emphasize different strengths and that MCS should not be interpreted as a standalone proof of architectural superiority.

The sensitivity analysis reduces, but does not eliminate, the circularity concern. Training objectives can indirectly influence theme and narrative behavior even when their corresponding evaluation components are excluded. For this reason, the main interpretation also considers perplexity, ROUGE-L, BERTScore, MAUVE, diversity, wrong-entity rate, controlled baselines, and blinded human ratings.

5.6. Entity Consistency and Theme Retention Analysis

Table 21 separates the effects on character preservation and thematic continuity. Adding the BiLSTM adapter increases entity F1 from 0.711 to 0.738 and reduces the wrong-entity rate from 0.122 to 0.101. The larger change occurs after the entity-aware objective is added, with entity F1 increasing to 0.783 and the wrong-entity rate decreasing to 0.071. Full MythoBiLLM reaches an entity F1 of 0.812, theme retention of 0.752, and a wrong-entity rate of 0.052. This progression is consistent with recurrent memory providing a modest sequence-level contribution and the targeted objective providing stronger supervision against unsupported character introduction.

Table 21. Entity consistency and theme retention results.

Model	Entity Precision	Entity Recall	Entity F1	Theme Retention	Wrong Entity Rate
LLM + LoRA	0.744	0.681	0.711	0.654	0.122
LLM + LoRA + BiLSTM adapter	0.762	0.716	0.738	0.676	0.101
LLM + LoRA + BiLSTM + entity loss	0.809	0.758	0.783	0.708	0.071
MythoBiLLM	0.834	0.792	0.812	0.752	0.052

Relative to LLM+LoRA, the wrong-entity rate decreases from 0.122 to 0.052, corresponding to a 57.4% relative reduction in the reported error rate. This change has direct practical relevance because an unsupported character can alter kinship, responsibility, allegiance, or the causal structure of a mythological passage. Figure 8 provides information not available in the aggregate table by showing that the improvement extends across named characters and broader groups such as sages, kings, warriors, deities, and demons.

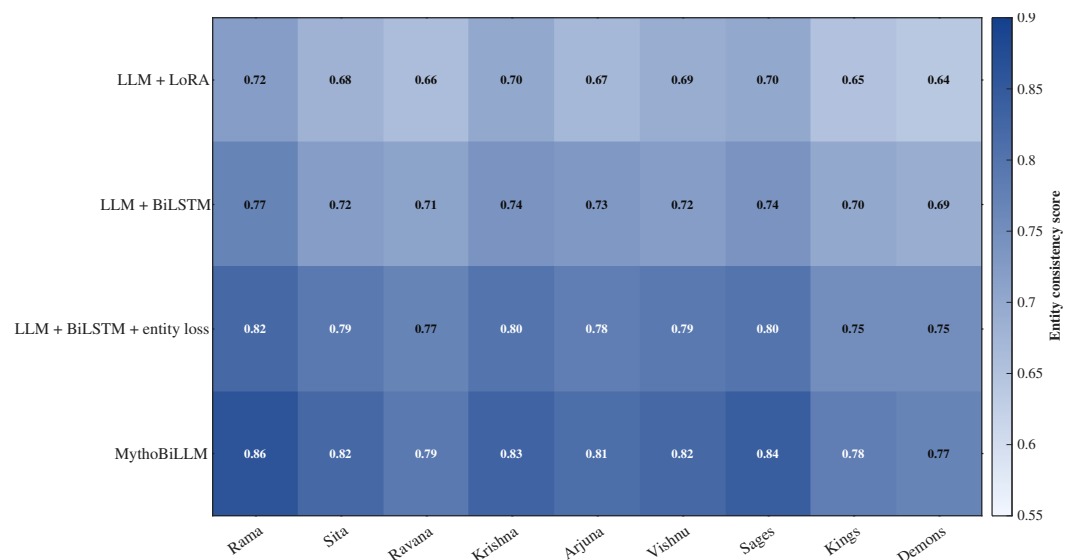


Figure 8. Entity consistency heatmap across major mythological character groups.

5.7. Ablation Study

Table 22 shows that the components contribute at different stages. LoRA produces the largest increase, raising MCS from 0.574 to 0.689, which confirms that domain adaptation is the primary source of improvement over the unadapted backbone. The BiLSTM adapter raises MCS to 0.713, while the parameter-matched MLP reaches 0.699 in the controlled comparison. The remaining difference is compatible with a modest benefit from recurrent sequence propagation rather than additional parameter capacity alone.

