Consolidating and Developing Benchmarking Datasets for the Nepali Natural Language Understanding Tasks

Anonymous ACL submission

Abstract

The Nepali language has distinct linguistic features, especially its complex script (Devanagari script), morphology, and various dialects, which pose a unique challenge for Natural Language Understanding (NLU) tasks. While the Nepali Language Understanding Evaluation (Nep-gLUE) benchmark provides a foundation for evaluating models, it remains limited in scope, covering four tasks. This restricts their utility for comprehensive assessments of Natural Language Processing (NLP) models. To address this limitation, we introduce twelve new datasets, creating a new benchmark, the Nepali Language Understanding Evaluation (NLUE) benchmark for evaluating the performance of models across a diverse set of Natural Language Understanding (NLU) tasks. The added tasks include Single-Sentence Classification, Similarity and Paraphrase Tasks, Natural Language Inference (NLI), and General Masked Evaluation Task (GMET). Through extensive experiments, we demonstrate that existing top models struggle with the added complexity of these tasks. We also find that the best multilingual model outperforms the best monolingual models across most tasks, highlighting the need for more robust solutions tailored to the Nepali language. This expanded benchmark sets a new standard for evaluating, comparing, and advancing models, contributing significantly to the broader goal of advancing NLP research for low-resource languages.

1 Introduction

Nepali is written in the Devanagari script and is a highly inflected language.The Nepali language incorporates a complex system of noun, adjective, and verb inflections, including gender, case, and number (Bal, 2004). It has a rich vocabulary with many homonyms and is spoken in different dialects across various regions, and there are variations in vocabulary, grammar, and pronunciation. Developing and establishing robust models for Nepali requires reliable methods to evaluate their quality and effectiveness and it is essential to have tools that can assess how well these models address the language's unique linguistic challenges while identifying their limitations.

Despite Nepalis importance as a primary or secondary language for millions of speakers, research efforts and resources dedicated to its computational processing and evaluation remain relatively sparse. Existing benchmarks, such as Nep-gLUE (Timilsina et al., 2022), have made significant progress in this direction, providing a foundation for evaluating models on fundamental tasks. However, these benchmarks are limited in scope, primarily addressing four basic tasks and overlooking critical aspects of linguistic understanding such as coreference resolution, paraphrase interpretation, and advanced inference capabilities. To address this need, we introduce a new benchmark comprising 12 Natural Language Understanding (NLU) tasks for Nepali. The tasks are grouped into four categories:

Single-Sentence Tasks: Sentiment Analysis (SA), Corpus of Linguistic Acceptability (CoLA), and WinoGrande (WG)

Similarity and Paraphrase Tasks: Quora Question Pairs (QQP), Microsoft Research Paraphrase Corpus (MRPC), Semantic Textual Similarity Benchmark (STS-B), and Query-Ad Matching (QADSM)

Natural Language Inference (NLI) Tasks: Multi-Genre NLI (MNLI), Question Answer NLI (QNLI), Recognizing Textual Entailment (RTE), and Coreference Resolution (CR)

General Masked Evaluation Task (GMET): A diagnostic task for testing factual and contextual understanding.

This suite includes a broader range of linguistic tasks, enabling more comprehensive evaluation of

Corpus	Train	Test	Task	Metrics Used	Domain
Single Sentence Tasks					
SA	65.1K	16.3K	Sentiment	Macro F1, Acc	Movie Reviews
CoLA	7.8K	1.95K	Acceptability	Macro F1, Acc	Books, Journal
WG	32.5K	8.14K	Commonsense Reasoning and Pronoun Coreference	Macro F1, Acc	Misc.
Similarity and Paraphrase Tasks					
QQP	26K	6.5K	Paraphrase	Macro F1, Acc	Social QA
MRPC	4.19K	1.05K	Paraphrase	Macro F1, Acc	News
STS-B	5.45K	1.36K	Sentence Similarity	Pearson Corr, Spearman Corr, R^2	News, Video Cap.
QADSM	59.4K	14.9K	Similarity	Macro F1, Acc	News
Natural Language Inference Tasks					
MNLI	40.8K	10.2K	NLI	Macro F1, Acc	Misc.
QNLI	28K	7K	QA/NLI	Macro F1, Acc	Wikipedia
RTE	2.01K	503	NLI	Macro F1, Acc	News, Wikipedia
CR	564	142	Coreference/NLI	Macro F1, Acc	Fiction Books
General Masked Evaluation Task					
GMET	-	1.5K	Mask Filling	Acc, Combined Score	Books, News

Table 1: Task descriptions and dataset statistics in the NLUE benchmark.

NLU capabilities for Nepali language models (Appendix A). Table 1 provides an overview of tasks, dataset sizes, evaluation metrics, and domains covered in the NLUE Benchmark.

The datasets in the NLUE Benchmark are inspired by the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) and XGLUE benchmark (Liang et al., 2020), and were developed through a combination of automated and manual processes to ensure high-quality task-specific datasets. Our contributions involved translating datasets with Large Language Models (LLMs), particularly GPT-4o-mini (OpenAI, 2024) and Gemini-2.5-flash (GeminiTeam, 2025), and ensuring the accuracy and contextual relevance of these translations (Appendix B & C). We also conducted a thorough review of the availability of existing Nepali datasets for each task. Where datasets were available, we integrated them with translated data, carefully eliminating duplicates to form a unified and comprehensive dataset. For tasks like Acceptability Judgments and Coreference Resolution, where suitable datasets or high-quality translations were unavailable, we performed manual translations to ensure linguistic accuracy and consistency. These efforts collectively ensure that the final dataset is robust, comprehensive, and reflective of the linguistic diversity in the Nepali language.

To assess the effectiveness of the NLUE benchmark and performance of models, we conducted experiments by fine-tuning both monolingual models trained exclusively on Nepali-language data and multilingual models that include Nepali as one of their supported languages. Each model was fine-tuned on tasks introduced in the NLUE Benchmark and evaluated using metrics provided in Table 1, providing a comprehensive understanding of their performance on various aspects of NLU.

2 Related Works

Benchmarks like GLUE (Wang et al., 2018) and its successor Super General Language Understanding Evaluation (SuperGLUE) benchmark (Wang et al., 2020) have been instrumental in advancing research in Natural Language Understanding (NLU). GLUE introduced a multi-task framework for evaluating diverse NLU capabilities, such as single-sentence classification, sentence-pair similarity, and inference tasks. SuperGLUE extended this with more challenging tasks, including causal reasoning and coreference resolution, addressing GLUE's limitations for state-of-the-art models. These benchmarks set a standard for evaluating linguistic and semantic understanding in high-resource languages like English, inspiring adapta-

tions in other languages and low-resource settings. Efforts like (Liang et al., 2020) and (Hu et al., 2020) expanded these concepts to multilingual contexts, enabling cross-lingual transfer learning.

Nep-gLUE (Timilsina et al., 2022) is the first comprehensive benchmark for Natural Language Understanding (NLU) tasks in Nepali. It includes four core tasks: Named Entity Recognition (NER), Part-of-Speech Tagging (POS), Content Classification (CC), and Categorical Pair Similarity (CPS). Although Nep-gLUE offers a robust foundation with its multi-task dataset, it falls short in addressing more advanced NLP tasks necessary for comprehensive evaluations of models at the linguistic level. The advanced and complex tasks are crucial for further progress in low-resource languages like Nepali.

Nepali Sentiment Analysis (NepSA) (Singh et al., 2020) is a targeted aspect-based sentiment analysis dataset, comprising 3,068 comments extracted from 37 YouTube videos across 9 channels. The dataset is annotated using a binary sentiment polarity schema across six aspects: General, Profanity, Violence, Feedback, Sarcasm, and Out-of-scope. Another dataset, Aspect-Based Sentiment Analysis (Tamrakar et al., 2020), contains 1,576 sentences, equally divided between positive and negative sentiments. Additionally, sentiment analysis datasets like Nepali Language Sentiment Analysis - Movie Reviews (Ghimire) with 602 data points, and Nepali Sentiment Analysis (Acharya) with 2,161 data points found on Kaggle, are limited in size and domain-specific. For our benchmark, we utilized the NepCOV19Tweets dataset (Sitaula et al., 2021), which includes ~33.5k sentiments labeled as positive, negative, or neutral. From this, we selected 14.9k positive and 13.5k negative data points for the SA dataset. A more recent dataset, Sentiment of Election-Based Nepali Tweets (Pokharel), contains ~17.8k tweets but includes English characters and numbers, making it less suitable for our benchmarked dataset. To our knowledge, there are no publicly available datasets for coreference resolution, acceptability judgment, paraphrase and similarity detection, commonsense reasoning, pronoun coreference resolution, general masked evaluation, or NLI in the Nepali language. Despite some studies focusing on Nepali grammar, the lack of datasets for these advanced tasks limits the development of comprehensive NLU benchmarks.

3 Model Selection

To evaluate the performance of Natural Language Processing (NLP) models on the Nepali Language Understanding Evaluation (NLUE) benchmark, we selected ten publicly available models that support devanagari script, carefully chosen to represent a diverse range of architectures, parameter sizes, and pretraining strategies including the state-of-the-art encoder model for language understanding and best monolingual models for the Nepali Language. Evaluating these models on the NLUE benchmark serves multiple purposes. First, it provides a comprehensive assessment of their capabilities across a diverse set of tasks. This enables us to identify which architectures and pretraining strategies are best suited for Nepali NLP, particularly for tasks that demand robust handling of the languages morphological complexity and dialectal variations. Second, comparing monolingual and multilingual models highlights the trade-offs between language-specific pretraining and cross-lingual generalization, offering insights into the optimal approach for low-resource languages. By identifying the strengths and weaknesses of existing models, this study informs the development of more robust solutions tailored to Nepalis unique linguistic challenges.

