Benchmarking the Performance of Pre-trained LLMs across Urdu NLP Tasks

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, Sarmad Hussain

Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore, Pakistan
firstname.secondname@kics.edu.pk

Abstract

Large Language Models (LLMs) pre-trained on multilingual data have revolutionized natural language processing research, by transitioning from languages and task specific model pipelines to a single model adapted on a variety of tasks. However majority of existing multilingual NLP benchmarks for LLMs provide evaluation data in only few languages with little linguistic diversity. In addition these benchmarks lack quality assessment against the respective state-of the art models. This study presents an in-depth examination of 7 prominent LLMs: GPT-3.5-turbo, Llama 2-7B-Chat, Llama 3.1-8B, Bloomz 3B, Bloomz 7B1, Ministral-8B and Whisper (Large, medium and small variant) across 17 tasks using 22 datasets, 13.8 hours of speech, in a zero-shot setting, and their performance against state-of-the-art (SOTA) models, has been compared and analyzed. Our experiments show that SOTA models currently outperform encoder-decoder models in majority of Urdu NLP tasks under zero-shot settings. However, comparing Llama 3.1-8B over prior version Llama 2-7B-Chat, we can deduce that with improved language coverage, LLMs can surpass these SOTA models. Our results emphasize that models with fewer parameters but richer language-specific data, like Llama 3.1-8B, often outperform larger models with lower language diversity, such as GPT-3.5, in several tasks.

1 Introduction

The rapid increase in the application of Artificial Intelligence (AI) across a diverse spectrum of research areas including machine translation, natural language understanding and question answering can be attributed to the remarkable performances exhibited by Foundation Models (FM) (Bommasani et al., 2021). Based on the framework of transformers (Vaswani et al., 2017), multilingual large language models (LLM) are a prominent category of foundation models that can be utilized in multiple downstream tasks. A number of studies have have evaluated the potential of LLMs on various Natural Language Processing (NLP) tasks. LLMRec, a LLM-based recommender system (Liu et al., 2023) evaluated 3 LLMs including Llama, ChatGPT and ChatGLM on 5 recommendation tasks. (Zhong et al., 2021) conducted a human evaluation encompassing 10 LLMs with variations in pre-training methods, prompts, and model scales evaluated the zero-shot summarization capability. (Bian et al., 2023) used 11 datasets covering 8 domains to evaluate the LLMs' ability in answering common sense questions. (Hendy et al., 2023) conducted evaluations on 3 GPT models: ChatGPT, GPT3.5 (text-davinci-003), and text-davinci002 using 9 language pairs including low resource languages, to evaluate 18 machine translation directions. Holistic Evaluation of Language Models (HELM) project (Liang et al., 2023) evaluated 30 LLMs (open, limited-access, and closed models) for English across 42 NLP tasks. (Ahuja et al., 2023) conducted a multilingual evaluation of GPT 2.5 and Bloomz, comparing their performance with SOTA on 8 NLP tasks involving 33 languages. (Srivastava et al., 2023) conducted a comprehensive evaluation of 214 tasks, including 48 non-English low-resource languages using 13 transformer models and 8 GPT-3 series models with varying parameters from 125 million to 175 billion. Another notable effort was conducted by (Abdelali et al., 2024) for evaluation of 3 LLMs on 33 unique tasks for Arabic Language.

Our study, focuses on evaluating the potential of both closed and open LLMs for supporting Urdu, a low resource language with limited data coverage in LLM's pre-training. In our experiments we utilize GPT3.5 turbo by OpenAI, Llama 2 and Llama 3.1 by Meta , Bloomz 3B and 7B1 by Big Science, Ministral 8B by Mistral AI and Whisper by OpenAI in zero-shot setting, and perform evaluation on 17 Urdu NLP tasks analyzing their performances with the existing SOTA models. To the best of our knowledge, this is the first in depth evaluation of prominent LLMs in Urdu Language context.

2 Approach

For benchmarking of Urdu NLP tasks, we perform experiments using GPT 3.5, Bloomz 3B and Bloomz 7B1 , Llama 2 and Llama 3.1, Ministral 8B and Whisper in zero-shot setting and comparatively analyse the results with the respective SOTA models. Model selection was based on factors like accessibility (open/closed), infrastructure requirement, performance and language support. GPT 3.5 was selected because of its superior performance on English tasks. Among open models, popular multilingual models i.e. Llama 2 , Llama 3.1, Ministral 8B and Bloomz were evaluated for text processing tasks and Whisper models were evaluated for speech recognition task. Due to budget limitations and lack of Urdu data in the pre-training, other closed LLMs models were not investigated.

The evaluation of LLMs involved prompting and significant post-processing to extract the output in desired format. A number of prompts were curated for all NLP tasks following the recommended format and instruction pattern proposed by LAraBench (Abdelali et al., 2024). The prompts for each model were optimized after testing them on a few samples for each task. These prompts have been reported in Appendix. A After obtaining a reasonable prompt, we used the LLM models in different settings. OpenAI's API was used for GPT 3.5. For Bloomz, we ran the model on Google Colab utilizing 16GB GPU and for Llama 2, Llama 3.1 and Ministral 8B, we used on premises hosted versions utilizing 2X40GB A100 GPUs. Results were post-processed in all cases to align with the test set's output. The following section elaborates the LLMs (including prompting and post-processing details), NLP Tasks, Datasets, SOTA Models and evaluation metrics, used in the study.

2.1 Models

2.1.1 GPT 3.5

GPT 3.5 Turbo has been trained on 175B parameters, encompassing both text and code data. GPT 3.5 despite being closed-source and less powerful than GPT-4 (OpenAI and et al., 2023), is more cost-effective, as its provides free access for experimentation. Additionally, at the time of research it was the most advanced model available from OpenAI for fine-tuning.

2.1.2 Bloomz 3B and 7.1B

Bloomz (Muennighoff et al., 2023), a Multitask Prompting Fine Tuned (MTF) version of the BLOOM (BigScienceWorkshop and et al., 2023), is trained on ROOTS corpus (Laurençon et al., 2023) covering 59 languages (including 13 programming languages, and 2.59TB of Urdu language data). For evaluation, the Bloomz 3B and 7.1B models from HuggingFace were used due to their open-source availability, and optimal balance between size and computational resources.

2.1.3 Llama 2 and Llama 3.1

Llama 2 (Touvron et al., 2023), released by Meta, is trained on 2 trillion tokens, with 89.70% of its content in English. Llama 3.1 (AI@Meta, 2024), available in three variants with 8 billion, 70 billion and 405 billion parameters, is trained on over 15 trillion tokens. Both models support 8k context lengths. For evaluation, the Llama 2-7b and Llama 3.1-8b models were used due to their open-source availability and potential for transfer learning and generalization to languages with limited data.

2.1.4 Ministral 8B

The Ministral 8B (Mistral AI Team, 2024) is trained on a mixture of multilingual and code datasets, supporting a context window of up to 128k facilitated by an interleaved sliding-window attention mechanism and a vocabulary of 131k. We benchmarked this model due to its open-source availability and its capability for low-memory inference, and its ability to be fine-tuned and adapted to a variety of tasks.