Table 22. Ablation study of the proposed MythoBiLLM framework.

Model Variant	ROUGE-L	BERTScore	Entity Consistency	Theme Retention	Style Alignment	MCS
Base LLM only	0.318	0.845	0.596	0.548	0.615	0.574
LLM + LoRA	0.405	0.889	0.702	0.654	0.691	0.689
LLM + LoRA + BiLSTM adapter	0.417	0.892	0.737	0.676	0.704	0.713
LLM + LoRA + BiLSTM + entity loss	0.429	0.898	0.784	0.707	0.721	0.738
LLM + LoRA + BiLSTM + style loss	0.435	0.899	0.771	0.716	0.758	0.746
Full MythoBiLLM	0.447	0.904	0.813	0.752	0.776	0.762

The auxiliary objectives produce more targeted changes. The entity-aware objective increases entity consistency and reduces unsupported entity introduction, whereas the style-aware objective has its largest effect on style alignment and theme retention. The complete configuration benefits from these complementary effects rather than from uniform improvement produced by a single component. The ablation pattern therefore supports a layered interpretation: LoRA provides domain adaptation, the recurrent adapter adds ordered contextual bias, and the auxiliary objectives direct the model toward mythology-specific constraints.

5.8. Robustness Analysis

Table 23 evaluates conditions that place different demands on narrative memory. The largest gains occur for character-heavy prompts (+0.086) and moral-conflict prompts (+0.080). These cases contain multiple character roles, duties, vows, or competing moral constraints, making them more dependent on ordered contextual relations than a simple descriptive prompt. The pattern is therefore consistent with the complete framework being most useful when the input contains dense narrative dependencies.

Table 23. Robustness testing under different mythological input conditions.

Robustness Condition	Test Description	LLM + LoRA MCS	MythoBiLLM MCS	Gain
Unseen text	Train on epics, test on Puranic text	0.622	0.674	+0.052
Different style	Verse-like to prose-like transfer	0.638	0.697	+0.059
Character-heavy prompt	Many entities in same input	0.601	0.687	+0.086
Moral-conflict prompt	Dharma/karma/sacrifice-based input	0.641	0.721	+0.080
Prompt perturbation	Slightly modified prompt wording	0.655	0.707	+0.052
Short context	Reduced input context length	0.616	0.661	+0.045
Long context	Extended input context length	0.670	0.735	+0.065

The smaller gain for short-context inputs (+0.045) shows that the framework cannot recover information that is absent from the prompt. The stronger gain under long-context input (+0.065) suggests that the additional modelling path benefits when more narrative evidence is available. Improvements under unseen-text and altered-style conditions further

show that the model is not restricted to one prompt formulation. For educational and digital-humanities use, this stability is relevant because source passages vary considerably in length, style, number of characters, and moral complexity.

5.9. Human Evaluation

Automatic metrics do not fully capture mythological correctness, moral continuity, or the acceptability of a generated narrative. A blinded human evaluation was therefore conducted on 100 randomly selected outputs. Model identities were concealed, and each output was rated for correctness, coherence, style, theme preservation, fluency, and hallucination control using a five-point Likert scale. Table 24 reports the mean ratings.

Table 24. Human evaluation results.

Model	Correctness	Coherence	Style	Theme	Fluency	Hallucination Control
LLM + LoRA	3.71	3.64	3.68	3.55	4.08	3.53
IndicBART	3.39	3.32	3.28	3.25	3.91	3.20
Llama 3.2 + LoRA	3.83	3.76	3.78	3.68	4.17	3.66
Gemma 3 + LoRA	3.88	3.81	3.84	3.72	4.21	3.70
Qwen3 + LoRA	3.94	3.87	3.90	3.80	4.25	3.77
MythoBiLLM	4.21	4.28	4.16	4.19	4.32	4.13

Scores are mean evaluator ratings on a 1–5 scale over 100 generated samples. Higher values indicate better quality. For hallucination control, a higher score indicates fewer unsupported entities or events.

The difference in fluency between MythoBiLLM and Qwen3+LoRA is relatively small, with scores of 4.32 and 4.25. Larger differences appear for coherence, theme preservation, and hallucination control, where MythoBiLLM scores 4.28, 4.19, and 4.13 compared with 3.87, 3.80, and 3.77. This concentration of improvement agrees with the automatic evaluation: recent LLMs already produce readable text, while the principal benefit of the complete framework lies in maintaining characters, themes, and narrative constraints. Human assessment therefore provides complementary evidence rather than merely repeating the lexical and semantic metrics.