4 Tasks

NLUE is a benchmark designed to evaluate the performance of language understanding models across a diverse set of tasks, addressing the limitations of its predecessor, Nep-gLUE. The objective of NLUE is to provide a robust evaluation metric applicable to a broad range of language understanding challenges. We describe the tasks below and in Table 1.

4.1 Single-Sentence Tasks

Single-sentence tasks in the NLUE benchmark focus on assessing a model's ability to understand and analyze individual sentences. These tasks evaluate a model's ability to understand and interpret the meaning, sentiment, and grammatical structure of individual sentences.

4.1.1 Sentiment Analysis (SA)

The Sentiment Analysis dataset has been added to evaluate models' ability to classify the emotional tone (Positive & Negative) of Nepali text. We created the dataset for sentiment analysis by translat-

Model	Params	SA		CoLA		WG	
		Acc	F1	Acc	F1	Acc	F1
Distilbert-Nepali (Maskey et al., 2022)	67M	86.34	86.33	84.51	80.96	58.20	58.08
NepBERT (Rajan, 2021)	82M	83.34	83.34	80.51	74.80	52.49	52.04
NepaliBERT (Pudasaini et al., 2023)	110M	87.06	87.06	84.51	80.92	54.77	54.75
BERT Nepali (Thapa et al., 2025)	110M	87.73	87.72	84.76	80.65	66.81	66.13
NepBERTa (Timilsina et al., 2022)	110M	86.62	86.62	84.15	80.60	67.12	50.52
RoBERTa Nepali (Thapa et al., 2025)	125M	87.75	87.74	85.44	82.14	**68.07**	**68.07**
DeBERTa-Nepali (Maskey et al., 2022)	139M	87.43	87.42	85.08	81.86	59.76	59.75
Multilingual BERT (Devlin et al., 2019)	172M	86.35	86.34	82.41	78.95	67.12	50.52
XLM-R Base (Conneau et al., 2020)	270M	88.33	88.34	85.64	82.03	50.77	50.52
m-DeBERTa-v3 (He et al., 2023)	276M	**88.94**	**88.93**	**88.31**	**85.64**	67.45	67.44

Table 2: Model Performance across Single-Sentence Tasks

ing Stanford Sentiment Treebank (Socher et al., 2013) from the GLUE Benchmark, which includes ~53k sentence-level data points from movie reviews with human-annotated sentiment labels, using GPT-4o-mini, and manually translating instances that could not be accurately translated (Appendix B & C). We incorporated this dataset with pre-existing sentiment analysis of Nepali COVID-19-related tweets (Sitaula et al., 2021), adding ~28.4k data points. The combined SA dataset totals 81.4k data points, equally distributed between the positive and negative classes. Models are evaluated using Accuracy and Macro F1-score metrics, as reported in Table 1.

4.1.2 Corpus of Linguistic Acceptability (CoLA)

The Acceptability Judgments dataset determines whether a given sentence follows the linguistic rules of Nepali, ensuring the model can assess grammaticality. The dataset was created by translating the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) from the GLUE Benchmark, which includes 9.75k data points sourced from books and journal articles on linguistic theory, using GPT-4o-mini. Manual corrections were applied to sections where translations were inaccurate (Appendix B & C). The dataset is divided into correct and incorrect classes in a 70:30 ratio, respectively. Models are evaluated using Accuracy and Macro F1-score metrics, as reported in Table 1.

4.1.3 Wino-Grande (WG)

The WinoGrande dataset evaluates a models ability to perform commonsense reasoning by identifying the correct referent in a sentence with a blank referring to one of two candidate entities. The dataset for this benchmark is converted to Nepali by translating the xGLUE benchmarks WinoGrande dataset (Sakaguchi et al., 2019) using Gemini-2.5-flash, with manual corrections applied to ensure translation accuracy (Appendix B & C). The final dataset contains 40.7k data points, with each instance labeled to indicate the correct referent, and is equally split between both classes. The dataset preserves the original format and balance of the English version. Models are evaluated using Accuracy and Macro F1-score metrics, as reported in Table 1.

4.2 Similarity and Paraphrase Tasks

Similarity and Paraphrase Task in the NLUE benchmark evaluates a model's ability to determine whether two sentences convey the same meaning or are paraphrases of each other. By focusing on this aspect of language comprehension, these tasks provide valuable insights into a model's proficiency in handling diverse expressions of similar ideas.

4.2.1 Quora Question Pairs (QQP)

The QQP dataset tests whether the model can identify if pairs of questions from the community question-and-answer website Quora have similar meanings. The dataset was created by translating the Quora Question Pairs dataset (Iyer et al., 2017) from the GLUE Benchmark into Nepali using GPT-4o-mini and Gemini-2.5-flash, with manual corrections applied (Appendix B & C). The dataset contains 32.5k question pairs, labeled as similar or dissimilar, with a class distribution of 40% similar and 60% dissimilar. Models are assessed using accuracy and Macro F1-score metrics, as reported in Table 1.

Model	Params	QQP		MRPC		STS-B			QADSM	
		Acc	F1	Acc	F1	Sp. corr	Pr. corr	R^2	Acc	F1
Distilbert-Nepali (Maskey et al., 2022)	67M	81.63	81.07	80.52	77.95	84.57	83.13	71.13	65.74	65.63
NepBERT (Rajan, 2021)	82M	71.17	69.61	67.34	56.91	41.44	41.06	12.71	61.03	60.80
NepaliBERT (Pudasaini et al., 2023)	110M	77.37	76.67	66.86	58.49	75.91	73.76	57.62	63.27	63.18
BERT Nepali (Thapa et al., 2025)	110M	80.88	80.48	76.50	73.61	84.57	83.87	71.54	64.35	64.34
NepBERTa (Timilsina et al., 2022)	110M	81.83	81.57	80.42	78.67	87.54	86.44	76.53	63.71	63.64
RoBERTa Nepali (Thapa et al., 2025)	125M	81.15	80.60	79.47	75.25	87.75	86.32	76.68	65.02	65.00
DeBERTa-Nepali (Maskey et al., 2022)	139M	82.85	82.33	80.61	78.23	81.62	80.00	66.35	65.23	65.21
Multilingual BERT (Devlin et al., 2019)	172M	82.31	81.78	81.47	78.02	87.75	86.62	76.93	63.91	63.84
XLM-R Base (Conneau et al., 2020)	270M	83.06	82.68	82.71	80.59	87.68	86.79	76.77	63.70	63.70
m-DeBERTa-v3 (He et al., 2023)	276M	**84.34**	**83.82**	**83.48**	**81.93**	**90.22**	**89.57**	**81.33**	**66.42**	**66.42**

Table 3: Model Performance across Similarity and Paraphrase Tasks

4.2.2 Microsoft Paraphrase Research Corpus (MPRC)

We introduced the MRPC dataset, intending to identify whether the sentence pairs extracted from news articles are paraphrases of each other, based on the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005). Using GPT-4o-mini, we translated the MRPC dataset into Nepali with manual correction whenever needed (Appendix B & C). The dataset contains 5.23k sentence pairs, with the class distribution of 70-30, with a higher proportion of paraphrase pairs. We report the Accuracy and Macro F1 score, as shown in Table 1.

4.2.3 Semantic Textual Similarity Benchmark (STS-B)

The STS-B dataset measures a models proficiency in predicting the degree of semantic relatedness between pairs of sentences drawn from sources such as news headlines and video captions. Each pair is annotated with a similarity score on a continuous scale from 0 (no meaning overlap) to 5 (complete semantic equivalence), framing the task as a regression problem. The dataset was created by translating the STS-B dataset (Cer et al., 2017) from the GLUE Benchmark into Nepali using Gemini-2.5-flash, with manual corrections applied to ensure translation accuracy (Appendix B & C). The dataset contains 6.82k sentence pairs. We evaluate the model using Pearson correlation, Spearman correlation, and R^2 metrics, as reported in Table 1.

4.2.4 Query-Ad Matching (QADSM)

The QADSM dataset assesses a models capability to align the semantic meaning between queries and advertisements. The dataset was created by translating the QADSM dataset from the XGLUE Benchmark (Liang et al., 2020) into Nepali using Gemini-2.5-flash, with manual refinements to ensure linguistic precision (Appendix B & C). The dataset contains 74.3k data points, equally split between relevant and irrelevant classes, based on ad-query relevance. Models are evaluated using accuracy and Macro F1-score metrics, as reported in Table 1.

4.3 Inference Tasks

The NLI tasks in this benchmark assess a model's ability to understand relationships between sentences, such as entailment, contradiction, and neutral alignment. These tasks are crucial because they evaluate a model's comprehension of contextual meaning, logical inference, and its ability to handle complex linguistic structures, making them essential for advancing robust language understanding.

4.3.1 Multi-Genre NLI (MNLI)

The MNLI dataset tests a models capability to predict the relationship between sentence pairs, determining whether a premise entails, contradicts, or is unrelated to a hypothesis (neutral). The dataset was created by translating the Stanford Natural Language Inference Corpus (Bowman et al., 2015) from the GLUE Benchmark into Nepali using GPT-4o-mini and Gemini-2.5-flash, with manual intervention for precision (Appendix B & C). The dataset contains 51k sentence pairs, equally divided among entailment, contradiction, and neutral classes. We report accuracy and Macro F1-score, as described in Table 1.