2.1.5 Whisper

Whisper (Radford et al., 2022), an Automatic Speech Recognition (ASR) model developed by OpenAI, is trained on an extensive dataset comprising 680,000 hours of multilingual and multitask supervised data collected from the web. Among the diverse languages included, Whisper incorporates only 104 hours of Urdu speech corpus. For inference , we utilized the small, medium and large variants of the pre-trained Whisper model. The small variant has 12 layers, 12 attention heads, a width of 768 with 244 million parameters. The medium variant is characterized by 24 layers, 16 attention heads, a width of 1024, and consists of 769 million parameters while, the large variant features

Task	Dataset	Dataset Size	Testset Size
Name Entity Recognition	MK-PUCIT (Kanwal et al., 2019)	99718	4165
News Categorization	COUNTER (Sharjeel et al., 2017)	1200	360
Intent Detection	Urdu Web Queries Dataset (UWQ-22) (Shams and Aslam, 2022)	6819	850
Hate Speech Detection	ISE-Hate corpus (Akram et al., 2023)	21759	2176
Hate Speech Detection	CLE-Hatespeech dataset (Ali et al., 2021)	5432	1087
Propaganda Detection	ProSOUL (Kausar et al., 2020)	11574	1737
Abusive Language Detection	HASOC - Task A(Das et al., 2021)	2400	240
Threat Detection	HASOC - Task B(Das et al., 2021)	9950	1975
Cyber Bullying Identification	Cyberbullying corpus (Adeeba et al., 2024)	12759	2480
Fake News Detection	(Khan et al., 2023)	4097	820
Hate Speech Categorization	ISE-Hate corpus(Akram et al., 2023)	8702	871
Text Summarization	CORPURES (Humayoun and Akhtar, 2022)	2649	311
Sentiment Analysis	(Muhammad and Burney, 2023)	10008	2002
Sentiment Analysis	Corpus of Aspect-based Sentiment for Urdu Political Data (ul Haq et al., 2020)	8760	1450
Multi-label Emotion Classification	Overview of EmoThreat (Task A) (Ashraf et al., 2022)	9750	1950
Emotion Classification	Urdu Nastalique Emotions Dataset (UNED) (Bashir et al., 2023)	4000	397
Machine Translation(Quran)	English-Urdu Religious Parallel Corpus (Jawaid and Zeman, 2011)	6414	200
Machine Translation(Bible)	English-Urdu Religious Parallel Corpus (Jawaid and Zeman, 2011)	7957	257
Abstractive Summarization	CLE Meeting Corpus (Sadia et al., 2024)	240	10
POS Tagging	Sense Tagged CLE Urdu Digest Corpus (Urooj et al., 2014)	100000	22522
ASR (Read Speech)	Urdu Speech Corpus (Farooq et al., 2019)	-	9.5 hours
ASR (Broadcast)	Urdu Broadcast (BC) Corpus(Khan et al., 2021)	-	4.3 hours

Table 1: NLP Tasks and Dataset Statistics

32 layers, 20 attention heads, a width of 1280, and comprises 1550 million parameters.

2.2 Tasks and Datasets

This study has focused on a comprehensive evaluation of pre-trained open and closed LLMs on Urdu NLP tasks. This study utilizes 22 publicly available datasets (see Table 1) to evaluate 17 Urdu NLP tasks as discussed in the following sections.

2.2.1 Name Entity Recognition

Name Entity Recognition (NER) is a sequence tagging task that involves identifying entities, such as names of people, organizations, locations, dates, etc. For its evaluation, we used the MK-PUCIT dataset and its SOTA model reported in (Kanwal et al., 2019).

2.2.2 News Categorization

News categorization classify news articles into topics based on their content. For its evaluation, COUNTER dataset (Sharjeel et al., 2017) was used that consisted of articles from 5 different domains and its SOTA is reported in (Khan et al., 2023).

2.2.3 Intent Detection

Intent detection focuses on determining the communicative intent behind a user's input query in the form of text or speech. For our evaluation, we used the UWQ-22 dataset and SOTA model reported in (Shams and Aslam, 2022).

2.2.4 Ethics and NLP: Factuality and Harmful Content Detection

These tasks aim to evaluate the accuracy of information, identify and combat misinformation, and detect harmful content. We benchmark several tasks such as i) Hate Speech Detection using the ISE-Hate corpus by (Akram et al., 2023) and CLE-Hatespeech dataset (Ali et al., 2021). ii) Propaganda Detection on the ProSOUL dataset developed by (Kausar et al., 2020). iii) Abusive Language Detection in Urdu, on the dataset by (Das et al., 2021) for their Subtask A. iv) Threat Detection on the dataset of (Das et al., 2021) for Subtask B. v) Cyber Bullying Identification using Cyberbullying corpus (Adeeba et al., 2024) vi) Fake News Detection using dataset prepared by (Khan et al., 2023) vii) Hate Speech Categorization using ISE-Hate corpus by (Akram et al., 2023).

2.2.5 Text Summarization

Text summarization involves extracting the most important sentences from a document to create a condensed version retaining essential information. We evaluated the LLMs on:

- **Extractive Summarization**

  Extractive summarization condenses text by

Task	Dataset	Metric	GPT 3.5	Bloomz 3B	Bloomz 7B1	Llama 2	Llama 3.1	Ministral 8B	SOTA	Delta
Name Entity Recognition	MK-PUCIT	Macro-F1	**0.55**	0.25	0.27	0.15	0.41	0.25	0.77	0.22
News Categorization	COUNTER	Macro-F1	**0.87**	0.58	0.48	0.13	0.64	0.67	0.70	-0.17
Intent Detection	Urdu Web Queries Dataset (UWQ-22)	Macro-F1	0.30	0.22	0.18	0.07	**0.42**	0.34	0.90	0.56
Hate Speech Detection	ISE-Hate corpus	Macro-F1	**0.72**	0.52	0.53	0.48	0.70	0.53	0.83	0.11
Hate Speech Detection	CLE-Hatespeech dataset	Macro-F1	0.67	0.35	0.43	0.51	**0.72**	0.54	0.98	0.26
Propaganda Detection	ProSOUL	Macro-F1	0.31	0.47	0.47	0.44	**0.66**	0.53	0.83	0.17
Abusive Language Detection	HAOSOC - Task A	Macro-F1	0.23	**0.51**	0.47	0.44	0.50	0.48	0.88	0.37
Threat Detection	HAOSOC - Task B	Macro-F1	**0.49**	0.35	0.20	0.21	0.40	0.46	0.54	0.05
Cyber Bullying Identification	(Adeeba et al., 2024)	Macro-F1	0.19	0.15	0.10	0.06	**0.22**	0.08	0.84	0.41
Fake News Detection	(Khan et al., 2023)	Macro-F1	0.55	0.52	0.51	0.47	**0.72**	0.57	0.93	0.21
Hate Speech Categorization	ISE-Hate corpus	Macro-F1	**0.40**	0.28	0.15	0.21	0.30	0.22	0.83	0.43
Extractive Summarization	CORPURES	Average Rouge-2 F1 score	0.54	0.46	0.55	0.59	**0.62**	0.52	0.57	-0.04
Abstractive Summarization	CLE Meeting Corpus	Rouge-1 Score (Avg)	0.22	0.02	0.07	0.006	**0.24**	0.06	0.31	0.07
Sentiment Analysis	(Muhammad and Burney, 2023)	Macro-F1	**0.62**	0.35	0.33	0.3	0.44	0.36	0.88	0.26
Sentiment Analysis	Corpus of Aspect-based Sentiment for Urdu Political Data	Macro-F1	0.31	0.20	0.21	0.13	**0.37**	0.28	0.70	0.37
Multi-label Emotion Classification	Overview of EmoThreat (Task A)	Macro-F1	0.20	0.17	0.26	–	**0.40**	0.29	0.68	0.28
Emotion Classification	Urdu Nastalique Emotions Dataset (UNED)	Macro-F1	0.32	0.25	0.21	0.18	0.24	**0.41**	0.87	0.46
Machine Translation (Quran)	English-Urdu Religious Parallel Corpus	BLEU	**3.75**	1.91	2.36	2.49e-78	3.44	0.004	13.24	9.49
Machine Translation(Bible)	English-Urdu Religious Parallel Corpus	BLEU	5.96	2.28	2.47	0.097	**6.43**	1.31e-78	13.99	8.03
POS Tagging	CLE Urdu POS Tagset	Accuracy	**0.49**	0.11	0.06	0.09	0.31	0.14	0.96	0.47

Table 2: Results from zero-shot experiments of GPT 3.5, Bloomz 3B, Bloomz 7B1, Llama 2, Llama 3.1 and Ministral 8B Models Compared to SOTA over NLP tasks. **Bold** text indicates the best score among models.

selecting and combining key sentences directly from the original content. For the evaluation of this task, we used the CORPURES dataset by (Humayoun and Akhtar, 2022).

- **Abstractive Summarization**

  Abstractive summarization generates concise summaries by understanding and paraphrasing the core meaning of a text into new, shorter sentences. For its evaluation, we have used CLE Meeting Corpus and its SOTA available in (Sadia et al., 2024).

2.2.6 Sentiment and Emotion Analysis

These tasks include understanding and interpreting human expressions in textual data. For Sentiment analysis, datasets from (Muhammad and Burney, 2023) and CLE (ul Haq et al., 2020) are used. For emotion analysis we used dataset from (Ashraf et al., 2022) for their Task A: Multi-label Emotion Detection consisted of "Neutral" label and Ekman's six basic emotions (Ekman, 1999). The other dataset used was Urdu Nastalique Emotions Dataset (UNED) by (Bashir et al., 2023).