5.10. Statistical Significance Analysis

The experiments were repeated across five random seeds. The mean score across K runs is:

$$\mu = \frac{1}{K} \sum_{i=1}^K x_i. \quad (33)$$

The standard deviation is:

$$\sigma = \sqrt{\frac{1}{K-1} \sum_{i=1}^K (x_i - \mu)^2}. \quad (34)$$

Equations (33) and (34) summarize the variation across runs. Paired bootstrap resampling was used to estimate significance because it can evaluate paired system outputs without assuming a normal distribution [33]. Bootstrap testing was performed with 1000 paired resamples over the evaluated outputs, and a result was considered statistically significant when $p < 0.05$. Effect size was reported using Cohen's d [34].

Table 25 reports statistically significant differences for the tested complete-framework comparisons at $p < 0.05$. These tests support differences between full MythoBiLLM and the listed baselines under the evaluated protocol. They do not isolate the BiLSTM as the cause of those differences because the full configuration also includes LoRA adaptation, gated fusion, and auxiliary objectives. The MCS significance result is interpreted cautiously because MCS contains dimensions related to the training objectives; the independent ROUGE-L, BERTScore, human-rating, perplexity, MAUVE, and diversity results provide complementary evidence.

Table 25. Statistical significance comparison between MythoBiLLM and baseline models.

Comparison	Metric	Mean Difference	p-Value	Effect Size	Significant
MythoBiLLM vs LLM + LoRA	MCS	+0.073	0.012	0.81	Yes
MythoBiLLM vs LLM + LoRA	Entity consistency	+0.111	0.009	0.93	Yes
MythoBiLLM vs LLM + LoRA	Theme retention	+0.098	0.014	0.76	Yes
MythoBiLLM vs IndicBART	ROUGE-L	+0.089	0.011	0.84	Yes
MythoBiLLM vs Qwen3 + LoRA	Human score	+0.30	0.021	0.64	Yes

Significance was estimated using 1000 paired bootstrap resamples over the evaluated samples. Effect size is reported as Cohen's *d*.

5.11. Qualitative Output Analysis

Table 26 illustrates the error patterns represented by the quantitative measures. The baseline continuations remain fluent but may introduce an unsupported adviser, weaken the relation between a sage and a king, or omit an earlier vow. These errors preserve surface readability while changing the narrative structure. The MythoBiLLM examples retain the relevant characters, moral relation, and earlier narrative commitment more consistently. The examples therefore clarify why entity consistency, theme retention, coherence, and hallucination-control ratings improve more strongly than general fluency.

Table 26. Qualitative comparison of generated mythological continuations.

Input Context	Reference Continuation	Baseline Output	MythoBiLLM Output
A prince prepares to leave the city after accepting exile as a duty.	The continuation keeps the focus on duty, family sorrow, and acceptance of the father's command.	The output remains fluent but shifts quickly to a generic journey and introduces a new adviser not present in the input.	The output preserves the prince, family grief, and duty-based decision. It continues the exile theme without introducing unsupported characters.
A sage warns a king that pride may disturb the moral order of the kingdom.	The continuation develops the warning through a moral dialogue between royal power and spiritual discipline.	The output gives a broad moral lesson but weakens the sage-king relation and repeats similar sentences.	The output keeps the dialogue structure, preserves the sage's authority, and links the warning to dharma and restraint.
A warrior stands before battle while remembering a divine promise and earlier vows.	The continuation connects the battle scene with the earlier vow and the expected divine support.	The output describes battle action but misses the earlier vow and adds unrelated heroic claims.	The output connects the battle with the earlier vow, retains the divine promise, and continues the heroic tone with fewer unsupported additions.

5.12. Discussion

The performance pattern reflects complementary contributions rather than a single dominant mechanism. LoRA produces the largest initial gain by adapting the pretrained model to the lexical, stylistic, and discourse distribution of the mythological corpus. The BiLSTM adapter then introduces an ordered recurrent representation that differs from both transformer attention and a position-wise MLP. Its advantage over the parameter-matched MLP, longer-context LoRA, and RAG configurations is present but modest. The fusion gate allows this recurrent representation to modify the transformer state selectively rather than replacing it at every token.

The auxiliary objectives explain much of the difference between the BiLSTM-only configuration and full MythoBiLLM. Entity-aware training directly penalizes missing or unsupported characters, while style-aware training discourages shifts away from the source narrative style. The ablation results match these functions: entity supervision has its clearest effect on entity consistency and wrong-entity introduction, whereas style supervision has a

larger effect on style alignment and theme retention. The complete model performs best because these objectives act on different error types.