4.3.2 Question-Answering NLI (QNLI)

The QNLI dataset evaluates a models capability to determine whether a context sentence contains the answer to a given question. The dataset has been adapted for Nepali from the GLUE benchmark by translating the original English dataset using GPT-4o-mini and Gemini-2.5-flash, with manual verification for accuracy (Appendix B & C), which originates from the Stanford Question

Model	Params	MNLI		QNLI		RTE		CR	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Distilbert-Nepali (Maskey et al., 2022)	67M	68.57	68.61	79.61	79.46	56.06	55.63	52.82	52.63
NepBERT (Rajan, 2021)	82M	51.93	51.09	60.04	59.96	53.88	53.85	55.63	39.70
NepaliBERT (Pudasaini et al., 2023)	110M	63.27	63.25	77.58	77.43	51.89	51.80	55.63	55.63
BERT Nepali (Thapa et al., 2025)	110M	71.80	71.92	81.26	81.22	53.28	53.27	59.15	51.63
NepBERTa (Timilsina et al., 2022)	110M	71.87	71.85	81.24	81.15	55.07	53.33	**58.52**	**57.33**
RoBERTa Nepali (Thapa et al., 2025)	125M	73.10	73.07	81.86	81.78	52.49	52.44	49.29	49.12
DeBERTa-Nepali (Maskey et al., 2022)	139M	74.01	74.01	82.64	82.64	53.88	53.04	50.70	50.67
Multilingual BERT (Devlin et al., 2019)	172M	71.60	71.85	83.47	83.46	**68.19**	**68.00**	47.89	47.38
XLM-R Base (Conneau et al., 2020)	270M	75.23	75.22	83.13	83.13	57.06	54.86	50.00	49.80
m-DeBERTa-v3 (He et al., 2023)	276M	**78.76**	**78.84**	**86.65**	**86.65**	57.85	57.80	46.48	32.84

Table 4: Model Performance across Inference Tasks

Answering Dataset (Rajpurkar et al., 2016) that contains question-paragraph pairs sourced from Wikipedia. The dataset contains 35k question-sentence pairs, equally split between entailment and non-entailment pairs, ensuring a balanced class distribution, and evaluated using accuracy and Macro F1-score metrics, as reported in Table 1.

4.3.3 Recognizing Textual Entailment (RTE)

The RTE dataset evaluates a model's ability to predict whether a hypothesis logically follows from a given premise. The dataset for this benchmark is converted to Nepali by translating the GLUE benchmarks RTE dataset, combined from RTE1 (Dagan et al., 2006), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009) using GPT-4o-mini, with manual corrections to maintain translation accuracy (Appendix B & C), containing 2.51k data points, equally distributed between two classes (entailment and non-entailment). We evaluate the model using Accuracy and Macro F1-score, as discussed in Table 1.

4.3.4 Coreference Resolution (CR)

This CR dataset tests the models ability to resolve coreference relationships within a Nepali text. We developed the coreference resolution dataset by manually translating the Winograd Schema Challenge (Levesque et al., 2011) from the GLUE Benchmark. The dataset has 706 data points, balanced between two classes, evaluated with Accuracy and Macro F1-score, as mentioned in Table 1.

4.4 General Masked Evaluation Task (GMET)

The General Masked Evaluation Task (GMET) dataset evaluates the zero-shot capabilities of language models in recognizing word relationships, understanding contextual nuances, and maintaining grammatical precision without fine-tuning. It serves as a benchmark for assessing logical reasoning and proficiency with complex linguistic constructs, critical for reliable language understanding across diverse scenarios. The GMET dataset comprises 1,500 authentic sentences from real-world contexts, ensuring ecological validity. These sentences are organized into 75 distinct categories, with 20 sentences per category, covering various topics and regional linguistic variations. Each sentence contains a missing word, challenging models to predict the appropriate word based on context, testing their inherent contextual understanding and language comprehension, particularly with nuanced expressions across communities. As the missing word may not always have a single correct answer, native speakers assisted in manual evaluations to ensure accurate and fair assessment.

Model	Params	Acc	C. Acc
Distilbert-Nepali (Maskey et al., 2022)	67M	51.47	42.84
NepBERT (Rajan, 2021)	82M	13.60	12.52
NepaliBERT (Pudasaini et al., 2023)	110M	44.53	37.99
BERT Nepali (Thapa et al., 2025)	110M	49.40	42.63
NepBERTa (Timilsina et al., 2022)	110M	46.40	39.89
RoBERTa Nepali (Thapa et al., 2025)	125M	**57.27**	**48.76**
DeBERTa-Nepali (Maskey et al., 2022)	139M	52.60	44.56
Multilingual BERT (Devlin et al., 2019)	172M	14.13	12.80
XLM-R Base (Conneau et al., 2020)	270M	53.27	44.75
m-DeBERTa-v3 (He et al., 2023)	276M	45.33	42.77

Table 5: Model Performance in GMET

Model performance on the GMET dataset is evaluated using two key metrics: overall accuracy and a combined score. Overall accuracy measures the proportion of correct predictions across all sentences, providing a straightforward performance indicator. The combined score integrates overall accuracy with an equality score, reflecting consistency across categories and penalizing uneven

performance to ensure balanced contextual understanding across diverse topics and linguistic variations. Further details are provided in Appendix F.

5 Experiments

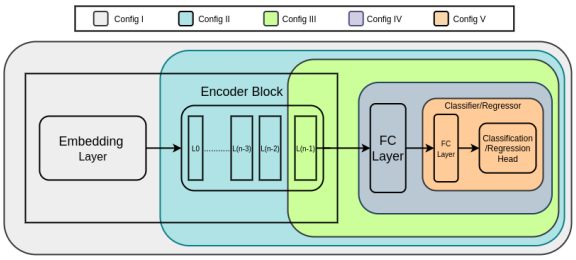

Figure 1: Different training config based on parameters with initial FC Layer

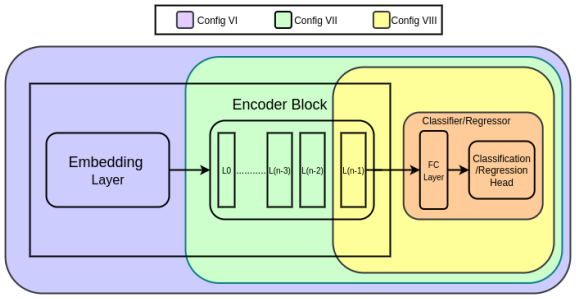

Figure 2: Different training config based on parameters without initial FC Layer

We experimented with eight distinct finetuning configurations (Configs IVIII), each controlling the subset of model parameters that are updated during training, as illustrated in Figure 1 and Figure 2. These configuration choices were driven by our available datasets to mitigate overfitting risks, which ranged widely from just under 1,000 to 80,000 data points. For larger datasets like QADSM and SA, training only the classification layer was insufficient, while for smaller datasets like CR and RTE, training all layers risked overfitting. Therefore, at least three configurations were tested per dataset to ensure robust performance comparisons and validate generalizability.

As part of our ablation study, we systematically examined how performance was affected by varying the number and type of layers updated during training, ranging from tuning only the top classification layer to progressively unfreezing intermediate and lower transformer layers. This analysis helped isolate the contributions of different layers to downstream performance. Specifically, we also experimented with Config IV both with and without the initial fully connected (FC) layer to assess its specific role in feature transformation.

Hyperparameter Search Space:

- Learning rate: $\{1e-5, 2e-5, 1e-4, 2e-4\}$
- Batch size: $\{8, 16, 32\}$
- Training epochs: Up to 15, with early stopping after three consecutive epochs of non-improving validation loss

For each configuration, we performed 5-fold cross-validation to select optimal hyperparameters and evaluate model performance. For each fold, we trained models with all hyperparameter combinations and selected the configuration that achieved the lowest validation loss. The best average hyperparameters across folds were used for final training. Test evaluation was done only after hyperparameter selection. Optimal hyperparameter settings and configuration for each dataset and model are reported in Appendix E.

6 Result and Analysis

We evaluate 10 language models on the NLUE benchmark across four task categories: Single-Sentence Classification, Similarity and Paraphrase Detection, Inference Tasks, and the GMET. For Classification Tasks, we report accuracy to measure overall correctness and macro-F1 to ensure balanced performance across potentially imbalanced classes. For the Regression Task, we use Spearman and Pearson correlation coefficients to assess monotonic and linear relationships, respectively, between predicted and actual continuous scores, and R^2 to quantify the proportion of variance in actual similarity scores explained by the models predictions. For GMET, we report a combined score, integrating overall accuracy with an equality score, to evaluate consistency across diverse categories.

Across Single Sentence tasks, m-DeBERTa-v3 achieves the highest overall scores, with an SA Macro-F1 of 88.93, WG Macro-F1 of 67.44, and CoLA Macro-F1 of 85.64. Among Nepali-specific models, RoBERTa-Nepali performs competitively in SA and WG, indicating that moderate-scale models can effectively handle single-sentence understanding in Nepali. Results reported in Table 2.

m-DeBERTa-v3 consistently achieves the highest scores across Similarity and Paraphrase Tasks with top Macro-F1 scores in QQP (83.82), MPRC (81.93), QADSM (66.42), and the highest correlation metrics in STS-B (90.22 Spearman, 89.57 Pearson). Among Nepali-specific models, DeBERTa-Nepali performs strongly on QQP and MRPC, while RoBERTa-Nepali shows better results on STS-B and QADSM. Overall, multilingual models dominate in performance, which suggests that semantic similarity detection in Nepali demands sophisticated representational capabilities beyond what current Nepali-specific models provide. Results reported in Table 3.