2.2.7 Machine Translation

Machine translation of Urdu is challenging due to its morphological complexity. To evaluate the translation capabilities of LLMs for English Urdu pair, we utilized the dataset by (Jawaid and Zeman, 2011) for Quran and Bible translations containing 200 and 257 testing samples respectively.

2.2.8 Part of Speech (POS) Tagging

POS tagging is a fundamental task in NLP that involves labeling each word in a sentence with its corresponding part of speech, such as noun, verb, adjective, etc. To evaluate this task we have used the CLE Urdu POS Tagset with the SOTA reported in (Ahmed et al., 2014).

2.2.9 Automatic Speech Recognition (ASR)

ASR automatically converts spoken language into text. For its evaluation, we utilized the small, medium and large variant of the pre-trained Whisper model (Radford et al., 2022). We benchmarked this model against the SOTA models using its pre-trained weights for both broadcast and read speech recognition tasks using following two corpora:

- Urdu Broadcast (BC) corpus: A broadcast speech corpus (Khan et al., 2021) with 4.3 hours of data from 25 speakers (14 males and 11 females). This dataset includes recordings from five different broadcast channels and YouTube, covering genres such as entertainment, health and science, current affairs, and politics.

- Urdu Speech corpus: A read speech corpus (Farooq et al., 2019) consisting of 9.5 hours of Urdu speech from 62 speakers. The dataset is balanced in terms of gender and recording channels.

2.3 Zero-Shot Setup

For all LLMs; GPT 3.5, Bloomz 3b and 7b, Llama 2 and Llama 3.1 and Ministral 8B we use zero-shot prompting giving natural language instructions describing the task and specify the expected output. Prompts allow LLMs to learn context and narrows the inference space to produces accurate output as further elaborated in the section 2.5.

2.4 Inference Settings

The inference experiments for Llama 2, Llama 3.1 and Ministral 8B were conducted using two parallel NVIDIA A100-PCIE-40GB GPUs, providing a combined computational capacity of 80GB. During the inference, nearly 90 percent of the total GPU capacity was utilized. For experiments of GPT-3.5, API from OpenAI was utilized. Inference experiments with GPT-3.5 were conducted using Google Colab. Inference experiments with Bloomz's 3B and 7.1B models, available on huggingface, were also conducted using Google Colab. For Speech processing experiments using Whisper, two NVIDIA RTX3060-12GB GPUs were employed, providing a combined computational capacity of 24GB.

2.5 Prompt Engineering and Post Processing

In our experimentation with different LLMs, we tweaked the prompts based on the models input. Prompts for tasks such as News categorization A.2 and Hate speech Categorization A.11 were challenging because they required outputs from predefined ground-truth categories. Prompts for Machine Translation task A.19 had to be engineered so that the model's output only includes the translated text. Thus optimal prompts were curated by testing against each model on few samples, while ensuring no bias in decision-making.

Despite careful prompting, model responses required post-processing to align with desired outcomes e.g. capitalization ("fake" vs. "Fake"), standardizing output formats ("1. Propaganda" to "1"), and omitting "explanations" and "note" produced with the models' responses, specifically in Hate speech detection A.5 task. Some model outputs didn't match desired outcomes, e.g. News categorization included 5 domains i.e. sports, showbiz, foreign , national , business however the models output out of context domains such as "politics" and "entertainment". Among all the models, Llama 2 required the most output post-processing.

For a thorough description of the prompts crafted for each LLM, please refer to Appendix A.

2.6 SOTA Models

In this study, we benchmark the capabilities of LLMs in a zero-shot scenario by comparing them with SOTA models as reported in respective studies. These SOTA models employed diverse architectures including Capsule NN, Support Vector Machine (SVM), Random Forest (RF), Decision Tree (J48), Sequential Minimal Optimization (SMO), Convolutional Neural Networks (1D-CNN), LSTM with CNN features , Naive Bayes classifier and various multilingual transformer models such as m-BERT and frameworks like XGboost and LGBM.

2.7 Evaluation Metrics

The evaluation metrics used for the experiments have been kept identical to the one used in the respective state of the art references. They are Macro-F1, Rouge 2 F1 score, BLEU [1], accuracy and Word Error Rate (WER). We have also computed the delta to highlight the differential between best performing LLM's output with the SOTA model.

[1] https://www.nltk.org/api/nltk.translate.bleu

Task	Domain	Metric	Whisper (Large)	Whisper (Medium)	Whisper (Small)	SOTA	Delta
ASR	Read Speech	WER	**23.51**	27.88	36.90	16.94	-6.57
ASR	Broadcast	WER	**27.97**	35.57	42.57	18.59	-9.38

Table 3: Performance Metrix of ASR for Whisper Large, Medium and Small Models Compared to SOTA.

3 Results and Discussion

The results on text processing tasks of our experimentation have been summarized in Figure 1. The Figure presents a grid of bar graphs for each NLP task, with the y-axis showing evaluation metrics specific to each task. For classification and detection tasks, the y-axis represents the macro F1 score. For summarization tasks, it shows the average ROUGE-2 score, while for machine translation tasks, it displays the BLEU score. Each model is represented by a distinct color bar , as indicated in the Figure's legend, which is kept consistent across all tasks, with the bar of SOTA providing a reference point for comparison. Missing bars in certain tasks indicate that the model outputs were effectively zero (e.g., in Table 2 value is 2.49e-78 for Llama 2 on the Machine Translation (Quran) task, and 1.31e-78 for Mistral 8B on the Machine Translation (Bible) task), reflecting negligible performance.

Our results show that LLMs differ in their applicability to different data regimes and tasks. LLM models were able to surpassed the SOTA model for news categorization with GPT 3.5 and Llama 3.1 for Extractive Summarization. In all other experiments, LLMs remained lower than the SOTA models (reference Table 2). Across all experiments, Llama 3.1 outperformed in 10 of the 17 tasks, while GPT-3.5 excelled in 8 tasks. In comparison, Bloomz and Ministral 8B each led in only one task. The minimum delta obtained was 0.05 between GPT 3.5 and SOTA model for threat detection task. In comparison with other the open LLMs, Llama 3.1 performed better in majority of the NLP tasks which is due to its extensive multilingual data, architecture and advanced training techniques, enabling it to effectively generalize across languages and tasks.

In choice and evaluation of LLMs, Bloomz 3B and Bloomz 7B1 were initially chosen for experimentation due to their early introduction and multilingual capabilities. However, they have not kept pace with advancements seen in other models like Llama. Analysis reveals that there is no significant performance efficiency gained from transitioning from Bloomz-3B to Bloomz-7B1 as evident from Table 2. In contrast, the performance of Llama models has notably improved, particularly from Llama 2 to Llama 3.1, indicating a more effective evolution in their design and capabilities.

Based on our evaluations, the top two performing models for NLP tasks are GPT-3.5 and Llama 3.1 with comparable performances as evident from Figure 1. Llama 3.1 outperformed GPT 3.5 in 11 NLP tasks and was even better than SOTA in Extractive Summarization. On the other hand, GPT 3.5 was better than Llama 3.1 in 8 tasks and surpassed SOTA in News Categorization task. Llama 3.1's superior performance is due to the increased coverage of Non-English data in the model as well as increased amount of pretraining data i.e. 15 trillion tokens.

The performance of Ministral 8B is comparable to GPT 3.5 and Llama 3.1. It outperformed GPT 3.5 in 6 NLP Tasks i.e. Intent Detection, Propaganda Detection, Abusive Language Detection, Fake News Detection, Multi Label Emotion Classification and Emotion Classification. And outperformed Llama 3.1 in News Categorization and it was best among all models in Emotion Classification. However its performance was quite inadequate on generation tasks like Machine Translation (reference Figure 1). Overall its performance is good in detection tasks and is attributed to its interleaved sliding-window attention mechanism, which enables it to efficiently handle extended contexts with reduced memory usage. Detection tasks such as Fake News Detection and Propaganda Detection often require recognizing patterns across longer texts. This enhanced ability to retain and utilize extended contextual information allowed Ministral 8B to excel in detection tasks.