The comparison with retrieval clarifies that recurrent memory and external evidence solve different problems. RAG performs well when a relevant training passage supplies explicit character or theme information. The BiLSTM adapter instead learns ordered relationships from the current sequence and performs better on perplexity, MAUVE, and diversity. A hybrid design that combines retrieval with recurrent memory may therefore be more useful than treating the two mechanisms as competitors.

The MCS sensitivity analysis shows that the aggregate ranking is not produced only by entity consistency and style alignment. Full MythoBiLLM remains the highest-scoring tested system when either component is omitted and when both are removed together. At the same time, the small reversal between RAG and the BiLSTM adapter when narrative coherence is omitted confirms that the composite score depends on its component definition. MCS is therefore most useful as a compact diagnostic interpreted alongside task metrics, controlled comparisons, and human assessment.

The practical importance of the improvement is concentrated in domain-specific errors. Relative to LLM+LoRA, the wrong-entity rate decreases by 57.4%, while the largest robustness gains occur for character-heavy and moral-conflict prompts. These changes correspond to fewer unsupported character insertions, stronger preservation of moral and thematic relations, and better continuity between earlier commitments and later events. Such properties are relevant for educational summaries, digital-humanities analysis, corpus exploration, and assisted preparation of mythological content, where a fluent but incorrectly attributed event can be more harmful than a minor stylistic imperfection.

The model should nevertheless be used as an assistive system rather than an authoritative source. The experiments cover one compact backbone, English public-domain translations, a limited expert-annotated subset, and a finite set of mythological works. Historical variants, translation choices, regional traditions, and interpretive differences may not be represented fully. Expert verification remains necessary when generated material is used for scholarship, cultural interpretation, or educational publication.

6. Conclusions

MythoBiLLM combines a LoRA-adapted Llama 3.2 3B-Instruct backbone, a gated BiLSTM narrative-memory adapter, and mythology-aware auxiliary objectives for summarization and continuation generation. The experiments used 3,684,838 word-level tokens and 6057 passages derived from public-domain English translations of the Ramayana, Mahabharata, Bhagavad-Gita, Vishnupuranam, Harivamsha, Hindu Tales, and Indian Myth and Legend.

Within this corpus and experimental protocol, the complete configuration reduces average perplexity from 23.4 for LLM+LoRA to 19.8, obtains a summarization MCS of 0.762, and reaches a continuation-generation MCS of 0.744. It also achieves an entity F1 of 0.812 and reduces the wrong-entity rate from 0.122 to 0.052. Improvements in ROUGE-L, BERTScore, MAUVE, diversity, human ratings, and cross-dataset evaluation provide evidence beyond the composite MCS.

The MCS sensitivity analysis shows that the full configuration remains the highest-scoring tested system when individual components are removed. When entity consistency and style alignment are excluded together, it retains scores of 0.751 for summarization and 0.731 for continuation generation. This result reduces concern that the overall ranking is produced only by dimensions directly related to the auxiliary objectives. Nevertheless, MCS remains a diagnostic composite and is not treated as independent evidence by itself.

The controlled comparisons provide limited rather than decisive evidence for the BiLSTM mechanism. The BiLSTM adapter reaches an MCS of 0.713, compared with 0.699 for the parameter-matched MLP, 0.704 for long-context LoRA, and 0.708 for RAG. These differences indicate a modest complementary contribution under the tested conditions. The larger improvement of full MythoBiLLM cannot be attributed solely to the BiLSTM because the complete system also includes LoRA adaptation, gated fusion, and entity-, style-, and task-aware objectives.

The findings are limited to one compact LLM backbone, English public-domain translations, a relatively small expert-annotated subset, and the selected model and evaluation settings. They do not establish universal superiority of BiLSTM adapters over transformer-only, longer-context, feed-forward, or retrieval-based alternatives.

Future research should proceed in four specific directions. First, multilingual corpora should be constructed for Sanskrit, Hindi, Tamil, Bengali, and other Indian languages using source-level separation and expert verification of translated or parallel passages. Second, the same controlled protocol should be applied to larger open LLMs and additional parameter-matched adapters, including unidirectional recurrent, convolutional, state-space, and attention-based memory modules. Third, hybrid retrieval–memory systems should be tested to determine whether explicit retrieved evidence and recurrent sequence representations provide complementary gains on long documents. Fourth, the evaluation framework should be validated using independently constructed external corpora, larger blinded expert panels, and expert-derived or empirically calibrated MCS weighting schemes. These extensions would clarify the generality of the current findings and support more reliable use across languages, textual traditions, and narrative tasks.