In Inference Tasks, m-DeBERTa-v3 achieves the strongest performance on MNLI (78.84 Macro-F1) and QNLI (86.65 Macro-F1), while Multilingual BERT achieves a strong 68 Macro-F1 on RTE, suggesting that multilingual pretraining enhances entailment and contradiction processing capabilities. However, their performance drops notably on the CR task, with no model exceeding 59.15% Accuracy (BERT Nepali), mainly due to its complexity in Nepali and limited dataset size (706 data points), which indicates that all models struggle with generalization from small datasets. Results reported in Table 4.

All evaluated models demonstrated suboptimal performance on the GMET dataset. These results indicate that zero-shot tasks in Nepali present significant challenges for current language models, even when processing straightforward conversational sentences. Notably, multilingual models and those with larger parameter counts failed to achieve superior performance compared to their monolingual counterparts. This performance gap may be attributed to tokenization limitations, as multilingual models typically contain fewer Devanagari tokens in their vocabularies relative to monolingual Nepali models. Results reported in Table 5.

7 Conclusion

The NLUE benchmark reveals distinct performance trends across models and tasks, with model size correlating strongly with performance. Larger models with multilingual pretraining, such as m-deberta-v3 (276M parameters) and XLM-r-base (270M parameters), consistently outperform smaller Nepali-specific models, particularly in tasks requiring nuanced semantic understanding (e.g., STS-B, QNLI). However, RoBERTa-Nepali (125M parameters) achieves competitive results despite its smaller size, suggesting that quality pretraining can outweigh parameter count.

Tasks like RTE and CR remain challenging, due to smaller dataset sizes, highlighting the need for enhanced datasets and improved modeling of Nepali textual entailment and coreference resolution. These results underscore the potential of multilingual models for low-resource languages like Nepali, while also revealing the importance of better Nepali-specific models to address language-specific challenges. Future work should prioritize creating larger, more diverse Nepali datasets and exploring techniques like cross-lingual transfer to enhance model robustness. The NLUE benchmark provides a valuable framework for evaluating and improving language models, paving the way for advancements in Nepali NLP.

8 Limitations

While the Nepali Language Understanding Evaluation (NLUE) benchmark significantly advances the evaluation of Natural Language Processing models for the Nepali Language, several limitations must be acknowledged to contextualize the findings and guide future research.

First, the datasets introduced in the NLUE benchmark were primarily created by translating existing English-language datasets from benchmarks such as GLUE and xGLUE, using automated tools like GPT-4o-mini and Gemini2.5-flash, supplemented by manual corrections. Although efforts were made to ensure translation accuracy, subtle linguistic nuances, cultural contexts, and idiomatic expressions specific to Nepali may not have been fully captured. The small size of certain datasets (e.g., CR and RTE) limits model performance and shows models' lack of generalization in smaller datasets. Second, the study evaluates a range of models with varying parameter sizes (67M to 276M), but resource constraints prevented the inclusion of larger, state-of-the-art models or extensive hyperparameter tuning. Finally, the reliance on specific evaluation metrics (e.g., accuracy, Macro-F1 score, Spearman, and Pearson correlations) may not fully capture the models performance across all dimensions of language understanding. For example, the GMET task relies on manual evaluations by native speakers, which might introduce subjectivity and potential inconsistencies.

References

Mahesh Acharya. Nepali language sentiment analysis.

Bal Krishna Bal. 2004. *Structure of Nepali Grammar*.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Text Analysis Conference (TAC)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Preprint*, arXiv:1508.05326.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

GeminiTeam. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Shikhar Ghimire. Nepali language sentiment analysis - movie reviews.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B. Dolan. 2007. The third pascal recognizing textual entailment challenge. In *ACL-PASCAL@ACL*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. *QuoraData*.

Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for Nepali text classification. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 106–111, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

Durga Pokharel. Sentiment of election based nepali tweets.

Shushanta Pudasaini, Subarna Shakya, Aakash Tamang, Sajjan Adhikari, Sunil Thapa, and Sagar Lamichhane. 2023. Nepalibert: Pre-training of masked language model in nepali corpus. In *7th International Conference on IoT in Social, Mobile, Analytics and Cloud*.

Rajan. 2021. Nepalibert.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of*

the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Chiranjibi Sitaula, Anish Basnet, Ashish Mainali, and Tej Shahi. 2021. Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets. *Computational Intelligence and Neuroscience*, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Sujan Tamrakar, Bal Krishna Bal, and Rajendra Thapa. 2020. *Aspect Based Sentiment Analysis of Nepali Text Using Support Vector Machine and Naive Bayes*. Ph.D. thesis.

Prajwal Thapa, Jinu Nyachhyon, Mridul Sharma, and Bal Krishna Bal. 2025. Development of pre-trained transformer-based models for the Nepali language. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 9–16, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. NepBERTa: Nepali language model trained in a large corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–284, Online only. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

A Dataset Description

A.1 Sentiment Analysis (SA)

To evaluate the sentiment understanding capabilities of language models, we developed a sentiment analysis dataset by combining existing Nepali sentiment datasets and translating several high-quality examples from English to Nepali.

{
text': 'त्यो दुर्लभ नाटक जसले सबैले अनुभव गरेको वा भविष्यमा अनुभव गर्ने प्रकारको पीडा बारे विचारशील र इनामस्वरूप झलक प्रदान गर्छ।',
'label': 1
}

{
'text': 'कोभिड र घरबन्दी बन्दाबन्दीका कारण नेपालीहरूमा पारेको असरबारे हाम्रो अर्को रिपोर्ट लगातार तीन महिना लामो घरबन्दीका कारण एक तिहाइ नेपालीहरूको आयआर्जन गुमेको र खाद्य संकटमा परेको अध्ययनले देखाएको छ।',
'label': 0
}

Figure 3: SA Positive (1) and Negative (0) Sample

Sentiment analysis requires models to grasp not just the literal meaning of words but also their emotional undertones and contextual implications. It is particularly challenging in Nepali, where sentiment is often conveyed through subtle linguistic cues and cultural context that may not be immediately apparent. Using this, we can better understand whether models truly comprehend the nuanced ways emotions are expressed in Nepali text, rather than simply memorizing surface-level patterns. This evaluation helps us gauge how well these models might perform on real-world applications involving subjective content analysis.

A.2 Corpus of Linguistic Acceptability (CoLA)

The Corpus of Linguistic Acceptability (CoLA) is a dataset originally developed for the GLUE benchmark to assess a models ability to judge the grammatical acceptability of English sentences. We incorporated CoLA-Nepali into our evaluation suite because understanding grammatical structure is fundamental to language comprehension. Unlike

other benchmarks that primarily test semantic understanding or task performance, CoLA directly probes whether models have internalized the syntactic rules that govern sentence formation.

{
'text': 'त्यहाँ एक अग्लो, रातो कपाल भएको र अविश्वसनीय रूपमा राम्रो लुगा लगाएको मानिस आइपुग्यो।',
'label': 1
}

{
'text': 'जति धेरै किताबहरू म सोध्छु कसलाई उसले दिनेछ, त्यति धेरै ऊ पढ्छ।',
'label': 0
}

Figure 4: CoLA Positive (1) and Negative (0) Sample

This evaluation is especially crucial for Nepali, where limited training data often contains grammatical inconsistencies or errors. By testing linguistic acceptability, we can determine whether our models have learned to distinguish well-formed Nepali sentences from those that violate grammatical constraints. This provides valuable insight into how deeply the models understand Nepali's structural patterns, beyond their ability to perform specific NLP tasks.

A.3 WinoGrande (WG)

The Winogrande dataset is a large-scale collection of Winograd Schema Challenge-style problems designed to evaluate commonsense reasoning and contextual disambiguation in natural language understanding.

{
'sentence': 'हामी गोठको भित्रपट्टि काम गर्नुको सट्टा घरको भित्रपट्टि काम गर्यौं, किनभने _ बस्नयोग्य थिएन।',
'option1': 'घर',
'option2': 'गोठ',
'answer': 2
}

Figure 5: WG Sample

We translated & added the Winogrande dataset into our evaluation benchmark to assess models' ability to perform commonsense reasoning and resolve linguistic ambiguities. The datasets adversarial construction reduces reliance on superficial statistical patterns, ensuring models rely on deep semantic and commonsense reasoning, which is essential for real-world applications like dialogue systems or conversational agents in Nepali. Including Winogrande in the Nepali, benchmark allows us to evaluate model strengths and limitations in handling nuanced linguistic structures and cultural contexts specific to Nepali, thereby improving their robustness for practical, context-sensitive applications.

A.4 Quora Question Pairs (QQP)

This dataset consists of pairs of questions that are labeled as either paraphrases (semantically equivalent) or not.

{
'question1': 'स्टार्टअपका लागि ब्रान्डको बारेमा सबैभन्दा राम्रो पुस्तक के हो?',
'question2': 'स्टार्टअप ब्रान्डिङको लागि कुन पुस्तक सर्वोत्तम हो?',
'label': 1
}

{
'question1': 'दैनिक जीवनको मनोविज्ञान: जीवनको बारेमा सिक्न र स्वीकृत गर्न सबैभन्दा कठिन कुरा के हो?',
'question2': 'तपाईंले आफ्नो जीवनमा गरेका सबैभन्दा कठिन कुरा के हो?',
'label': 0
}

Figure 6: QQP Positive (1) and Negative (0) Sample

We incorporated QQP into our benchmark to evaluate paraphrase detection capabilities, which are fundamental for robust language understanding systems. This task is particularly challenging in Nepali due to its morphological richness and limited resources. The dataset allows us to examine whether models can identify semantic equivalence beyond surface-level token matching or basic lexical similarity.