The results of Speech Processing tasks are summarized in Table 3. The analysis of the results indicates that the SOTA models, which were trained

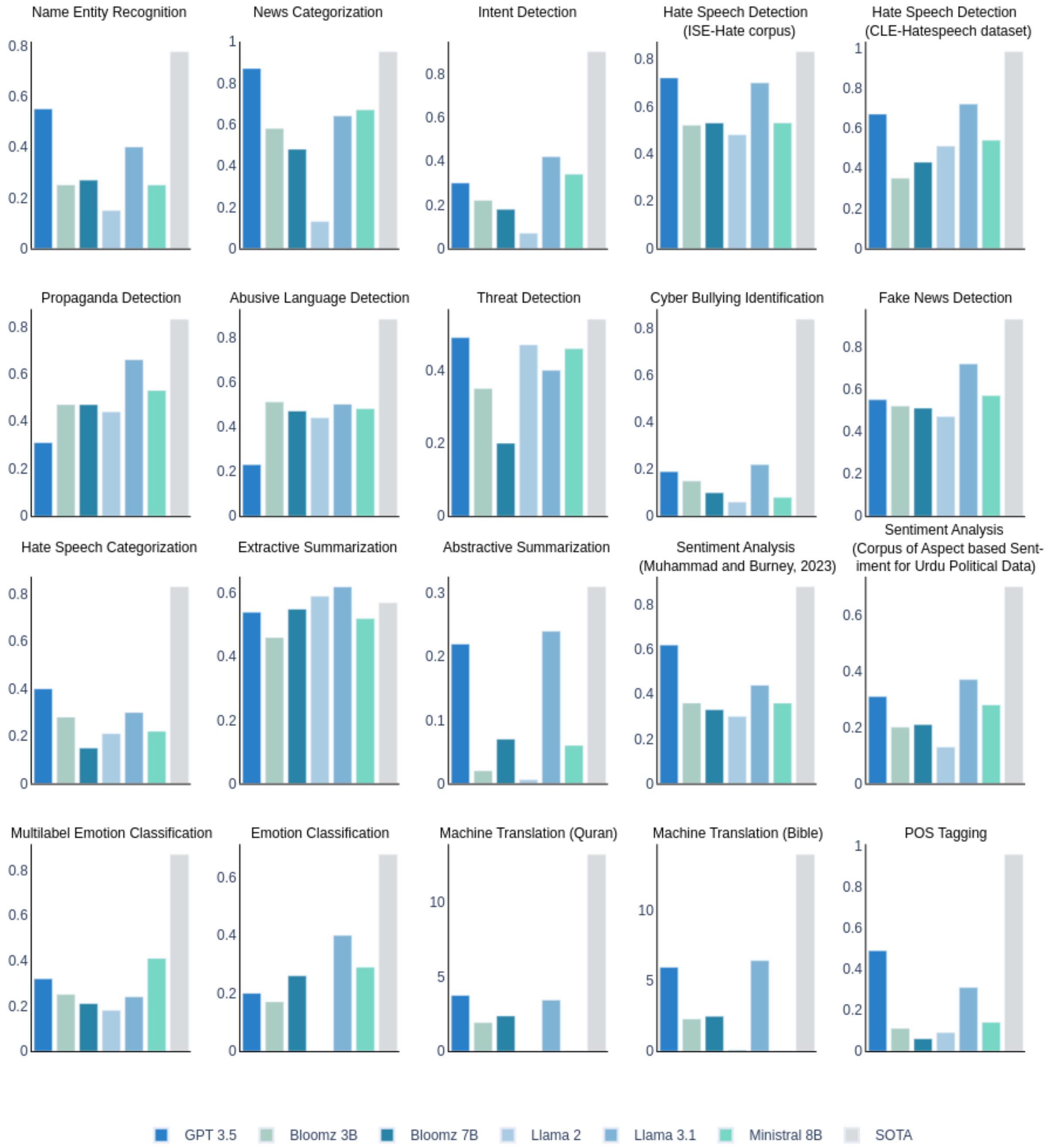

Figure 1: The performance of different models in zero-shot scenario as compared to SOTA. Missing bars in some tasks mean that the specific model cannot perform the specified task.

on a larger corpus of Urdu data, outperformed all variants of the Whisper model across the evaluated datasets. These findings suggest that while the SOTA models trained on more extensive Urdu datasets exhibited superior performance, the larger variant of the Whisper also demonstrated improved performance compared to its medium and small counterpart, underscoring the importance of model size and complexity in ASR tasks. The negative delta values indicate that the model's performance falls below the SOTA benchmarks, highlighting a gap to be addressed.

Error analysis of the LLMs' output against the ground truth revealed two main factors that account for the decline in overall F1 scores of LLMs. The factors include i) discrepancies in the output format, where the output contained extra or omitted tokens, and ii) the generation of out-of-scope labels. These observations imply that the seamless deployment of LLMs may be challenging, requiring substantial efforts either in formulating precise prompts for accurate outputs or engaging in post-processing to align the outputs with reference labels.

Thus, performance of LLMs significantly depends on well-curated prompts and intelligent post-processing of the outputs. While Llama 2 and Bloomz show a notable performance deficit compared to the SOTA, the newer Llama version i.e. Llama 3.1 and GPT 3.5 succeeds in mitigating this gap to a considerable extent.

4 Conclusion and Future Work

In this study, we benchmark the potential of both open and closed LLMs on 17 Urdu NLP tasks employing a substantial number of publicly accessible datasets. Through our experiments we provide a comparative performance analysis for each task and dataset against the SOTA. These findings will assist the Urdu NLP community in selecting suitable models for usage and fine-tuning within specific contexts. As future work, we aim to develop a public leader board for Urdu benchmarking and explore integration of additional models, tasks, and datasets. Also, after evaluating multiple models, we are focusing on pretraining Llama 3.1 to enhance Urdu language support and expand token coverage for greater adaptability across diverse languages and domains. This will also includes domain-specific fine-tuning to further boost performance, leveraging Llama 3.1's compact size, active community, and robust results.

Limitations

Our study is confined to seven LLMs and does not include the heavier versions of models such as Bloomz-170B or Llama 3.1 405B due to hardware and computational resource limitations which may impact the comprehensiveness of the analysis. This limitation may affect the generalization of the findings to models with higher parameters, potentially missing insights into the performance of more robust versions of these language models. Our study also primarily concentrates on evaluating the models in a zero-shot setting. While this setting provides valuable insights into the models' out-of-the-box performance, it may not capture the full potential of fine-tuned models for specific tasks. Our study also does not extensively delve into the quality and representativeness of the training data for Urdu language used in these models.

Acknowledgments

We would like to express our sincere gratitude to Dr. Miriam Butt for allowing us to utilize the infrastructure at the University of Konstanz for our benchmarking experiments. Additionally, we extend our acknowledgment to Miss Ayesha Khalid for supplying the ASR dataset and corresponding SOTA techniques.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. Larabench: Benchmarking arabic ai with large language models.

Farah Adeeba, Muhammad Irfan Yousuf, Izza Anwer, Sardar Umair Tariq, Abdullah Ashfaq, and Malik Naqeeb. 2024. Addressing cyberbullying in urdu tweets: a comprehensive dataset and detection system. *PeerJ Comput. Sci.*, 10:e1963.

Tafseer Ahmed, Saba Urooj, Sarmad Hussain, Asad Mustafa, Rahila Parveen, Farah Adeeba, Annette Hautli, and Miriam Butt. 2014. The cle urdu pos tagset.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai.

AI@Meta. 2024. Llama 3 model card.

Muhammad Akram, Khurram Shahzad, and Maryam Bashir. 2023. Ise-hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in urdu. *Information Processing & Management*, 60.

Muhammad Ali, Ehsan Ul Haq, Sahar Rauf, Kashif Javed, and Sarmad Hussain. 2021. Improving hate speech detection of urdu tweets using sentiment analysis. *IEEE Access*, PP:1–1.

Noman Ashraf, lal khan, Sabur Butt, Hsien-Tsung Chang, Grigori Sidorov, and Alexander Gelbukh. 2022. Multi-label emotion classification of urdu tweets. *PeerJ Computer Science*, 8:e896.

Muhammad Farrukh Bashir, Abdul Rehman Javed, Muhammad Umair Arshad, Thippa Reddy Gadekallu, Waseem Shahzad, and Mirza Omer Beg. 2023. Context-aware emotion detection from low-resource urdu language using deep neural network. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models.