Author Contributions: Conceptualization, S.B., S.Y. and S.F.A.R.; methodology, S.B.; software, S.B.; validation, S.B., S.Y. and S.F.A.R.; formal analysis, S.B. and S.Y.; investigation, S.B.; data curation, S.B.; writing—original draft preparation, S.B.; writing—review and editing, S.B., S.Y. and S.F.A.R.; visualization, S.B.; supervision, S.Y. and S.F.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived by the K.R. Mangalam University because the study involved anonymous evaluation of model-generated textual outputs and did not collect sensitive, clinical, intervention-related, or personally identifiable information.

Informed Consent Statement: Informed consent was obtained from all evaluators involved in the study. Before participation, each evaluator received the Participant Information Sheet and was informed about the study purpose, voluntary participation, the use of anonymized model outputs, the right to withdraw, and the non-collection of personally identifiable information.

Data Availability Statement: The data and supporting files presented in this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.20697625>. These data were derived from public-domain Indian mythological texts obtained through Project Gutenberg, as described in the manuscript.

Acknowledgments: The authors acknowledge Multimedia University, Melaka, Malaysia, for institutional support. The authors also acknowledge the publicly accessible source repositories and benchmark providers used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
2. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 248.
3. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8342–8360.
4. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022), Virtual Event, 25–29 April 2022.
5. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
6. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681.
7. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783.
8. Gemma Team. Gemma 3 Technical Report. *arXiv* **2025**, arXiv:2503.19786.
9. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388.
10. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
11. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), Virtual Event, 26 April–1 May 2020.
12. Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; Harchaoui, Z. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Online, 6–14 December 2021; Volume 34, pp. 4816–4828.
13. Gorman, R. Morphosyntactic annotation in literary stylometry. *Information* **2024**, *15*, 211.
14. He, X.; Lashkari, A.H.; Vombatkere, N.; Sharma, D.P. Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information* **2024**, *15*, 131.
15. Yang, L.; Wang, G.; Wang, H. Reimagining literary analysis: Utilizing artificial intelligence to classify modernist French poetry. *Information* **2024**, *15*, 70.
16. Vargas-Solar, G. Processing the Narrative: Innovative Graph Models and Queries for Textual Content Knowledge Extraction. *Electronics* **2024**, *13*, 3688.
17. Han, S.; Wang, M.; Zhang, J.; Li, D.; Duan, J. A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. *Electronics* **2024**, *13*, 5040.
18. Peykani, P.; Ramezanlou, F.; Tanasescu, C.; Ghanidel, S. Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. *Appl. Sci.* **2025**, *15*, 8103.
19. He, Y.; Zhu, X.; Li, D.; Wang, H. Enhancing large language models for specialized domains: A two-stage framework with parameter-sensitive LoRA fine-tuning and chain-of-thought RAG. *Electronics* **2025**, *14*, 1961.
20. Repede, S.E.; Brad, R. LLaMA 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. *Computers* **2024**, *13*, 292.
21. Abou Baker, N.; Rohrschneider, D.; Handmann, U. Parameter-efficient fine-tuning of large pretrained models for instance segmentation tasks. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2783–2807.
22. Kazlaris, I.; Antoniou, E.; Diamantaras, K.; Bratsas, C. From illusion to insight: A taxonomic survey of hallucination mitigation techniques in LLMs. *AI* **2025**, *6*, 260.
23. Dabre, R.; Shrotriya, H.; Kunchukuttan, A.; Puduppully, R.; Khapra, M.M.; Kumar, P. IndicBART: A pre-trained model for indic natural language generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 1849–1863.
24. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
25. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. Technical Report. OpenAI, 2019. Available online: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 20 July 2026).

26. Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonnell, K.; Phang, J.; et al. Gpt-neox-20b: An open-source autoregressive language model. In Proceedings of the BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models, Dublin, Ireland and Virtual Event, 27 May 2022; pp. 95–136.
27. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
28. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880.
29. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
30. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
31. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, W.B. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 110–119.
32. Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; Yu, Y. Texygen: A benchmarking platform for text generation models. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 1097–1100.
33. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1994.
34. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Routledge: New York, NY, USA, 2013.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.