A.5 Microsoft Research Paraphrase Corpus (MRPC)

The MRPC dataset consists of pairs of sentences extracted from news sources, labeled as either paraphrase (semantically equivalent) or not. It is widely used to evaluate a model's ability to detect semantic equivalence between sentence pairs, particularly in formal and factual text domains.

MRPC challenges models to identify nuanced semantic similarities beyond superficial word overlap, requiring a deep understanding of sentence structure and meaning in Nepali. Its inclusion in

the Nepali benchmark ensures robust evaluation of models ability to handle formal news text.

{
'sentence1': 'सुरक्षा विज्ञहरूले चेतावनी दिएका छन् कि एउटा नयाँ मास-मेलिङ वर्म इन्टरनेटभरि व्यापक रूपमा फैलिएको छ, कहिलेकाहीँ माइक्रोसफ्टका संस्थापकको इमेलको रूपमा प्रस्तुत हुँदै।',
'sentence2': 'एक नयाँ वर्म इन्टरनेटभरि तीव्र गतिमा फैलिरहेको छ, कहिलेकाहीँ माइक्रोसफ्टका अध्यक्ष बिल गेट्सको इमेल भएको नाटक गर्दै, एन्टिभाइरस विक्रेताहरूले सोमबार भने।',
'label': 1
}

{
'sentence1': 'तर डेभिसका सल्लाहकार रोजर सालाजारले भने कि राज्यपालको ध्यान आफ्नो काममा छ, न कि फिर्ता बोलाउने निवेदनमा।',
'sentence2': 'तर डेभिसका सल्लाहकार रोजर सालाजारले भने कि राज्यपालको ध्यान उनले पारिश्रमिक पाएको काम गर्नेमा छ।',
'label': 0
}

Figure 7: MRPC Positive (1) and Negative (0) Sample

A.6 Semantic Textual Similarity Benchmark (STS-B)

The STS-B dataset consists of pairs of sentences annotated with a similarity score that reflects their semantic closeness on a continuous scale, typically from 0 (completely dissimilar) to 5 (semantically equivalent).

{
'sentence1': 'होभान बोलेनन्, तर उनका वकिल, जोन स्पेरान्जाले भने कि उनका ग्राहक "त्यो दिन कसैलाई चोट पुर्याउने उद्देश्यले उठेका थिएनन्"।',
'sentence2': 'होभान "त्यो दिन कसैलाई चोट पुर्याउने उद्देश्यले उठेका थिएनन्," प्रतिरक्षा वकिल जोन स्पेरान्जाले भने।',
'label': 4.5
}

{
'sentence1': 'सिरियामा विश्वसनीयता खतरामा',
'sentence2': 'ओबामा: कांग्रेस, विश्वको विश्वसनीयता खतरामा',
'label': 1.4
}

Figure 8: STS-B High (4.5) and Low (1.4) Similarity Sample

Unlike binary paraphrase detection tasks, the Semantic Textual Similarity Benchmark (STS-B) requires models to assess fine-grained semantic similarity between sentence pairs on a continuous scale. This makes STS-B a valuable benchmark for evaluating nuanced language understanding. This task challenges models to understand subtle differences and degrees of meaning overlap, which is essential for many real-world applications such as information retrieval, question answering, and summarization.

A.7 Query-Ad Matching (QADSM)

The QADSM dataset is incorporated into the NLUE benchmark to assess models ability to align semantic meaning between queries and advertisements in a binary classification task.

{
'query': 'सेतो पखेटा',
'ad_title': 'सेतो पखेटा: सस्तो',
'ad_description': 'सेतो पखेटाका लागि डिलहरू फेला पार्नुहोस्। तपाईंले किन्नु अघि उचित मूल्य हेर्नुहोस्!',
'relevance_label': 1
}

{
'query': 'स्कूल नर्स एड्स भिडियो',
'ad_title': 'स्कूल नर्स शिक्षा',
'ad_description': 'शैक्षिक पाठ्यक्रमहरूको बारेमा जानकारी प्राप्त गर्नुहोस्। स्कूल नर्स डिग्री कार्यक्रमहरू ब्राउज गर्नुहोस्',
'relevance_label': 0
}

Figure 9: QADSM Positive (1) and Negative (0) Sample

QADSM challenges models to discern semantic relevance beyond superficial keyword matching, requiring a detailed understanding of user intent and contextual meaning in Nepali. This is a critical capability for many real-world applications such as targeted advertising, search result optimization, and personalized content delivery in Nepali, where accurate query-ad alignment is critical.

A.8 Multi-Genre Natural Language Inference (MNLI)

The MNLI dataset consists of sentence pairs, each containing a premise and a hypothesis, labeled with one of three classes: entailment, contradiction, or neutral.

We include MNLI in our benchmark to evaluate a models ability to reason about the relationship between sentences across diverse domains. This task extends beyond surface-level similarity, requiring models to capture subtle semantic distinctions, such as entailment, contradiction, and neutrality, which are essential for applications like question answering, summarization, and dialogue systems.

{
'premise': 'गरिबीका कारण स्वास्थ्य सेवा प्रदायकहरूबाट गर्भपतनसम्बन्धी सेवाहरू प्राप्त गर्नबाट वञ्चित छन्।',
'hypothesis': 'गर्भपतन सेवाहरू गरिबीका कारण निषेधित छन्।',
'label': 0
}

{
'premise': 'उनीहरू यहाँ आउँदैछन्! कादान चिच्याए।',
'hypothesis': 'कादान इङ्गल्याण्डकी रानीको आगमनबारे चिच्याइरहेका थिए।',
'label': 1
}

{
'premise': 'सुपर-मूल्य भोजनको लागि ८० हङकङ डलर र म्याकडोनाल्ड्सले बेच्ने प्रत्येक स्नुपी खेलौनाको लागि ६ हङकङ डलर।',
'hypothesis': 'स्नुपी म्याकडोनाल्ड्सबाट खरिदका लागि उपलब्ध छैन।',
'label': 2
}

Figure 10: MNLI Entailment (0), Neutral (1) and Contradiction (2) Sample

A.9 Question-Answering Natural Language Inference (QNLI)

The QNLI dataset is a sentence pair classification benchmark designed to evaluate a models ability to perform natural language inference in the context of question answering.

{
'question': '१३ औं शताब्दीको मोजाइक सान सिप्रियानो चर्चबाट कसले किने?',
'sentence': '१९ औं शताब्दीमा चर्च भत्काइएपछि, मोजाइक प्रसियाका फ्रेडरिक विलियम चौथोले किनेका थिए।',
'label': 0
}

{
'question': 'कुन चाडमा मुख्यतया बालबालिकाहरू सहभागी हुन्छन्?',
'sentence': 'तेस्रो र सबैभन्दा ठूलो कार्निवलको अन्तिम दिनमा हुन्छ र यसमा सयौं मानिसहरू शहरको सबैभन्दा लामो मार्गमा पोशाक लगाएर हिंड्छन्।',
'label': 1
}

Figure 11: QNLI Entailment (0) and Non-Entailment (1) Sample

We include the QNLI dataset in our benchmark to evaluate a models ability to reason over questionanswer pairs. This task requires understanding the intent behind a question and determining whether a candidate sentence contains information that answers it, thereby testing the models grasp of both question semantics and contextual relevance.

A.10 Recognizing Textual Entailment (RTE)

The Recognizing Textual Entailment (RTE) dataset is a benchmark designed to evaluate a models ability to determine whether the meaning of one text fragment (the hypothesis) can be inferred from another text fragment (the text).

{
'sentence1': 'केसलरको टोलीले १४ देशका वयस्कहरूसँग ६०,६४३ वटा प्रत्यक्ष अन्तर्वार्ता लिएको थियो।',
'sentence2': 'केसलरको टोलीले १४ देशमा ६०,००० भन्दा बढी वयस्कहरूसँग अन्तर्वार्ता लिएको थियो।',
'label': 0
}

{
'sentence1': 'यदि कुनै मेक्सिकन सिमानामा पुग्छ भने, उसले अवैध रूपमा सिमाना पार गर्न खोजिरहेको मानिन्छ।',
'sentence2': 'मेक्सिकनहरूले अवैध रूपमा सिमाना पार गर्न जारी राखेका छन्।',
'label': 1
}

Figure 12: RTE Entailment (0) and Non-Entailment (1) Sample

Including RTE in a Nepali benchmark is important because entailment recognition is a core aspect of natural language understanding, especially in low-resource settings where explicit reasoning and semantic alignment are critical. It helps assess whether models trained on Nepali data can capture subtle logical relationships.

A.11 Co-reference Resolution (CR)

Co-reference resolution is the task of identifying when two or more expressions in a text refer to the same entity. This is essential for understanding the meaning of a passage, as natural language often relies on pronouns and noun phrases that depend on previous context.

We include the co-reference resolution dataset in our evaluation benchmark to assess a models ability to understand and maintain coherence across sentences. In the context of the Nepali language, this task is particularly challenging due to the flexible and context-sensitive nature of referential expressions shaped by discourse. Evaluating models on this task allows us to probe their understanding of entity continuity, pronoun grounding, and broader contextual reasoning.