BigScienceWorkshop and et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021. Abusive and threatening language detection in urdu using boosting based and BERT based models: A comparative approach. *CoRR*, abs/2111.14830.

Paul Ekman. 1999. *Basic Emotions*, chapter 3. John Wiley & Sons, Ltd.

Muhammad Farooq, Farah Adeeba, Sahar Rauf, and Sarmad Hussain. 2019. Improving large vocabulary urdu speech recognition system using deep neural networks. pages 2978–2982.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Muhammad Humayoun and Naheed Akhtar. 2022. Corpures: Benchmark corpus for urdu extractive summaries and experiments using supervised learning. *Intelligent Systems with Applications*, 16:200129.

Bushra Jawaid and Daniel Zeman. 2011. Word-order issues in english-to-urdu statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95.

Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2019. Urdu named entity recognition: Corpus generation and deep learning applications.

Soufia Kausar, Bilal Tahir, and Amir Mehmood. 2020. Prosoul: A framework to identify propaganda from online urdu content.

E. Khan, S. Rauf, F. Adeeba, and S. Hussain. 2021. A multi-genre urdu broadcast speech recognition system. In *Proceedings of The O-COCOSDA 2021*, Singapore.

Sajid Khan, Mehmoon Anwar, Huma Ayub, Farooq Ali, and Marriam Nawaz. 2023. Fake news classification using machine learning: Count vectorizer and support vector machine. *Journal of Computing & Biomedical Informatics*, 4.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Junling Liu, Chao Liu, Peilin Zhou, Qichen Ye, Dading Chong, Kang Zhou, Yueqi Xie, Yuwei Cao, Shoujin Wang, Chenyu You, and Philip S. Yu. 2023. Llmrec: Benchmarking large language models on recommendation task.

Mistral AI Team. 2024. Un ministral, des ministraux. Accessed: 2024-10-29.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Khalid Bin Muhammad and S. M. Aqil Burney. 2023. Innovations in urdu sentiment analysis using machine and deep learning techniques for two-class classification of symmetric datasets. *Symmetry*, 15(5).

OpenAI and et al. 2023. Gpt-4 technical report.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Bareera Sadia, Farah Adeeba, Sana Shams, and Kashif Javed. 2024. Meeting the challenge: A benchmark corpus for automated urdu meeting summarization. *Information Processing & Management*, 61(4):103734.

Sana Shams and Muhammad Aslam. 2022. Improving user intent detection in urdu web queries with capsule net architectures. *Applied Sciences*, 12:11861.

Muhammad Sharjeel, Rao Nawab, and Paul Rayson. 2017. Counter: corpus of urdu news text reuse. *Language Resources and Evaluation*, 51.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ehsan ul Haq, Sahar Rauf, Sarmad Hussain, and Kashif Javed. 2020. Corpus of aspect-based sentiment for urdu political data.

Saba Urooj, Sana Shams, Sarmad Hussain, and Farah Adeeba. 2014. Sense tagged cle urdu digest corpus.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Appendix

A.1 Prompts - Name Entity Recognition

A.1.1 Bloomz

Perform Name Entity Recognition for the words using the following technique: - Mark names, nicknames, cast, family, and relational names as Person. - Mark names of companies, media groups, teams, and political parties as Organization. - Mark all man-made structures and politically defined locations, such as names of countries, cities, and places like railway stations, as Location. - Mark all remaining words, such as prepositions, adjectives, adverbs, and names of books and movies, as Other. No explanation is required. Just output the Entity name.Word: Entity:

A.1.2 GPT 3.5

Perform Name Entity Recognition corresponding to each word using the following annotation technique: Person : name,nickname,cast,family,relational names and titles. God's name should NOT be marked as Person. Organization : name of company, media group, team,political party. Name of product or brand should NOT be marked as Organization. Location : all man-made structures and politically defined locations such as names of countries,city and places like railway station etc. A generic reference to location should NOT be marked as Location. Other : all remaining words, such as prepositions, adjectives, adverbs, names of books and movies etc. No explanation is required. Just output the tag name. word =

A.1.3 Llama 2

You are Performing Name Entity Recognition for the urdu words.«/SYS» Human: Word: Please select one of the following entity: Person Organization Location Other No explanation or further assistance is required. Only entity name is required Assistant: The entity is

A.1.4 Llama 3.1

You are a name entity recognition model. Your task is to mark the entity as Person, Organization, Location, or Other in Urdu text samples. Ensure that your outputs are Person, Organization, Location, or Other. No explanation is required.

A.1.5 Ministral 8B

Perform Name Entity Recognition for the words using the following technique: - Mark names, nicknames, cast, family, and relational names as Person. - Mark names of companies, media groups, teams, and political parties as Organization. - Mark all man-made structures and politically defined locations, such as names of countries, cities, and places like railway stations, as Location. - Mark all remaining words, such as prepositions, adjectives, adverbs, and names of books and movies, as Other. No explanation is required. Just output the Entity name.Word: Entity:

A.2 Prompts - News Categorization

A.2.1 Bloomz

News: Classify the given news into one of the following category 0. sports 1. national 2. foreign 3. showbiz 4. business Choose the best suited label from above. Your output should be 0-4 only. No explanation. Only 0-4. No other label or additional text. Label (0,1,2,3,4):

A.2.2 GPT 3.5

News: Classify the given news into one of the following category 0. sports 1. national 2. foreign 3. showbiz 4. business Choose the best suited label from above. Your output should be the name of the category only. No explanation.No other label or additional text. Category:

A.2.3 Llama 2

Provide the label of the above news from the following: 0. sports 1. national 2. foreign 3. showbiz 4. business No explanation. Please answer in numbers News : Answer:

A.2.4 Llama 3.1

News: Classify the given news into one of the following category 0. sports 1. national 2. foreign 3. showbiz 4. business Choose the best suited label from above. Your output should be 0-4 only. No explanation. Only 0-4. No other label or additional text. Label (0,1,2,3,4):

A.2.5 Ministral 8B

News: Classify the given news into one of the following category 0. sports 1. national 2. foreign 3. showbiz 4. business Choose the best suited label from above. Your output should be 0-4 only. No explanation. Only 0-4. No other label or additional text. Label (0,1,2,3,4):

A.3 Prompts - Intent Detection

A.3.1 Bloomz

You are an intent classification model. Your task is to identify the intent in the following urdu sentence. Intents are: 0. Informational 1. Navigational 2. Transitional Output (0,1,2):

A.3.2 GPT 3.5

"system": "You are an intent detection classification model. You are an intent classification model. Your task is to identify the intent in the following urdu sentence. Intents are: 0. Informational 1. Navigational 2. Transitional Output (0,1,2):

A.3.3 Llama 2

You are an intent classification model. Your task is to identify the intent in the following urdu sentence. Intents are: 0. Informational 1. Navigational 2.

Transitional Dont write any explanation or reason for answer. Output (0,1,2):

A.3.4 Llama 3.1

You are an intent classification model. Your task is to mark the intent as 0 or 1 or 2 in Urdu text samples. Ensure that the model outputs '0' for Informational intent , '1' for Navigational intent and '2' for Transitional intent. Ensure that your outputs 0 or 1 or 2 only. No explanation is required.

A.3.5 Ministral 8B

You are an intent classification model. Your task is to identify the intent in the following urdu sentence. Intents are: Informational Navigational Transitional Only output the name of the intent. No explanation is required.

A.4 Prompts - Hate Speech Detection ISE-Hate corpus

A.4.1 Bloomz

Classify the hate sentence into the category it falls: Ethnic Interfaith Sectarian Other Output "0" for Other, "1" for "Sectarian", "2" for "Interfaith" and "3" for "Ethnic" Sentence: Class:

A.4.2 GPT 3.5

"system": "You are an expert in detecting hate speech in the urdu samples " Classify the hate sentence into the category it falls: Ethnic Interfaith Sectarian Other Output "0" for Other, "1" for "Sectarian", "2" for "Interfaith" and "3" for "Ethnic". No explanation is required Sentence: Output (0,1,2,3):

A.4.3 Llama 2

You are a hate speech classification model. Labels: 1: Sectarian hate 2: Interfaith hate 3: Ethnic hate 0: None of the above Instructions: To distinguish between hate speech and non-hate speech in text samples. Ensure that the model outputs "1" for hate related to "Sectarian", "2" for hate related to "Interfaith" and "3" for hate related to "Ethnic" and "0" if you think it does not fall in these three categories. Your output should be only 0, 1, 2 or 3. No explanation is required. Sentence: Label(0,1,2,3):

A.4.4 Llama 3.1

You are a hate speech classification model. Your task is to mark 0 or 1 or 2 or 3 in Urdu text samples. Ensure that the model outputs '1' for Sectarian hate text , '2' for Interfaith hate text , '3' for Ethnic hate text and '0' for any other hate text. Ensure that your outputs 0 or 1 or 2 or 3 only. No explanation is required.

A.4.5 Ministral 8B

You are a hate speech classification model. Labels: 1: Sectarian hate 2: Interfaith hate 3: Ethnic hate 0: None of the above Instructions: To distinguish between hate speech and non-hate speech in text samples. Ensure that the model output "1" for hate related to "Sectarian", "2" for hate related to "Interfaith" and "3" for hate related to "Ethnic" and "0" if you think it does not fall in these three categories. Your output should be only 0, 1, 2 or 3. No explanation is required. Sentence: Label(0,1,2,3):

A.5 Prompts - Hate Speech Detection CLE corpus

A.5.1 Bloomz

Classify the sentence as hate speech or non-hate speech. Output "1" for non-hate speech and "-1" for hate speech. No explanation is required. Sentence: Output(1 or -1):

A.5.2 GPT 3.5

"system": "You are an expert in detecting hate speech in the urdu samples" Classify the sentence as hate speech or non-hate speech. Output "1" for non-hate speech and "-1" for hate speech. No explanation is required. Sentence:

A.5.3 Llama 2

You are a hate speech classification model. Labels: 1: Non-hate speech -1: Hate speech Instructions: To distinguish between hate speech and non-hate speech in text samples. Ensure that the model outputs 1 for non-hate speech and -1 for hate speech. Your output should be only 1 or -1. No explanation is required Sentence: Label(1 or -1):

A.5.4 Llama 3.1

You are a hate speech classification model. Your task is to mark 0 or 1 in Urdu text samples. Ensure that the model outputs '0' for Non-hate text and '1' for Hate text. Ensure that your outputs 0 or 1 only. No explanation is required.

A.5.5 Ministral 8B

You are a hate speech classification model. Labels: 0: Non-hate speech 1: Hate speech Instructions: To distinguish between hate speech and non-hate speech in text samples. Ensure that the model outputs 0 for non-hate speech and 1 for hate speech. Your output should be only 0 or 1 Sentence: Label:

A.6 Prompts - Propaganda Detection

A.6.1 Bloomz

Classify the article as Propaganda or Non-Propaganda. Output '1' for Propaganda and '0' for Non-Propaganda. Don't concatenate input with output. No explanation is required. The article is: . Class(0 or 1):

A.6.2 GPT 3.5

Classify the article as Propaganda or Non-Propaganda. Output '1' for Propaganda and '0' for Non-Propaganda. No explanation is required. The article is: . Class(0 or 1):

A.6.3 Llama 2

Classify the article as Propaganda or Non-Propaganda. Output '1' for Propaganda and '0' for Non-Propaganda. Don't concatenate input with output. No explanation is required. The article is: . Class(0 or 1):

A.6.4 Llama 3.1

Classify the article as Propaganda or Non-Propaganda. Output '1' for Propaganda and '0' for Non-Propaganda. Don't concatenate input with output. No explanation is required. The article is: . Class(0 or 1):

A.6.5 Ministral 8B

Classify the article as Propaganda or Non-Propaganda. Output '1' for Propaganda and '0' for Non-Propaganda. No explanation is required. The article is: . Class (0 or 1):

A.7 Prompts - Abusive Language Detection

A.7.1 Bloomz

You are an abusive language detection model. Labels: 0: non-abusive language 1: abusive language Instructions: To distinguish between abusive and non-abusive language in text samples. Ensure that the model outputs 0 for non-abusive language and 1 for abusive language. Your output should be only 0 or 1 Sentence: Label:

A.7.2 GPT 3.5

"system": "You are an expert in detecting abusive language in the urdu samples Classify the sentence as abusive language or non-abusive language. Output "1" for non-abusive language and "0" for abusive language. No explanation is required. Sentence:

A.7.3 Llama 2

You are a abusive language detection model. Labels: 0: non-abusive language 1: abusive language Instructions: To distinguish between abusive and non-abusive language in text samples. Ensure that the model outputs 0 for non-abusive language and 1 for abusive language. Your output should be only 0 or 1. No explanation Sentence: Label(0 or 1):

A.7.4 Llama 3.1

You are a abusive language detection model.

Labels:

0: non-abusive language 1: abusive language Instructions: To distinguish between abusive and non-abusive language in text samples. Ensure that the model outputs 0 for non-abusive language and 1 for abusive language. Your output should be only 0 or 1. No explanation Sentence: Label(0 or 1):

A.7.5 Ministral 8B

You are a abusive language detection model. Labels: 0: non-abusive language 1: abusive language Instructions: To distinguish between abusive and non-abusive language in text samples. Ensure that the model outputs 0 for non-abusive language and 1 for abusive language. Your output should be only 0 or 1. No explanation Sentence: Label(0 or 1):

A.8 Prompts - Threat Detection

A.8.1 Bloomz

Classify the sentence as threatening or non threatening. Output class "1" for threatening and "0" for non threatening. Sentence: Class(1 or 0):

A.8.2 GPT 3.5

system: You are an expert in detecting threat in the urdu samples Classify the sentence as threatening or non threatening. Output "1" for threatening and "0" for non threatening. Sentence: :

A.8.3 Llama 2

Classify the sentence as threatening or non threatening. Output class "1" for threatening and "0" for non threatening. No explanation required. Sentence: Output(1 or 0):

A.8.4 Llama 3.1

You are a classification model. Your job is to classify the sentences as threatening or non-threatening. Output "1" if the sentence is threatening and "0" if it is non-threatening. No explanation is required

A.8.5 Ministral 8B

Classify the sentence as threatening or non threatening. Output "1" for threatening and "0" for non threatening. No explanation is required. Sentence:

A.9 Prompts - Cyber Bullying Identification

A.9.1 Bloomz

Your task is to classify the nature of cyberbullying with one of the labels: INSULT OFFENSIVE NAMECALLING PROFANE THREAT CURSE NONE Output only label name. no explanation is required. Sentence . Output label:

A.9.2 GPT 3.5

Your task is to classify the nature of cyberbullying with one of the labels: INSULT OFFENSIVE NAMECALLING PROFANE THREAT CURSE NONE Output only label name. no explanation is required. Sentence . Output label:

A.9.3 Llama 2

You are a helpful assistant in classification of cyberbullying. You should always provide answer from given labels without explanation. «/SYS» Human: Sentence . classify the nature of cyber bullying present in sentence with one of the following label: INSULT OFFENSIVE NAMECALLING PROFANE THREAT CURSE NONE Assitant:

A.9.4 Llama 3.1

You are a cyberbullying classification model. Your task is to mark the input as INSULT or OFFENSIVE or NAMECALLING or PROFANE or THREAT or CURSE or NONE in Urdu text samples. Ensure that your output is INSULT or OFFENSIVE or NAMECALLING or PROFANE or THREAT or CURSE or NONE. No explanation is required.

A.9.5 Ministral 8B

Your task is to classify the nature of cyberbullying with one of the labels: INSULT OFFENSIVE NAMECALLING PROFANE THREAT CURSE NONE Output only label name. no explanation is required. Sentence . Output label:

A.10 Prompts - Fake News Detection

A.10.1 Bloomz

You are a fake news detection model. Labels: fake real Instructions: To distinguish between fake news and real news in text samples. Ensure that the model outputs 'fake' for fake news and 'real' for real news. No explanation is required Sentence: Label(fake or real):

A.10.2 GPT 3.5

"system": "You are an expert in detecting fake news in the urdu samples" You are a fake news detection model. Labels: fake real Instructions: To distinguish between fake news and real news in text samples. Ensure that the model outputs 'fake' for fake news and 'real' for real news. No explanation is required Sentence: Label(fake or real):

A.10.3 Llama 2

You are a fake news detection model. Labels: fake real Instructions: To distinguish between fake news and real news in text samples. Ensure that the model outputs 'fake' for fake news and 'real' for real news. No explanation is required Sentence: Label(fake or real):

A.10.4 Llama 3.1

You are a fake news detection model.Output "fake" for fake news and "real" for real news. No explanation is required.