{
'sentence1': 'प्रहरीले सबै गिरोहका सदस्यहरूलाई गिरफ्तार गऱ्यो। उनीहरू छिमेकमा लागूऔषध व्यापार रोक्न प्रयास गर्दै थिए।',
'sentence2': 'प्रहरीले छिमेकमा लागूऔषध व्यापार रोक्न प्रयास गर्दै थिए।',
'label': 1
}

{
'sentence1': 'टेबल ढोकाबाट अटाउँदैन किनभने यो धेरै साँघुरो छ।',
'sentence2': 'टेबल धेरै साँघुरो छ।',
'label': 0
}

Figure 13: CR Positive (1) and Negative (0) Sample

A.12 General Masked Evaluation Task (GMET)

We developed the General Masked Evaluation Task (GMET) dataset, and it is designed to test whether a model understands context. As the task is to predict masked tokens, we test our models on this task without fine-tuning. Given a mask, any word or phrase could plausibly fit the blank depending on the context, so the model must deeply understand the meaning and structure of the sentence to make an accurate prediction. Including GMET in the benchmark is important because it evaluates general language modeling capabilities, such as contextual comprehension, lexical choice, and syntactic fluency skills that are essential for strong performance across a wide range of downstream tasks in Nepali.

{
'sentence': 'म भान्सामा चिया पकाउँदै गर्दा अचानक ग्यास [MASK]।'
}

Figure 14: GMET Sample

B Dataset Translation Approach

Given the unfunded nature of this research, we relied on personal resources and utilized the APIs of two large language models, GPT-4o-mini and Gemini2.5-flash, to translate datasets into Nepali. We processed data in batches of 50 to 100 rows, each containing text and its corresponding label, using automated scripts to manage batching, API interactions, and output collection. For tasks requiring nuanced understanding, such as Co-reference Resolution, manual translation and review were employed to ensure accuracy.

B.1 Translation Problems

During the translation process, we encountered several challenges:

B.1.1 Label Mismatch

Despite instructions to preserve labels, models produced correct Nepali translations that differed in meaning from the English source, resulting in label mismatches. In some cases, models also corrected errors in the original English, requiring manual review to ensure consistency. Examples are provided in Figure 15, Figure 16 and Figure 17.

{
"English Sentence": "What did Bill buy potatoes and ?",
"EN Label": 0
"Nepali Sentence": "बिलले आलु र के किन्यो?",
"NE label": 1
}

Figure 15: Incomplete coordination in English becomes a well-formed question in Nepali, causing label mismatch.

{
"English Sentence": "Who do you think that will question Seamus first?",
"EN Label": 0
"Nepali Sentence": "तपाईंलाई के लाग्छ कसले सिमसलाई पहिले प्रश्न गर्नेछ?",
"NE label": 1
}

Figure 16: English ungrammaticality from complementizer-trace is absent in Nepali, leading to label mismatch.

{
"English Sentence": "many evidence was provided.",
"EN Label": 0
"Nepali Sentence": "धेरै प्रमाणहरू प्रदान गरिएको थियो।",
"NE label": 1
}

Figure 17: Plural-subject agreement error in English is resolved in Nepali, resulting in label mismatch.

B.1.2 Literal Translations

Some translations were overly literal, failing to capture contextual nuances. This issue arises when translation models prioritize word-by-word equivalence rather than interpreting the sentence as a whole. As a result, idiomatic expressions, culturally specific phrases, or context-dependent mean-

ings are mistranslated, leading to loss of intended meaning. Examples are provided in Figure 18, Figure 19, Figure 20 and Figure 21.


```
{
 "English Sentence": "on all cylinders",
 "Nepali Sentence": "सबै सिलिन्डरहरूमा",
}
```


Figure 18: English idiom on all cylinders meaning working perfectly or at full capacity becomes too literal in Nepali translation, losing its sentiment.


```
{
"English Sentence": "hide new secretions from the parental units",
"Nepali Sentence": "आमाबाबुको इकाईहरूबाट नयाँ स्रावहरू लुकाउनुहोस्",
}
```


Figure 19: Teen slang that means hiding new secrets from parents, when translated, talks about bodily fluid, losing the context.


```
{
 "English Sentence": "of saucy",
 "Nepali Sentence": "चटनी को",
}
```


Figure 20: Wordplay lost in translation, English slang becomes nonsensical in Nepali.


```
{
 "English Sentence": "a doa",
 "Nepali Sentence": "एक डीओए",
}
```


Figure 21: English slang phrase transliterated in Nepali, losing its negative sentiment.

B.1.3 Temporal Expression Mismatch

Abbreviated years in English, when translated literally into Nepali numerals, are often misinterpreted as regular numbers rather than references to specific years. This results in a loss of temporal context, which is especially problematic in historical or review texts where accurate time representation is crucial. Such misinterpretations can alter the meaning of the text and reduce the effectiveness of models trained on this data. Careful handling of these expressions is necessary to preserve the intended temporal information in Nepali translations. Examples are provided in Figure 22.


```
{
"English Sentence": “stale retread of the '53 original",
"Nepali Sentence": "५३ मूल संस्करणको बासी पुनरावृत्ति",
}

{
"English Sentence": “(like Saturday morning TV in the '60s)",
"Nepali Sentence": "(६० को शनिवार बिहानको टिभी जस्तै)",
}
```


Figure 22: Examples of temporal expression mismatches due to literal translation of abbreviated years.

B.1.4 Named Entities and Cultural References

Inconsistent translations of named entities and cultural references often disrupted the semantic integrity of the text and required manual corrections to maintain relevance and consistency within the Nepali context. These inconsistencies, if left uncorrected, could mislead models during training or evaluation. Examples are provided in Figure 23, Figure 24 and Figure 25.


```
{
"Premise": "The Old One always comforted Ca'daan, except today.",
"Hypothesis": "Ca'daan knew the Old One very well."
}

{
"Premise": "पुरानो व्यक्तिले सधैं का'दानलाई सान्त्वना दिन्थे, आज बाहेक।",
"Hypothesis": "का'दानले पुरानो व्यक्तिलाई राम्रोसँग चिन्थे।"
}
```


Figure 23: Example of cultural reference mismatch: mythological connotation of "The Old One" is weakened in Nepali translation.


```
{
 "Sentence1": "A plane is taking off.",
 "Sentence2": "An air plane is taking off."
}

{
 "Sentence1": "विमान उडिरहेको छ।",
 "Sentence2": "हवाई जहाज उडिरहेको छ।"
}
```


Figure 24: Example of lexical ambiguity in named entities: English terms like "plane" and "airplane" in Nepali have a subtle semantic distinction.

{
"English Sentence": "Yet another entry in the sentimental oh-those-wacky-Brits genre that was ushered in by The Full Monty and is still straining to produce another smash hit.",
"Nepali Sentence": "फुल मोंटकै द्वारा सुरु गरिएका भावुक ओ-त्यो-पागल-ब्रिटिशहरूको शैलीमा अझ एउटा प्रवेश, र अझै अर्को ठूलो सफलताको उत्पादन गर्न संघर्ष गर्दैछ।",
}

Figure 25: Example of idiomatic mismatch: oh-those-wacky-Brits referring to eccentric British cultural traits becomes oh-those-crazy-British-people, sounding awkward or negative in Nepali.

B.2 Suggestions on Translation

To improve translation quality and accuracy, we recommend the following strategies:

- **Use detailed prompts:** Instruct models to translate into clear, natural Nepali while preserving the original meaning and sentence structure.
- **Handle untranslatable terms:** For words or phrases without direct Nepali equivalents, allow romanization as a fallback strategy.
- **Batch size optimization:** Process **50–100 examples per API request** (assuming each has 80-100 tokens) to balance efficiency with translation quality. Avoid exceeding 100 examples to prevent degradation.
- **Class-wise translation:** Translate examples belonging to the same class in separate requests, and assign labels **after** translation to avoid mismatches due to grammatical differences across languages.
- **Ensure output consistency:** Implement automated checks to verify that the **number of translated outputs matches the input examples**, minimizing the risk of data loss during batch processing.

C Dataset Quality

C.1 Multilingual Content Filtering

We employed automated language detection to identify and remove any English or non-Nepali text remnants from the translated outputs. This filtering process ensures the purity of the translation by flagging code-switching instances, incomplete translations, or processing errors that leave the artifacts of the source language.

C.2 Statistical Quality Sampling with Manual Validation

A random sampling approach was used to select 1% of the translated corpus for manual quality assessment by native Nepali speakers. Each sampled translation was evaluated using standardized rubrics covering adequacy, fluency, and cultural appropriateness. It was decided that if more than 10% of the samples were found to be unacceptable, retranslation would be performed with an updated prompt. However, no such cases were encountered, indicating high confidence in the translation quality achieved using GPT-4o-mini and Gemini-2.5-flash for Nepali language translation.

C.3 Bidirectional Translation Validation (Back-translation)

Back-translation validation was performed on a randomly sampled 1% subset by translating Nepali outputs back to English using a different translation system. Semantic preservation was measured through automated similarity metrics, including BLEU scores between original and back-translated English texts. Some divergences indicated potential quality issues such as semantic drift or ambiguity resolution errors in the forward translations, but no significant concerns were seen.

D Experiment Stats

We utilized 1,200 GPU hours on NVIDIA T4 GPUs for our experiments. This includes fine-tuning 10 distinct model variants on 11 benchmark datasets, on the configurations outlined in the experiments (section 5).

E Hyperparameter Settings

This section details the optimal hyperparameters identified for each model and dataset combination in our benchmark evaluation for reproducibility. Tuning config is written in the following order: Model Config, Learning Rate, Epoch, Batch Size.

For model config, see Figure 1 and Figure 2.

E.1 Single Sentence Tasks

Best hyperparameter settings (model config, learning rate, epochs, batch size) for each model on SA (Sentiment Analysis), CoLA (Corpus of Linguistic Acceptability), and WG (Wino-Grande) tasks are reported in Table 6.

Model	Params	SA	CoLA	WG
Distilbert-Nepali (Maskey et al., 2022)	67M	I, 2e-5, 3, 16	I, 2e-5, 5, 16	II, 2e-5, 8, 32
NepBERT (Rajan, 2021)	82M	I, 2e-5, 2, 16	I, 2e-5, 8, 16	II, 2e-5, 4, 32
NepaliBERT (Pudasaini et al., 2023)	110M	I, 2e-5, 2, 16	I, 2e-5, 8, 16	II, 2e-5, 5, 32
BERT Nepali (Thapa et al., 2025)	110M	I, 2e-5, 2, 16	I, 2e-5, 10, 16	II, 2e-5, 10, 32
NepBERTa (Timilsina et al., 2022)	110M	I, 2e-5, 2, 16	I, 2e-5, 8, 16	I, 2e-5, 9, 32
RoBERTa Nepali (Thapa et al., 2025)	125M	I, 2e-5, 2, 16	I, 2e-5, 9, 16	I, 2e-5, 9, 32
DeBERTa-Nepali (Maskey et al., 2022)	139M	I, 2e-5, 3, 16	I, 2e-5, 5, 16	II, 2e-5, 8, 32
Multilingual BERT (Devlin et al., 2019)	172M	I, 2e-5, 3, 16	I, 2e-5, 10, 16	II, 2e-5, 10, 32
XLM-R Base (Conneau et al., 2020)	270M	I, 2e-5, 3, 16	I, 2e-5, 5, 16	I, 2e-5, 8, 32
m-DeBERTa-v3 (He et al., 2023)	276M	I, 2e-5, 3, 16	I, 2e-5, 5, 16	I, 2e-5, 8, 32

Table 6: Best hyperparameter settings for Single Sentence Tasks.

Model	Params	QQP	MRPC	STS-B	QADSM
Distilbert-Nepali (Maskey et al., 2022)	67M	VI, 2e-5, 3, 16	VI, 2e-5, 4, 16	I, 2e-5, 15, 16	VI, 2e-5, 4, 16
NepBERT (Rajan, 2021)	82M	VI, 2e-5, 2, 16	VI, 2e-5, 3, 16	I, 2e-5, 15, 16	VI, 2e-5, 3, 16
NepaliBERT (Pudasaini et al., 2023)	110M	VI, 2e-5, 4, 16	VII, 2e-5, 5, 16	I, 2e-5, 14, 16	VI, 2e-5, 5, 16
BERT Nepali (Thapa et al., 2025)	110M	VI, 2e-5, 2, 16	VII, 2e-5, 5, 16	I, 2e-5, 15, 16	VI, 2e-5, 5, 16
NepBERTa (Timilsina et al., 2022)	110M	VI, 2e-5, 4, 16	VII, 2e-5, 4, 16	I, 2e-5, 15, 16	VI, 2e-5, 3, 32
RoBERTa Nepali (Thapa et al., 2025)	125M	VI, 2e-5, 2, 16	VI, 2e-5, 6, 16	II, 2e-5, 12, 8	VI, 2e-5, 3, 32
DeBERTa-Nepali (Maskey et al., 2022)	139M	VI, 2e-5, 2, 16	VII, 2e-5, 4, 16	I, 2e-5, 5, 16	VI, 2e-5, 5, 16
Multilingual BERT (Devlin et al., 2019)	172M	VI, 2e-5, 2, 16	VII, 2e-5, 3, 16	I, 2e-5, 5, 16	VI, 2e-5, 5, 16
XLM-R Base (Conneau et al., 2020)	270M	VI, 2e-5, 2, 16	VI, 2e-5, 4, 16	I, 2e-5, 13, 16	VI, 2e-5, 3, 32
m-DeBERTa-v3 (He et al., 2023)	276M	VI, 2e-5, 3, 16	VII, 2e-5, 6, 16	I, 2e-5, 14, 16	VI, 2e-5, 3, 16

Table 7: Best hyperparameter settings for Similarity and Paraphrase Tasks.

Model	Params	MNLI	QNLI	RTE	CR
Distilbert-Nepali (Maskey et al., 2022)	67M	VI, 2e-5, 3, 16	VI, 2e-5, 2, 16	V, 2e-5, 10, 32	V, 1e-5, 4, 16
NepBERT (Rajan, 2021)	82M	VI, 2e-5, 4, 16	VI, 2e-5, 3, 16	V, 2e-5, 12, 32	V, 1e-5, 3, 16
NepaliBERT (Pudasaini et al., 2023)	110M	VI, 2e-5, 3, 16	VI, 2e-5, 2, 16	V, 2e-5, 9, 32	IV, 1e-5, 3, 16
BERT Nepali (Thapa et al., 2025)	110M	VI, 2e-5, 3, 16	VI, 2e-5, 2, 16	V, 2e-5, 11, 32	V, 1e-5, 2, 16
NepBERTa (Timilsina et al., 2022)	110M	VI, 2e-5, 3, 16	VI, 2e-5, 2, 16	V, 2e-5, 10, 32	V, 1e-5, 2, 32
RoBERTa Nepali (Thapa et al., 2025)	125M	VII, 2e-5, 7, 16	VI, 2e-5, 2, 16	V, 2e-5, 15, 32	V, 2e-5, 5, 32
DeBERTa-Nepali (Maskey et al., 2022)	139M	VII, 2e-5, 5, 16	VI, 2e-5, 4, 16	V, 2e-5, 10, 32	IV, 2e-5, 5, 32
Multilingual BERT (Devlin et al., 2019)	172M	VI, 2e-5, 3, 16	VI, 2e-5, 2, 16	V, 2e-5, 10, 32	IV, 2e-5, 3, 32
XLM-R Base (Conneau et al., 2020)	270M	VII, 2e-5, 5, 16	VI, 2e-5, 3, 16	V, 2e-5, 15, 32	IV, 2e-5, 3, 32
m-DeBERTa-v3 (He et al., 2023)	276M	VI, 2e-5, 3, 16	VI, 2e-5, 2, 16	IV, 2e-5, 12, 32	IV, 1e-5, 4, 8

Table 8: Best hyperparameter settings for Inference Tasks.

E.2 Similarity and Paraphrase Tasks

Best hyperparameter settings (model config, learning rate, epochs, batch size) for each model on QQP (Quora Question Pairs), MRPC (Microsoft Research Paraphrase Corpus), STS-B (Semantic Textual Similarity Benchmark), and QADSM (Query Ad Matching) tasks are reported in Table 7.

E.3 Inference Tasks

Best hyperparameter settings (model config, learning rate, epochs, batch size) for each model on MNLI (Multi-Genre Natural Language Inference), QNLI (Question-answering Natural Language Inference), RTE (Recognizing Textual Entailment), and CR (Co-reference Resolution) tasks are reported in Table 8.

F More on GMET

F.1 Dataset Categories

The GMET dataset is organized into the following 75 categories, grouped into seven thematic areas, presented here in English:

- **Daily Life & Home:** Family, House, Kitchen, Food, Clothing, Market, Shop, Daily Routine, Furniture, Health, Dream, Cleanliness, Medicine
- **Nature & Environment:** Weather, Animals, Birds, Insects, Fruits, Vegetables, Trees, Flowers, Nature, Water, River, Mountain, Forest, Sky, Earth, Ocean/Sea, Weather Conditions
- **Society & Culture:** School, Village, City, Sports, Festivals, Music, Art, Friendship, Society, Language, Nepali Culture, Movies, Books
- **Concepts & Knowledge:** Colors, Body Parts, Time, Numbers, Days of the Week, Months, Feelings, Shapes, Directions, Senses, Opposites, Geography, Science
- **Work & Activities:** Work, Professions, Agriculture, Learning, Cooking, Hobbies, Communication, Travel
- **Broader World:** Transportation, Money, History, Government, Technology, Space (sun, moon), Tools, Materials (wood, metal)
- **Nepal Specific:** Geography of Nepal, Animals of Nepal, Festivals of Nepal

F.2 Evaluation Metrics for GMET

The General Masked Evaluation Task (GMET) employs two primary metrics to assess language model performance: overall accuracy and a combined score. These metrics are formalized as follows:

F.2.1 Overall Accuracy

The overall accuracy is defined as the proportion of correct predictions across all sentences in the dataset. Given a dataset with N = 1500 sentences, where each prediction is scored as 1 (correct) or 0 (incorrect), let s_i represent the score for the i^{th} sentence. The overall accuracy A is calculated as:

$$A = \frac{1}{N} \sum_{i=1}^{N} s_i$$

F.2.2 Combined Accuracy

The combined score integrates the overall accuracy with an equality score that measures consistency across categories. The dataset is divided into K = 75 categories, with each category containing 20 sentences. For the k^{th} category, the category-wise accuracy A_k is computed as:

$$A_k = \frac{1}{20} \sum_{i=1}^{20} s_{i,k}$$

where $s_{i,k}$ is the score for the i^{th} sentence in the k^{th} category. The standard deviation of the category-wise accuracies, σ, is calculated as:

$$\sigma = \sqrt{\frac{1}{K} \sum_{i=1}^{K} (A_k - \bar{A})^2}$$

where $\bar{A}$ is the mean of the category-wise accuracies:

$$\bar{A} = \frac{1}{K} \sum_{i=1}^{K} A_k$$

The equality score, E, is derived by transforming the standard deviation to map lower deviation values to higher scores, with the score ranging between 0 and 1. The equality score is defined as:

$$E = e^{-\sigma}$$

This function ensures that lower standard deviations (indicating more consistent performance across categories) yield higher equality scores. In the edge case of a single category, where σ is undefined, E is set to 1. The combined score, C, is then

computed as the product of the overall accuracy and the equality score:

$$C = A.E$$

This combined score balances overall performance with consistency, penalizing models that exhibit uneven performance across the diverse linguistic and topical categories of the GMET dataset.