A.10.5 Ministral 8B

You are a fake news detection model. Labels: fake real Instructions: To distinguish between fake news and real news in text samples. Ensure that the model outputs 'fake' for fake news and 'real' for real news. No explanation is required Sentence: Label(fake or real):

A.11 Prompts - Hate Speech Categorization

A.11.1 Bloomz

You are a hate speech classification model. Labels: 0: Non-hate speech 1: Hate speech Instructions: To distinguish between hate speech and non-hate speech in text samples. Ensure that the model outputs 0 for non-hate speech and 1 for hate speech. Your output should be only 0 or 1 Sentence: Label:

A.11.2 GPT 3.5

You are a hate speech classification model. Labels: 0: Non-hate speech 1: Hate speech Instructions: To distinguish between hate speech and non-hate speech in text samples. Ensure that the model outputs 0 for non-hate speech and 1 for hate speech. Your output should be only 0 or 1 Sentence:

A.11.3 Llama 2

You are a hate speech classification model. Labels: 0: Non-hate speech 1: Hate speech Instructions:

To distinguish between hate speech and non-hate speech in text samples. Ensure that the model outputs 0 for non-hate speech and 1 for hate speech. Your output should be only 0 or 1. No explanation is required Sentence: Label(0 or 1):

A.11.4 Llama 3.1

You are a hate speech classification model. Your task is to mark 0 or 1 in Urdu text samples. Ensure that the model outputs '0' for Non-hate text and '1' for Hate text. Ensure that your outputs 0 or 1 only. No explanation is required.

A.11.5 Ministral 8B

You are a hate speech classification model. Your task is to mark 1 or -1 in Urdu text samples. Ensure that the model outputs '1' for Non-hate text and '-1' for hate text. Ensure that your outputs 1 or -1 only. No explanation is required.

A.12 Prompts - Extractive Summarization

A.12.1 Bloomz

You are an extractive summarization model. Label the sentence that you considered is important for Summarization as "1". If you think sentence should not be kept for extractive summary, label it as "0". Sentence Label:

A.12.2 GPT 3.5

"system": "You are an extractive summarization model for Urdu language" You are an extractive summarization model. Label the sentence that you considered is important for Summarization as "1". If you think sentence should not be kept for extractive summary, label it as "0". Sentence Label:

A.12.3 Llama 2

Passage: For extractive summarization, should this passage be kept or discarded? Act as a summarization model. Provide answer only (0 or 1) without explanation. Answer:

A.12.4 Llama 3.1

You are an extractive summarization model. You have to decide whether the given passage should be kept or discarded for summary? Output "1" if the passage should be kept and "0" for discarding (0 or 1) without explanation

A.12.5 Ministral 8B

You are an extractive summarization model. You have to decide whether the given passage should be kept or discarded for summary? Output "1" if the passage should be kept and "0" for discarding (0 or 1) without explanation

A.13 Prompts - Abstractive Summarization

A.13.1 Bloomz

Write summary of the given Urdu meeting. Meeting: . Summary:

A.13.2 GPT 3.5

You are a summarization model. Generate summary for the given meeting minutes in Urdu. Prompt: Meeting minutes: Summary:

A.13.3 Llama 2

You are a summarization model. Your job is to summarize the urdu meeting minutes. Meeting minutes: Summary:

A.13.4 Llama 3.1

You are a summarization model. Generate the summary for the meeting minutes in Urdu.

A.13.5 Ministral 8B

You are a summarization model. Generate the summary for the meeting minutes in Urdu.

A.14 Prompts - Sentiment Analysis

A.14.1 Bloomz

Do the sentimental analysis. Output should be "pos" for positive sentence , "neu" for neutral sentence and "neg" for negative sentence. No explanation is required. Sentence: Label:

A.14.2 GPT 3.5

Do the sentiment analysis. Output should be "pos" for positive sentences , "neu" for neutral sentences and "neg" for negative sentences. No explanation is required. Sentence:

A.14.3 Llama 2

You are a helpful assistant in sentiment analysis. You should always provide answer from given labels without explanation. Human: Do the sentiment analysis. Output "neu" for neutral sentence, "pos" for positive sentence , and "neg" for negative sentence. No explanation is required. Sentence: Assistant:

A.14.4 Llama 3.1

Perform sentiment analysis. Your output should be "pos" for positive sentence , "neu" for neutral sentence and "neg" for negative sentence. No explanation is required.

A.14.5 Ministral 8B

You are a helpful assistant in sentiment analysis. You should always provide answer from given labels without explanation. Tweet: . Perform sentiment analysis on the tweet and answer with one of the following label: -2 for Highly negative -1 for Negative 0 for Neutral 1 for Positive 2 for Highly positive No explanation is required. Output (-2,-1,0,1,2):

A.15 Prompts - Sentiment Analysis (CLE)

A.15.1 Bloomz

Your task is to perform sentiment analysis on the tweets. Labels are: -2 : Highly negative -1 : Negative 0 : Neutral 1 : Positive 2 : Highly positive Output only label name. no explanation is required. Tweet . Output label:

A.15.2 GPT 3.5

"system": "You are an expert in sentiment analysis on urdu tweets Your task is to perform sentiment analysis on the tweets. Labels are: -2 : Highly negative -1 : Negative 0 : Neutral 1 : Positive 2 : Highly positive Output only label name. no explanation is required. Tweet . Output label:

A.15.3 Llama 2

You are a helpful assistant in sentiment analysis. You should always provide answer from given labels without explanation.

Human: Tweet: . Perform sentiment analysis on the tweet and answer with one of the following label: -2 : Highly negative -1 : Negative 0 : Neutral 1 : Positive 2 : Highly positive Assitant:

A.15.4 Llama 3.1

You are a helpful assistant in performing sentiment analysis. Perform sentiment analaysis on the tweet and answer with one of the following label: -2 for Highly negative -1 for Negative 0 for Neutral 1 for Positive 2 for Highly positive Only output -2 to 2 based on the above mentioned scale. No explanation is required

A.15.5 Ministral 8B

You are a helpful assistant in performing sentiment analysis. Perform sentiment analaysis on the tweet and answer with one of the following label: -2 for Highly negative -1 for Negative 0 for Neutral 1 for Positive 2 for Highly positive Only output -2 to 2 based on the above mentioned scale. No explanation is required

A.16 Prompts - Multi-label Emotion Classification

A.16.1 Bloomz

Output the emotion or emotions(if multiple) for the sentence. Emotions: anger, disgust, fear, sadness, surprise, happiness, neutral. You can output multiple emotions as well but should only be the name of the emotions. Output:

A.16.2 GPT 3.5

Output the emotion or emotions(if multiple) for the sentence. Emotions: anger, disgust, fear, sadness, surprise, happiness, neutral. You can output multiple emotions as well but should only be the name of the emotions. Output:

A.16.3 Llama 2

Output the emotion or emotions(if multiple) for the sentence. Emotions: anger, disgust, fear, sadness, surprise, happiness, neutral. You can output multiple emotions as well but should only be the name of the emotions. Output:

A.16.4 Llama 3.1

Output the emotion or emotions(if multiple) for the sentence. Emotions: anger, disgust, fear, sadness, surprise, happiness, neutral. You can output multiple emotions as well but should only be the name of the emotions. No explanation is required.