G Performance Visualization of Individual Models on each task

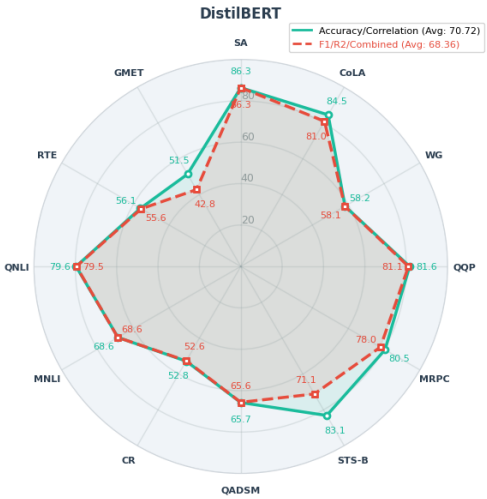

Figure 26: DistillBERT across all tasks

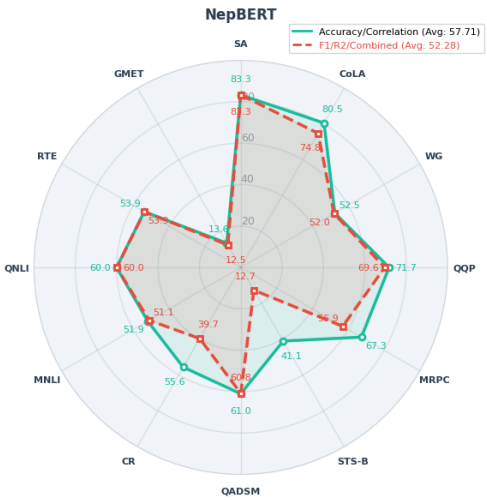

Figure 27: NepBERT across all tasks

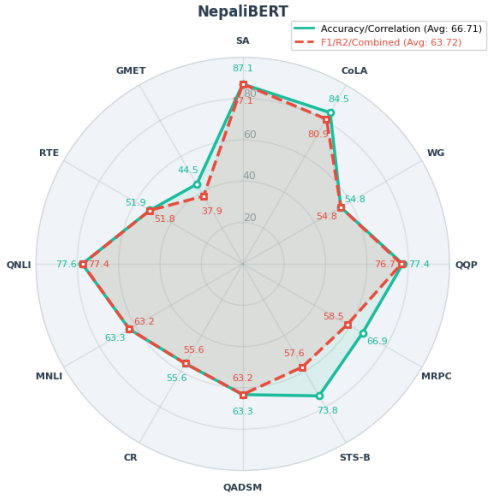

Figure 28: NepaliBERT across all tasks

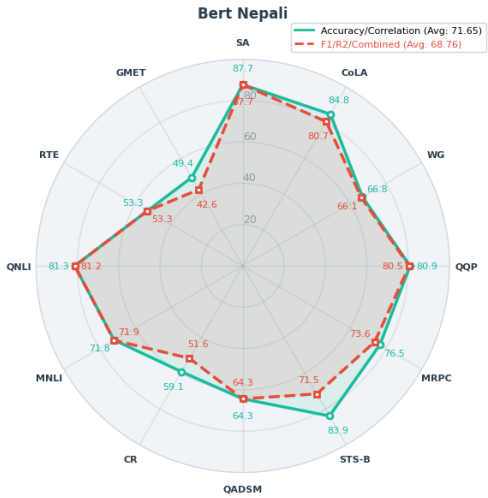

Figure 29: BERT Nepali across all tasks

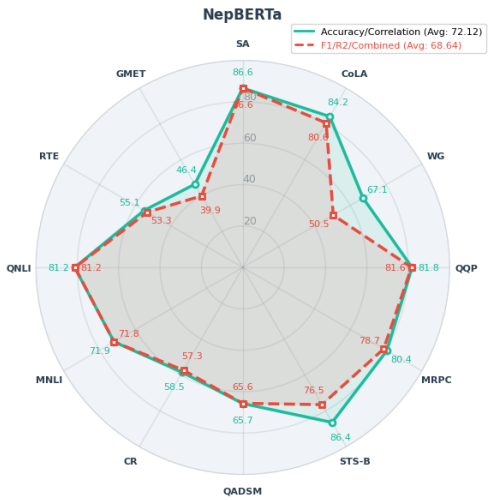

Figure 30: NepBERTa acrosss all tasks

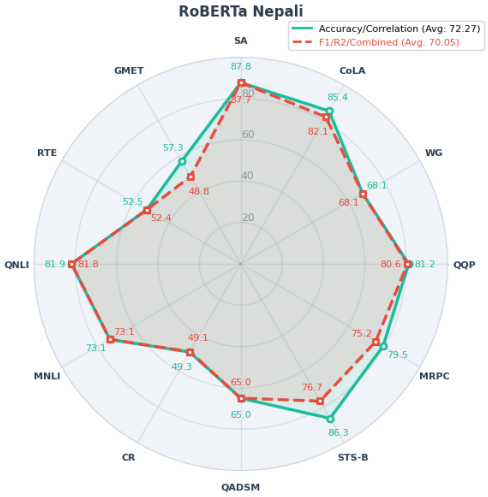

Figure 31: RoBERTa Nepali across all tasks

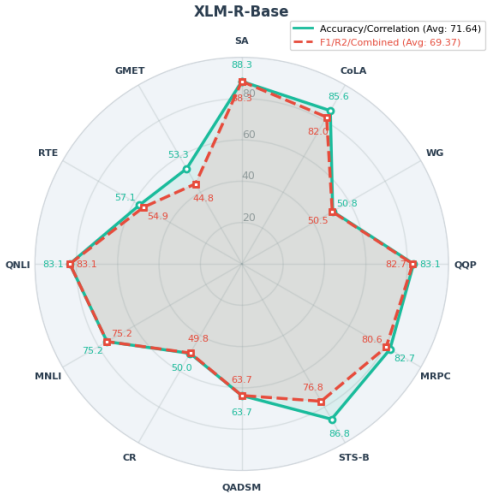

Figure 34: XLM-R-Base across all tasks

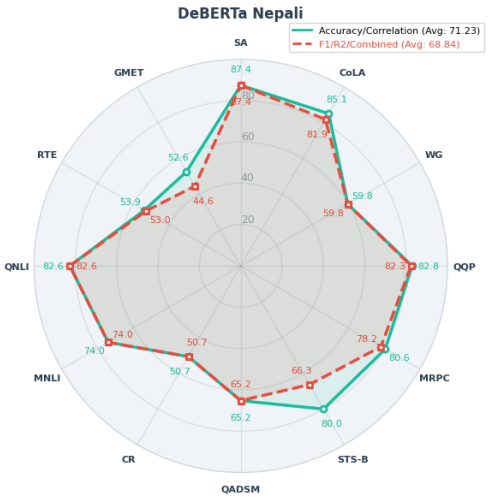

Figure 32: DeBERTa Nepali across all tasks

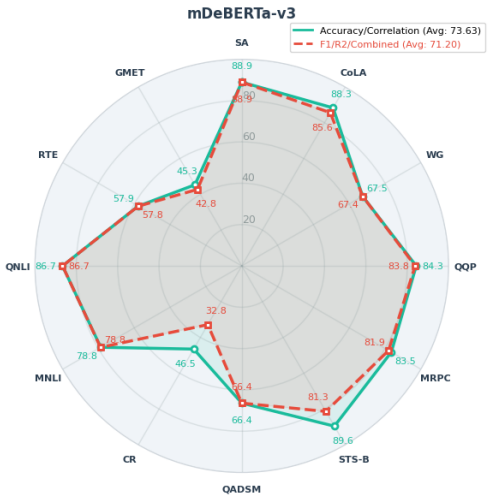

Figure 35: mDeBERTa-v3 across all tasks

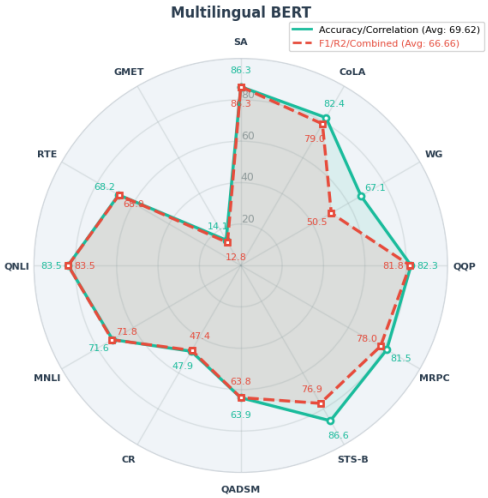

Figure 33: Multilingual BERT across all tasks