A.16.5 Ministral 8B

Output the emotion or emotions for the paragraph. Emotions: neutral, happy, fear, sad, anger, love. Your output should only be the name of one of the emotions. Output:

A.17 Prompts - Emotion Classification

A.17.1 Bloomz

Output the emotion or emotions for the paragraph. Emotions: neutral, happy, fear, sad, anger, love. Your output should only be the name of one of the emotions. Output:

A.17.2 GPT 3.5

"system", "content": "You are an expert in emotion recognition in the urdu samples " Output the emotion or emotions for the paragraph. Emotions: neutral, happy, fear, sad, anger, love. Your output should only be the name of one of the emotions. Output:

A.17.3 Llama 2

Output the emotion or emotions for the paragraph. Emotions: neutral, happy, fear, sad, anger, love. Your output should only be the name of only one of the emotions. Output:

A.17.4 Llama 3.1

Output the emotions for the paragraph from one of the following: neutral, happy, fear, sad, anger, love. Your output should only be the name of one of the given emotions. Don't provide any other apart from these six emotions. No explanation is required

A.17.5 Ministral 8B

Output the emotions for the paragraph from one of the following: neutral, happy, fear, sad, anger, love. Your output should only be the name of one of the given emotions. Don't provide any other apart from these six emotions. No explanation is required

A.18 Prompts - Machine Translation

A.18.1 Bloomz

You are an expert translator specialized in translating texts from English to Urdu .Translate the following English sentence to Urdu:

A.18.2 GPT 3.5

"system": "You are an expert translator specialized in translating texts from English to Urdu " Translate the following English sentence to Urdu:

A.18.3 Llama 2

No explanation or notes required. Just translate. English: Urdu:

A.18.4 Llama 3.1

You are an English to Urdu translator. Translate the english sentences into Urdu. No explanation is required. Just translate into Urdu

A.18.5 Ministral 8B

You are an expert translator specialized in translating texts from English to Urdu. Translate the following English sentence to Urdu: "". Provide only the Urdu translation, without any additional text or explanations.

A.19 Prompts - POS Tagging

A.19.1 Bloomz

Your task is to tag POS in input. You will use following Taggig scheme: Tag Proper Noun as NNP ,Tag Common Noun as NN,Tag Personal pronoun as PRP,Tag Demonstrative as PDM,Tag Possessive pronouns as PRS,Tag Reflexive pronouns as PRF,Tag Reflexive Apna as APNA,Tag Relative Personal as PRR,Tag Relative Demonstrative as PRD,Tag Main Verb Infinitive as VBI,Tag Main Verb Finite as VB,Tag Aspectual auxiliaries as AUXA,Tag Progressive auxiliaries as AUXP,Tag Tense auxiliaries as AUXT,Tag Modals auxiliaries as AUXM,Tag Foreign Fragment as FF,Tag Interjection as INJ,Tag Preposition as PRE,Tag Postposition as PSP,Tag Common as SYM,Tag Punctuation as PU,Tag Common as RB,Tag Negation as NEG,Tag Common as PRT,Tag Vala as VALA,Tag Coordinate Conjunction as CC,Tag Subordinate Conjunction as SC,Tag SC Kar as SCK,Tag Presentential as SCP,Tag Ordinal as OD,Tag Fraction as FR,Tag Multiplicative as QM,Tag Adjective as JJ,Tag Quantifier as Q,Tag Cardinal as CD. Your Output should be only one tag corrosponding to input word. no explaination is required. input:

A.19.2 GPT 3.5

"system": "You are an expert in Urdu pos tagging " Your task is to tag POS in input. You will use following Taggig scheme: Tag Proper Noun as NNP ,Tag Common Noun as NN,Tag Personal pronoun as PRP,Tag Demonstrative as PDM,Tag Possessive pronouns as PRS,Tag Reflexive pronouns as PRF,Tag Reflexive Apna as APNA,Tag Relative Personal as PRR,Tag Relative Demonstrative as PRD,Tag Main Verb Infinitive as VBI,Tag Main Verb Finite as VB,Tag Aspectual auxiliaries as AUXA,Tag Progressive auxiliaries as AUXP,Tag Tense auxiliaries as AUXT,Tag Modals auxiliaries as AUXM,Tag Foreign Fragment as FF,Tag Interjection as INJ,Tag Preposition as PRE,Tag Postposition as PSP,Tag Common as SYM,Tag Punctuation as PU,Tag Common as RB,Tag Negation as NEG,Tag Common as PRT,Tag Vala as VALA,Tag Coordinate Conjunction as CC,Tag Subordinate Conjunction as SC,Tag SC Kar as SCK,Tag Presentential as SCP,Tag Ordinal as OD,Tag Fraction as FR,Tag Multiplicative as QM,Tag Adjective as JJ,Tag Quantifier as Q,Tag Cardinal as CD. Your Output should be only one tag corrosponding to input word. no explaination is required. input:

A.19.3 Llama 2

Your task is to tag POS in input. You will use following Taggig scheme: Tag Proper Noun as NNP ,Tag Common Noun as NN,Tag Personal pronoun as PRP,Tag Demonstrative as PDM,Tag Possessive pronouns as PRS,Tag Reflexive pronouns as PRF,Tag Reflexive Apna as APNA,Tag Relative Personal as PRR,Tag Relative Demonstrative as PRD,Tag Main Verb Infinitive as VBI,Tag Main Verb Finite as VB,Tag Aspectual auxiliaries as AUXA,Tag Progressive auxiliaries as AUXP,Tag Tense auxiliaries as AUXT,Tag Modals auxiliaries as AUXM,Tag Foreign Fragment as FF,Tag Interjection as INJ,Tag Preposition as PRE,Tag Postposition as PSP,Tag Common as SYM,Tag Punctuation as PU,Tag Common as RB,Tag Negation as NEG,Tag Common as PRT,Tag Vala as VALA,Tag Coordinate Conjunction as CC,Tag Subordinate Conjunction as SC,Tag SC Kar as SCK,Tag Presentential as SCP,Tag Ordinal as OD,Tag Fraction as FR,Tag Multiplicative as QM,Tag Adjective as JJ,Tag Quantifier as Q,Tag Cardinal as CD. Your Output should be only one tag corrosponding to input word. no explaination is required. input:

A.19.4 Llama 3.1

Your task is to tag POS in input. You will use following Taggig scheme: Tag Proper Noun as NNP ,Tag Common Noun as NN,Tag Personal pronoun as PRP,Tag Demonstrative as PDM,Tag Possessive pronouns as PRS,Tag Reflexive pronouns as PRF,Tag Reflexive Apna as APNA,Tag Relative Personal as PRR,Tag Relative Demonstrative as PRD,Tag Main Verb Infinitive as VBI,Tag Main Verb Finite as VB,Tag Aspectual auxiliaries as AUXA,Tag Progressive auxiliaries as AUXP,Tag Tense auxiliaries as AUXT,Tag Modals auxiliaries as AUXM,Tag Foreign Fragment as FF,Tag Interjection as INJ,Tag Preposition as PRE,Tag Postposition as PSP,Tag Common as SYM,Tag Punctuation as PU,Tag Common as RB,Tag Negation as NEG,Tag Common as PRT,Tag Vala as VALA,Tag Coordinate Conjunction as CC,Tag Subordinate Conjunction as SC,Tag SC Kar as SCK,Tag Presentential as SCP,Tag Ordinal as OD,Tag Fraction as FR,Tag Multiplicative as QM,Tag Adjective as JJ,Tag Quantifier as Q,Tag Cardinal as CD. Your Output should be only one tag corrosponding to input word. no explaination is required. input:

A.19.5 Ministral 8B

Your task is to tag POS in input. You will use following Taggig scheme: Tag Proper Noun as NNP ,Tag Common Noun as NN,Tag Personal pronoun as PRP,Tag Demonstrative as PDM,Tag Possessive pronouns as PRS,Tag Reflexive pronouns as PRF,Tag Reflexive Apna as APNA,Tag Relative Personal as PRR,Tag Relative Demonstrative as PRD,Tag Main Verb Infinitive as VBI,Tag Main Verb Finite as VB,Tag Aspectual auxiliaries as AUXA,Tag Progressive auxiliaries as AUXP,Tag Tense auxiliaries as AUXT,Tag Modals auxiliaries as AUXM,Tag Foreign Fragment as FF,Tag Interjection as INJ,Tag Preposition as PRE,Tag Postposition as PSP,Tag Common as SYM,Tag Punctuation as PU,Tag Common as RB,Tag Negation as NEG,Tag Common as PRT,Tag Vala as VALA,Tag Coordinate Conjunction as CC,Tag Subordinate Conjunction as SC,Tag SC Kar as SCK,Tag Presentential as SCP,Tag Ordinal as OD,Tag Fraction as FR,Tag Multiplicative as QM,Tag Adjective as JJ,Tag Quantifier as Q,Tag Cardinal as CD. Your Output should be only one tag corrosponding to input word. no explaination is required. input: