Generalists vs. Specialists: Evaluating Large Language Models for Urdu

Samee Arif and **Abdul Hameed Azeemi** and **Agha Ali Raza**
{samee.arif, abdul.azeemi, agha.ali.raza}@lums.edu.pk
Lahore University of Management Sciences

Awais Athar
awais@ebi.ac.uk
EMBL European Bioinformatics Institute

Abstract

In this paper, we compare general-purpose models, `GPT-4-Turbo` and `Llama-3-8b`, with special-purpose models—`XLM-Roberta-large`, `mT5-large`, and `Llama-3-8b`—that have been fine-tuned on specific tasks. We focus on seven classification and seven generation tasks to evaluate the performance of these models on Urdu language. Urdu has 70 million native speakers, yet it remains underrepresented in Natural Language Processing (NLP). Despite the frequent advancements in Large Language Models (LLMs), their performance in low-resource languages, including Urdu, still needs to be explored. We also conduct a human evaluation for the generation tasks and compare the results with the evaluations performed by `GPT-4-Turbo`, `Llama-3-8b` and `Claude 3.5 Sonnet`. We find that special-purpose models consistently outperform general-purpose models across various tasks. We also find that the evaluation done by `GPT-4-Turbo` for generation tasks aligns more closely with human evaluation compared to the evaluation the evaluation done by `Llama-3-8b`. This paper contributes to the NLP community by providing insights into the effectiveness of general and specific-purpose LLMs for low-resource languages.

1 Introduction

In recent years the introduction of LLMs including `GPT` (Brown et al., 2020; OpenAI, 2024) and `Llama` (Touvron et al., 2023a,b) has led to a significant advancement in NLP. However, expanding the reach of NLP to low-resource languages is crucial for advancing multilingual AI systems and promoting technological inclusivity. Urdu, with over 70 million native speakers, stands as a significant yet underserved language in the NLP domain (Blasi et al., 2022).

For this study, we classify LLMs into two distinct categories:

1. **Generalists**: General-purpose models capable of performing a wide variety of tasks. We will use `GPT-4-Turbo` (abbreviated as `GPT`) trained on dataset up to Dec 2023 and `Llama-3-8b` (abbreviated as `Llama`) as the generalist models.

2. **Specialists**: Special-purpose models fine-tuned to perform specific tasks. We use `XLM-Roberta-large` (abbreviated as `XLM-R`) (Conneau et al., 2020), `mT5-large` (abbreviated as `mT5`) (Xue et al., 2021), and a fine-tuned version of `Llama-3-8b` (abbreviated as `Llama-FT`) as the specialist models.

We present a comprehensive evaluation of generalist and specialist models for classification and generation tasks, exploring their strengths and limitations. Table 1 outlines the sub-tasks associated with both categories, illustrating the scope of our evaluation.

Classification	**Generation**
Sentiment Analysis	Question Answering
Abuse Detection	Summarization
Sarcasm Detection	Paraphrasing
Fake News Detection	Transliteration
Topic Classification	Translation (en-ur)
Part-of-Speech Tagging	Translation (ur-en)
Named-entity Recognition	AI Assistant

Table 1: Sub-tasks for classification and generation.

In this paper, we aim to answer: (1) how each category of models performs on Urdu language tasks, and (2) which model type is more effective in practical applications for Urdu-speaking users. Specifically, we seek to determine if the added specialization of the specialist models translates into significant performance gains over the generalist models in the defined tasks.

Our contributions can be summarized as follows:

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7263–7280

1. We fine-tune XLM-R, mT5, and Llama on classification and generation tasks to optimize their performance for the defined tasks.

2. We compare the performance of both generalist and specialist models on a smaller, controlled test set consisting of min(1000, len(dataset["test"])) samples, and provide a detailed comparison using various metrics. The exact size of each test set is given in Appendix D. Given the inherent challenges of working with low-resource languages like Urdu, our test set size reflects the current limitations in available data.

3. We conduct a human evaluation of the outputs from the generation tasks and compare these results with the automated evaluations performed by GPT, Llama and Claude 3.5 Sonnet (abbreviated as Claude).

The code, model outputs, and the human and LLM evaluations are publicly available on GitHub[1].

2 Related Work

In recent years there has been a growing interest in the performance of LLMs across various languages. The MEGA benchmark (Ahuja et al., 2023) evaluates 16 NLP datasets across 70 languages. They compare the performance of BLOOMZ, GPT models, and State of the Art (SOTA) non-autoregressive models. MEGAVERSE (Ahuja et al., 2024) builds on top of the MEGA benchmark and evaluates the non-English capabilities of GPT-3.5-Turbo, GPT-4, PaLM2, Gemini-Pro, Mistral, Llama-2, and Gemma. IndicGenBench (Singh et al., 2024) evaluates LLMs on user-facing generation tasks across a set of 29 Indic languages. They perform evaluation on both proprietary and open-source LLMs including GP-3.5, GPT-4, PaLM-2, mT5, Gemma, BLOOM, and Llama. They cover the following tasks: cross-lingual summarization, machine translation, and cross-lingual question-answering. Zhao et al. (2024) conduct an evaluation of Llama's response quality based on LLM-Eval (Zhang et al., 2023), a benchmark comprising instruction tasks from 17 categories. Mujadia et al. (2024) presents a comprehensive translation evaluation using Llama for English and Indian languages. They found that LLM-based evaluator achieves a comparable score with human judgement.

Khondaker et al. (2023a) presents a in-depth evaluaton of BLOOMZ, ChatGPT and specialist models AraT5 and MARBERTv2. They evaluated these models for Arabic on 44 distinct language understanding and generation tasks. They also present a comparison between the human evaluation and GPT-4 evaluation. Additionally, they observed that the English prompt works better than the Arabic prompt. Abdelali et al. (2024) provides a benchmark of LLMs against SOTA models for Arabic NLP. They evaluated GPT-3.5-Turbo, GPT-4, BLOOMZ, and Jais-13b-Chat across 33 tasks using 61 publicly available datasets, resulting in 330+ experimental setups. They observed that SOTA models generally outperform LLMs in zero-shot learning, larger models with few-shot learning techniques significantly reduce the performance gaps.

Existing multilingual NLP benchmarks often lack extensive language-specific evaluation and comparison against task-specific models. Tahir et al. (2024) address this gap by evaluating GPT-3.5-Turbo, Llama-2-7B-Chat, and Bloomz (3B and 7B) across 14 tasks using 15 Urdu datasets in a zero-shot setting, comparing their performance against task-specific models. Their findings reveal that task-specific models (Support Vector Machine (SVM), Decision Tree, and m-BERT, etc) generally outperform the mentioned LLMs in Urdu NLP tasks with zero-shot learning. However, they do not perform few-shot prompting, chain-of-thought prompting, or LLM fine-tuning.

3 Datasets

3.1 Classification

Sentiment Analysis. We use the Urdu IMDB sentiment analysis dataset[2], which is a translated version of the original IMDB dataset (Maas et al., 2011). It is translated using Google Translator, and comprises 50,000 movie reviews.

Abuse Detection. We use the dataset by Akhter et al. (2020) which has 2,171 entries and the dataset by Amjad et al. (2022) which consists of 3,502 entries.

Sarcasm Detection. For this task we use Urdu Sarcastic Tweets Dataset (Khan and Na-

[1] https://github.com/sameearif/Generalists-vs-Specialists

[2] https://github.com/urduhack/resources/releases/tag/imdb_urdu_reviews_v1.0.0

jeeb, 2023) which consists of 19,955 tagged tweets.

Fake News Detection. We use the fake news dataset by Amjad et al. (2020) which is a mixture of real and translated data. It comprises 1,300 labeled news articles.

Topic Classification. For this task, we use a dataset[3] consisting of 137,161 news headlines categorized into the following topics: business, entertainment, health, politics, science, sports, world and other.

Part-of-Speech (PoS) Tagging. For this task, we use the Universal Dependencies (Nivre et al., 2020) dataset, which consists of 5,130 sentences, annotated with a PoS tag for every word. For `GPT` and `Llama`, we wrap the word for which we want to predict the PoS tag in `<hl>` tags. The structure of the data is given in Figure 1.

Input:
کمیٹی نے اپنی رپورٹ میں <hl>بتایا<hl> کہ ایک یونیورسٹی کے تحت مختلف کالجس میں علاحدہ علاحدہ فیس مقرر رہنے کی وجہ سے فیس ری۔ ایمبرسمنٹ میں ٹیکنیکل خامیاں پیدا ہو رہی ہیں۔

Label: Verb

Figure 1: PoS Data Structure for Llama

Named Entity Recognition (NER). We use the dataset[4] available on Hugging Face which consists of 33,748 sentences, annotated with a NER tag for every word. For `GPT` and `Llama`, we applied the same data structuring method as done in the PoS tagging task.

3.2 Generation

Given our resource constraints, we could not set the maximum sequence length for model training to the longest entries in each dataset. Instead, we filter the datasets to include only those entries that fell within a manageable maximum length of 2048. This approach allowed us to optimize the training process and ensure efficient use of computational resources while still maintaining a representative sample of the data.

Question-Answering. We use three datasets for question-answering: (1) UQA (Arif et al., 2024a), consisting of 88,829 answerable questions; (2) UQuAD[5], containing 139 questions; and (3) Wiki-UQA[6], a manually generated dataset from Wikipedia articles, comprising 210 questions.

Summarization. For this task, we use the XSUMUrdu (Munaf et al., 2023) dataset, selecting a subset of 76,626 entries based on the maximum length used during model training.

Paraphrasing. For paraphrasing we use the dataset[7] available on Hugging Face. We select 387,004 entries from the dataset based on the maximum length used while training the models.

Transliteration. We use the Dakshina dataset (Roark et al., 2020) from which we select 11,464 sentences based on the maximum length used while training the models.

Translation. We use OPUS-100 (Zhang et al., 2020) (Tiedemann, 2012) for English-to-Urdu and Urdu-to-English translation. We select 755,526 sentences from this dataset based on the maximum length used while training the models.

AI Assistant. We use UrduAssistant (Khalil, 2023) dataset which is translated to Urdu from English and has 67,017 prompts.

4 Methodology

4.1 Experimental Design

We utilize a controlled test set consisting of `min(1000, len(dataset["test"]))` samples for each task except AI assistant, ensuring that the evaluation is both comprehensive and cost-efficient. For AI assistant we do human evaluation of 50 samples. We ensure that the test dataset is as balanced as possible, with an effort to achieve equal representation of different classes within each task. We use Macro-F_1 Score to evaluate all the classification tasks, SQuAD F_1 (Rajpurkar et al., 2018) (Rajpurkar et al., 2016) to evaluate question-answering, SacreBLEU (Post, 2018) for paraphrasing, transliteration and translation, ROUGE-L (with word-level

[3] https://github.com/mwaseemrandhawa/Urdu-News-Headline-Dataset

[4] https://huggingface.co/datasets/mirfan899/urdu-ner

[5] https://github.com/ahsanfarooqui/UQuAD---Urdu-Question-Answer-Dataset/tree/main

[6] https://huggingface.co/datasets/uqa/Wiki-UQA

[7] https://huggingface.co/datasets/mwz/ur_para

tokenization) (Lin, 2004) for summarization and Wins (the number of times a model is ranked one by the evaluator) for AI assistant evaluation.

We fine-tune Llama and XLM-R for each classification task separately, ensuring that each model is specifically optimized for its respective task. Similarly, for generation tasks, we fine-tune the mT5 and Llama separately for each task. For the mT5 models, we use a learning rate of $5e^{-5}$. The Llama models are fine-tuned with a learning rate of $2e^{-4}$ for both generation and classification tasks. In the case of XLM-R, we use a learning rate of $5e^{-6}$. We use LoRA (Hu et al., 2021) to fine-tune int4 quantized Llama. The batch size, number of epochs each model is trained for and LoRA config for Llama is given in Appendix E. After fine-tuning all the models we perform evaluation on the test dataset.

To evaluate the performance of the generalist models, we design a series of experiments. We use GPT and Llama as our generalist models. GPT is chosen due its top performance on the LMSYS chatbot arena (Chiang et al., 2024) as of March 1st, 2024 when we started our research. Our experimental setup is as follows: We conduct experiments under zero-shot, three-shot, and six-shot settings for both generation and classification tasks. The examples for the three-shot and six-shot scenarios are selected from the training dataset of each task. Specifically, the examples are carefully selected by a human expert to ensure a representative and balanced sample, avoiding any unintended bias in the selection process. We also conduct experiments with Chain-of-Thought (CoT) reasoning for classification tasks in the six-shot setting. We create CoT reasoning for the six selected examples from the training dataset. These examples along with their CoT is given as a few-shot prompt to the model. Figure 2 presents an example of generated CoT reasoning. For all the evaluations, the temperature and nucleus sampling for GPT-4 is set to 1.0, which is the default for the GPT API. For Llama, the temperature is set to 0.6 and nucleus sampling is set to 0.9, i.e., the values used in the original code-base for Llama[8].

4.2 Prompt Design

Khondaker et al. (2023b) observe in their ChatGPT evaluation for Arabic that an English prompt performs better than the Arabic one. Therefore, following this study, we decide to use English prompts

[8] https://github.com/meta-llama/llama3/blob/main/example_chat_completion.py

Input:
جہاں سے بھی سے گزرو گے خوش رہو گے

CoT:
The phrase translates to "Wherever you go, you will be happy." This statement is not abusive as it conveys a positive and well-wishing sentiment. The language is kind and encouraging, with no offensive or derogatory terms, making it a non-abusive expression.

Figure 2: CoT example from abuse detection

for our evaluations of Urdu tasks as well. In the prompt templates, the following placeholders are used:

1. ROLE: It specifies the persona of the LLM. For example, it could be sentiment classifier, abuse detector, sarcasm detector, etc.

2. TASK DESCRIPTION: It provides a brief description of what the task is and what the model is expected to do.

3. LABEL LIST: It lists the possible labels the model can assign to the input text. For example, it could be `['positive', 'negative']` for sentiment analysis.

Figure 3 shows the prompt template for the classification task without CoT, and Figure 4 shows an example prompt for it. Figure 5 shows the CoT prompt template for the classification task. Figure 6 presents the prompt template for the generation task, and Figure 7 provides an example of it. Appendix G contains all the prompts for classification and generation tasks.

```
You are an Urdu ROLE. The input text
should be labeled according to TASK
DESCRIPTION. The label list is:
LABEL LIST
ALWAYS RETURN JSON OBJECT IN FOLLOWING
FORMAT ONLY: {"label": ...}
```

Figure 3: Classification Prompt Template

4.3 Human Evaluation

We select a subset of size 50 (as done by Khondaker et al. (2023a) for Arabic) from the test dataset of each task. For human evaluation, two annotators (native Urdu speakers) are presented with

```
You are an Urdu sentiment classifier.
The input text should be labeled
according to its sentiment. The label
list is:
['positive', 'negative']
ALWAYS RETURN JSON OBJECT IN FOLLOWING
FORMAT ONLY: {"label": ...}
```

Figure 4: Classification Prompt Example

```
You are an Urdu ROLE. The input text
should be labeled according to TASK
DESCRIPTION. The label list is:
LABEL LIST
Use chain of thought (cot) to reason
your answer. ALWAYS RETURN JSON OBJECT
IN FOLLOWING FORMAT ONLY:
{"cot": ..., "label": ...}
```

Figure 5: Classification Prompt Template (CoT)

anonymized outputs of Llama, Llama-FT, mT5 and GPT. They are asked to rank them from one to four based on the criteria in Appendix G, allowing multiple models to have the same rank. We prompt GPT, Llama and Claude with the same criteria, to assign a score to the outputs of the mentioned tasks, and then ranking is done based on this score. Since both GPT and Llama responses are being evaluated, we also include Claude as an additional evaluator because of the bias when LLMs assess their own outputs (Arif et al., 2024b). We compare ranking done by human annotators, and the three LLMs using Krippendorff's alpha to determine the inter-rater reliability to determine the determine whether LLMs are effective evaluators for Urdu tasks.

5 Evaluation and Discussion

5.1 Classification

We present the evaluation of the generalists (GPT, Llama) and the specialists (Llama-FT, XLM-R) for the classification tasks in Table 2. We observe that Llama-FT (fine-tuned) model achieves the highest scores in four tasks (sentiment analysis, topic classification, PoS tagging, and NER tagging), while XLM-R outperforms other models in three tasks (abuse detection, sarcasm detection, and fake news detection). GPT does not perform better than Llama-FT and XLM-R. For certain tasks like NER and PoS tagging, CoT reasoning leads to a better

```
You are a ROLE. Your task is to TASK
DESCRIPTION.
ALWAYS RETURN JSON OBJECT IN FOLLOWING
FORMAT ONLY: {"label": ...}
```

Figure 6: Generation Prompt Template

```
You are a text summarizer. Your task is
to summarize the given Pakistani Urdu
text in 1 to 2 sentences.
ALWAYS RETURN JSON OBJECT IN FOLLOWING
FORMAT ONLY: {"summary": ...}
```

Figure 7: Generation Prompt Example

Macro-F_1 score. We now discuss the performance on each classification task in detail.

Sentiment Analysis. For sentiment analysis, Llama-FT achieves the highest score with a Macro-F_1 of 95.30, outperforming XLM-R and GPT. GPT's performance improves with more shots, reaching a Macro-F_1 of 94.90 with 6-shot learning. The CoT setting provides a slight increase to 94.60. However the Macro-F_1 for Llama decreases with few-shot and CoT prompting. The small difference in the Macro-F_1 between Llama-FT and GPT indicates the effectiveness of GPT for Urdu sentiment analysis without requiring task-specific fine-tuning.

NER Tagging. For NER tagging, Llama-FT achieves the highest score with a Macro-F_1 of 90.90, significantly outperforming XLM-R and GPT. The highest score achieved by GPT is 63.95, and by Llama is 53.96, both using CoT reasoning. We uncover the differences in accuracy for various entities for GPT and Llama-FT in Figure 8. We observe a drop in the recognition accuracy of Person entities by GPT, with 95 correctly recognized compared to 139 by Llama-FT. We notice a similar trend for Organization entities, with 70 classified correctly compared to 134 by Llama-FT. This suggests that NER fine-tuning on a specialized Urdu dataset improves the model's ability to recognize Person and Organization entities in Urdu text.

Abuse Detection. In the task of abuse detection, XLM-R leads with a Macro-F_1 of 90.92, followed closely by Llama-FT at 89.15. GPT achieves its highest performance of 89.01 with 3-shot

	GPT				Llama				XLM-R	Llama-FT
Task	0	3	6	CoT	0	3	6	CoT		
Sentiment Analysis	90.98	91.17	94.90	94.60	87.88	60.62	59.71	57.54	92.90	**95.30**
Abuse Detection	86.27	89.01	88.71	87.62	44.73	65.85	71.64	75.82	**90.92**	89.15
Sarcasm Detection	58.17	69.18	66.56	65.47	44.94	34.25	50.67	49.75	**84.37**	81.48
Fake News Detection	78.88	75.95	78.14	76.45	66.36	63.48	71.45	40.98	**84.99**	72.14
Topic Classification	76.05	73.54	74.57	70.43	54.08	53.02	53.68	49.24	84.49	**84.74**
PoS Tagging	53.31	51.51	54.17	54.61	25.80	34.32	33.85	41.62	65.41	**67.55**
NER Tagging	61.96	62.18	62.98	63.95	42.13	40.80	45.29	53.96	70.41	**90.90**

Table 2: We report Macro-F_1 score for each classification task. `GPT-4-Turbo` and `Llama-3-8b` is evaluated in 0-shot, 3-shot, 6-shot and 6-shot with Chain-of-Thought settings. FT stands for fine-tuned.

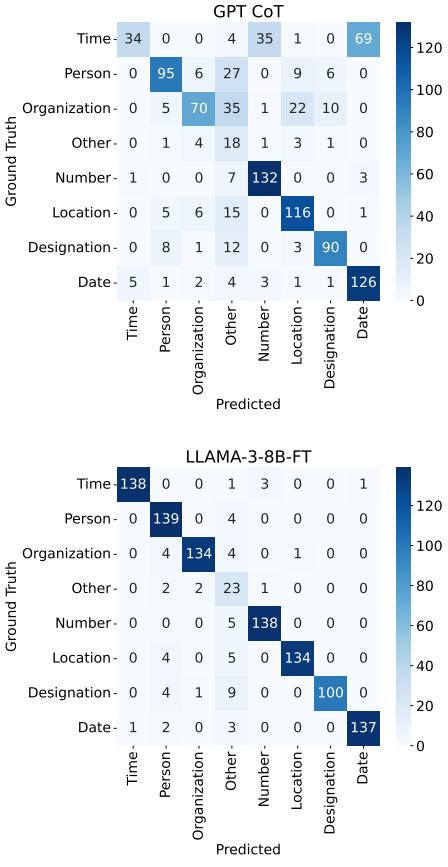

Figure 8: Performance of `GPT-CoT` and `Llama-FT` on recognizing different entities in the NER task.

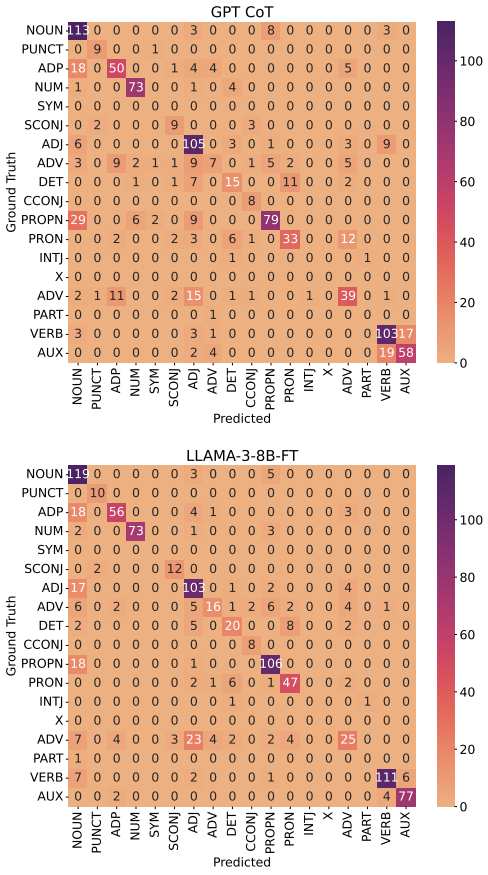

Figure 9: Performance of `GPT-CoT` and `Llama-FT` on recognizing different parts of speech in the POS task.

prompting, while `Llama` reaches a peak of 75.82 using CoT reasoning.

Sarcasm Detection. `XLM-R` performs best in sarcasm detection with a Macro-F_1 score of 84.37. `Llama-FT` also shows strong performance with 81.48. `GPT` achieves its best performance at 69.18 with 3-shot prompting, while `Llama` performs best at 50.67 with 6-shot prompting. The large difference between the Macro-F_1 scores of generalist and specialist models indicates that sarcasm detection in Urdu can be challenging without task-specific fine-tuning.

Fake News Detection. For fake news detection, `XLM-R` achieves the highest score with a Macro-F_1 of 84.99, while `Llama-FT` follows with 72.14. `GPT`'s performance is relatively consistent across different shot settings, with its highest score being 78.88 in the 0-shot setting. `Llama`'s highest score is 71.45 at 6-shot setting. The large margin by `XLM-R` indicates its effectiveness in discerning fake news.

Topic Classification. `Llama-FT` excels in topic classification with a Macro-F_1 score of 84.74,

Task	Metric	GPT			Llama			mT5	Llama-FT
		0	3	6	0	3	6		
Question-Answering	SQuAD-F_1	66.28	72.55	**72.78**	69.89	71.65	71.62	69.66	70.42
Summarization	ROUGE-L	22.54	23.35	23.34	25.76	25.84	26.17	30.72	**31.01**
Paraphrasing	SacreBLEU	3.59	3.92	4.01	2.42	6.40	6.11	**11.98**	10.17
Transliteration	SacreBLEU	30.93	32.38	32.09	12.37	18.51	21.35	**40.23**	37.95
Translation (en-ur)	SacreBLEU	12.07	12.62	11.59	5.14	6.56	5.83	**18.35**	14.63
Translation (ur-en)	SacreBLEU	16.29	18.05	19.18	7.83	13.11	12.32	21.55	**29.18**

Table 3: We report SQuAD-F_1, ROUGE-L or SacreBLEU depending on the generation tasks. GPT-4-Turbo and Llama-3-8b is evaluated in 0-shot, 3-shot and 6-shot setting.

significantly higher than the other models. XLM-R also performs well with 84.49.

PoS Tagging. In PoS tagging, Llama-FT leads with a Macro-F_1 of 67.55, followed by XLM-R at 65.41. GPT's highest score is 53.31 in the 0-shot setting and Llama's highest score is 41.62 with CoT reasoning. To understand the significant difference in Macro-F_1 between GPT and Llama-FT, we study the individual class performance (Figure 9). We find that GPT struggles with correctly tagging proper nouns, pronouns, and auxiliaries, while Llama-FT is able to identify most of them correctly, suggesting that task-specific fine-tuning improves tagging performance for these parts of speech.

5.2 Generation

We present the evaluation of the generalists and the specialists for the generation tasks in Table 3 and Table 4. We observe that the fine-tuned Llama-FT model achieves the highest scores in summarization and translation from Urdu to English. On the other hand GPT shows a consistent performance across all the tasks with its best results often appearing in the 6-shot setting. The mT5 model also demonstrates strong performance in tasks such as transliteration and paraphrasing, benefiting from its extensive multilingual training on translation tasks. We now discuss the performance on each generation task in detail.

5.2.1 Quantitative Evaluation

Question-Answering. All the models show similar performance on question-answering task. GPT achieves the highest score with a SQuAD-F1 of 72.78 in the 6-shot setting. Llama closely follows with a score of 71.65 in 3-shot setting. Llama-FT has a lower score of 70.42, indicating that the translated UQA dataset is not very effective in improving the performance of the model. mT5 also performs well with a score of 69.66. The consistent improvement of GPT with increasing shots suggests its capability to use more context effectively.

Summarization. In the summarization task, Llama-FT achieves the highest ROUGE-L score of 31.01, outperforming Llama and GPT. mT5 also shows strong performance with a score of 30.72. Non-fine-tuned Llama outperforms GPT in all few-shot settings. GPT's best performance is in the 3-shot setting with a score of 23.35 while Llama's best is at 6-shot setting with a score of 26.17. The fine-tuning of Llama contributes to its superior performance in capturing and summarizing Urdu content effectively.

Paraphrasing. For paraphrasing, mT5 achieves the highest SacreBLEU score of 11.98. This indicates the advantage mT5 has due to its massive multilingual pre-training. Llama-FT follows with a SacreBLEU score of 10.17. Llama's best score is 6.40 at 3-shot setting outperforming GPT's best score of 4.01.

Transliteration. In the task of transliteration, mT5 leads with a SacreBLEU score of 40.23, followed by Llama-FT at 37.95. GPT's performance peaks at 32.38 with 3-shot learning. Llama shows poor performace in this task with the highest score of 21.35 at 6-shot setting. Figure 10 shows the words with the highest mismatches in the transliterated text. To count these mismatches, we first tokenize the transliterated sentences and then count the instances where the predicted word differs from the ground truth. Smaller words such as "ye," "ke," "aik," and "wo" pose challenges for GPT, resulting in higher mismatch counts. In contrast, mT5 demonstrates lower mismatches for these words.

Task	GPT's Wins		Llama's Wins		mT5's Wins		Llama-FT's Wins	
	Annot. 1	Annot. 2	Annot. 1	Annot. 2	Annot. 1	Annot. 2	Annot. 1	Annot. 2
Summarization	41	34	9	17	7	12	4	11
Paraphrasing	42	39	12	15	4	3	4	4
Transliteration	40	38	5	6	13	21	19	21
Translation (en-ur)	46	46	3	1	7	10	13	10
Translation (ur-en)	40	45	12	10	14	15	18	15
AI Assistant	49	49	3	3	0	0	1	1

Table 4: We report the number of times each model is ranked 1 by each annotator for the generation tasks.

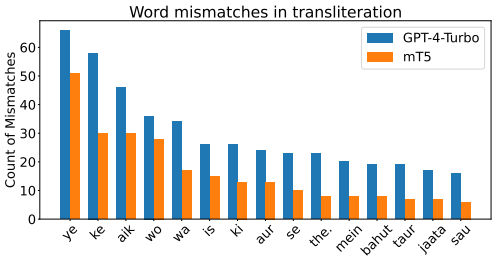

Figure 10: Comparison between mismatches in the words predicted by mT5 and GPT for the transliteration task.

Translation (en-ur). For English to Urdu translation, mT5 achieves the highest SacreBLEU score of 18.35, indicating its proficiency in translating English to Urdu. Llama-FT follows with a score of 14.63. GPT's performance is consistent, with its highest score being 12.62 in the 3-shot setting. Llama under-performs with a least highest score of 6.56 with 3-shot prompting. The superior performance of mT5 is likely due to the inclusion of high-quality Urdu data in the multilingual C4 corpus used for its pre-training (Xue et al., 2021).

Translation (ur-en). In Urdu to English translation, Llama-FT excels with a SacreBLEU score of 29.18, outperforming other models. Surprisingly, contrary to the results in en-ur translation, mT5 shows a lower SacreBLEU of 21.55 compared to Llama-FT. GPT's highest score is 19.18 in the 6-shot setting and Llama's higest score is 13.11 in the 3-shot setting.

5.2.2 Qualitative Evaluation

AI Assistant. GPT outperforms all the models with 49 out of 50 wins as shown in Table 4. Llama-FT is ranked one 0 times according to both annotators and mT5 is ranked number one only 1 time. The low performance of fine-tuned models, such as Llama-FT and mT5, can be attributed to the fact that the fine-tuning dataset was translated from English.

Based on quantitative evaluation, GPT only performs the best in the question-answering task based on the quantitative evaluation provided in Table 3. However, Table 4 shows that GPT performs the best across all other tasks as well, according to qualitative evaluations by human annotators with more than 90% win-rate for most of the tasks. Figure 11 presents the number of times each rank was assigned to GPT for Urdu to English translation. Appendix F contains presents the rank counts for the other tasks.

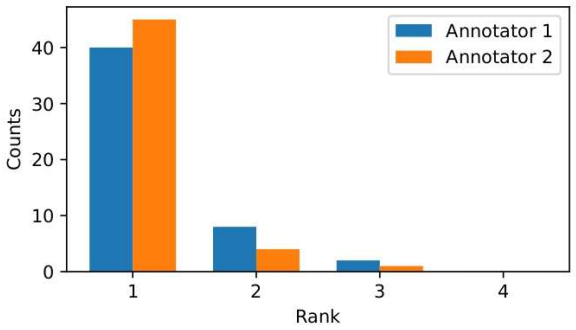

Figure 11: Rank counts for GPT in human evaluation of Urdu to English translation.

Based on human evaluation LLama outperforms mT5 and even the fine-tuned Llama-FT on summarization, paraphrasing and AI assistant task. On transliteration and both translation tasks Llama-FT outperforms mT5. This discrepancy between quantitative and qualitative evaluations highlights an important issue: the translation of datasets from other languages may introduce inconsistencies or artifacts that impact the quantitative metrics, especially for tasks like summarization, paraphrasing, and story generation. This observation stresses the need for the creation of native Urdu datasets, rather than relying on translations of existing datasets. It also reflects the limitations and lower quality of the currently available Urdu datasets, which further exacerbates these challenges.

5.3 Factors Contributing to Improved Performance

Task-specific fine-tuning on Urdu datasets significantly improves the performance of LLMs on Urdu tasks, as evidenced by the superior results of fine-tuned Llama, XLM-R, and mT5 on both classification and generation tasks. Additionally, the inclusion of high-quality Urdu data in pre-training datasets, such as the clean multilingual C4 corpus, enhances downstream performance, with mT5 achieving higher SacreBLEU scores on transliteration, translation, and paraphrasing tasks compared to GPT and fine-tuned Llama. Providing more context in prompts also boosts performance, as shown by GPT's improvement with increasing shots (0, 3, 6). Moreover, high-quality data for Indic languages like Hindi in pre-training improves performance on Urdu, especially for models like XLM-R, which excel in cross-lingual tasks, outperforming others in abuse detection, sarcasm detection, and fake news detection. The lack of native Urdu datasets poses a significant challenge for Urdu NLP, as most available datasets are translated from other languages.

5.4 Human Evaluators vs. LLMs

We compare human evaluation and LLM-based evaluation for summarization, paraphrasing, transliteration, English to Urdu translation, and Urdu to English translation, presented in Table 5.

Task	A	B	C	D
Summarization	0.684	0.504	0.502	0.624
Paraphrasing	0.710	0.592	0.471	0.565
Transliteration	0.694	0.510	0.253	0.603
Translation (en-ur)	0.728	0.592	0.392	0.515
Translation (ur-en)	0.730	0.474	0.307	0.510
AI Assistant	0.894	0.767	0.721	0.775

Table 5: **A**: Krippendorff's alpha between annotator 1 and annotator 2. **B**: Alpha between both annotators and GPT. **C**: Alpha between both annotators and Llama. **D**: Alpha between both annotators and Claude.

For each task, the Krippendorff's alpha value for human evaluation exceeds 0.67, which, according to Krippendorff's interpretation is sufficient for a tentative conclusion to be drawn. Table 5 illustrates that the inclusion of GPT's evaluation, Llama's evaluation or Claude's evaluation significantly reduces the alpha values meaning that the annotations done by GTP, Llama and Claude have a lower degree of agreement with the human annotators. For summarization, transliteration, Urdu-to-English translation, and AI assistant tasks, Claude has the highest agreement with humans, at 0.624, 0.604, 0.510, and 0.775, respectively. This showcases that the LLMs exhibit bias when the evaluated responses include their own generated content. For paraphrasing and Urdu-to-English translation tasks, GPT has the highest alpha values of 0.592 and 0.474, respectively. There is a higher agreement between the GPT rankings and human rankings as compared to the agreement between Llama rankings and human rankings.

6 Conclusion

In this paper, we present a comprehensive evaluation of generalist models and specialist models on 7 classification tasks and 7 generation tasks for Urdu NLP. Our evaluation covers prompting techniques such as few-shot, CoT reasoning as well as the fine-tuning of LLMs. We found that specialist models quantitatively outperformed generalist models on 12 out of the 14 tasks. The results highlight the importance of fine-tuning models to achieve higher performance in domain-specific applications in a low-resource setting. However, generalist models, such as GPT, showcased better performance in the human evaluation of the generation tasks, highlighting the importance of qualitative evaluation in accurately assessing model performance. We also performed a LLM based evaluation of the outputs of the models for the generation tasks. The low agreement between the rankings done by LLMs and human rankings shows that LLMs struggle when it comes to low-resource language understanding. Furthermore, the evaluation indicates a bias in LLMs when assessing responses that include their own outputs.

7 Future Work

One avenue for future research is to explore other strong general-purpose models (e.g., GPT-4o, GPT-4o1 and Claude) and expand the scope of the evaluation to more Urdu NLP tasks. Additionally, using Retrieval-Augmented Generation (RAG) to find examples from training dataset for few-shot prompting would be an interesting experiment to enhance the performance of generalist models. In conclusion, while specialist models currently hold an edge in classification tasks performance, generalist models' adaptability and better performance in generation tasks remains valuable, and continuous advancements in LLMs promise further improvements for low-resource NLP.

8 Limitations

While our study provides valuable insights into the performance of generalist and specialist models for Urdu NLP tasks, it is important to acknowledge several limitations. The question-answering and sentiment analysis datasets used for training the specialist models are translated from English to Urdu. This translation process can introduce inaccuracies that may affect the models' performance. Additionally, the lack of native, high-quality Urdu datasets poses a significant challenge. Translated datasets often fail to capture the linguistic and cultural nuances inherent to Urdu, which can impact both training and evaluation outcomes. The evaluation is conducted on a subset of 1000 data points for each task. While this size is manageable and allows for a cost-efficient evaluation, it may not be fully representative of the model's overall performance. Although we use multiple annotators and calculate inter-rater reliability using Krippendorff's alpha, there is still a degree of subjectivity that may influence the results.

9 Ethical Impact

In this paper we present a comprehensive evaluation of LLMs with the aim to enhance the accessibility of NLP applications for Urdu speakers. This has significant ethical implications, as it addresses the digital divide and promotes linguistic diversity in technology. Our findings indicate that specialist models perform better than the generalist models in most Urdu NLP tasks. Consequently, our work may inspire researchers to develop more resources for the Urdu language, including models and datasets.

The potential risks associated with the usage of LLMs include the amplification of existing biases present in their training data, which may lead to unfair and discriminatory outcomes (Ye et al., 2023). A comprehensive fairness evaluation of these models must be conducted before they are deployed for public use.

Acknowledgments

We are grateful for the time and effort put in by our research intern, Muhammad Suhaib Rashid who annotated our data and Mustafa Abbas who worked on creating Wiki-UQA dataset. We are also grateful to OpenAI for supporting our work through their Research Access Program.

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. *Preprint*, arXiv:2305.14982.

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. *Preprint*, arXiv:2303.12528.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *Preprint*, arXiv:2311.07463.

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, and Muhammad Tariq Sadiq. 2020. Automatic detection of offensive language for urdu and roman urdu. *IEEE Access*, 8:91213–91226.

Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020. Data augmentation using machine translation for fake news detection in the Urdu language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2537–2542, Marseille, France. European Language Resources Association.

Maaz Amjad, Alisa Zhila, Grigori Sidorov, Andrey Labunets, Sabur Butta, Hamza Imam Amjad, Oxana Vitman, and Alexander Gelbukh. 2022. Overview of abusive and threatening language detection in urdu at fire 2021. *Preprint*, arXiv:2207.06710.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024a. UQA: Corpus for Urdu question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17237–17244, Torino, Italia. ELRA and ICCL.

Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2024b. The fellowship of the llms: Multi-agent workflows for synthetic preference optimization dataset generation. *Preprint*, arXiv:2408.08688.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Mahwiz Khalil. 2023. Urduassistant dataset. `https://huggingface.co/datasets/mwz/UrduAssistant`.

Shumaila Khan and Fahad Najeeb. 2023. Urdu sarcastic tweets dataset.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023a. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *Preprint*, arXiv:2305.14976.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023b. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, and Dipti Misra Sharma. 2024. Towards large language model driven reference-less translation evaluation for english and indian languages. *Preprint*, arXiv:2404.02512.

Mubashir Munaf, Hammad Afzal, Naima Iltaf, and Khawir Mahmood. 2023. Low resource summarization using pre-trained language models. *Preprint*, arXiv:2310.02790.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Işin Demirşahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 2413–2423.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *Preprint*, arXiv:2404.16816.

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, and Sarmad Hussain. 2024. Benchmarking pre-trained large language models' potential across urdu nlp tasks. *Preprint*, arXiv:2405.15453.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul,

Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. 2023. Assessing hidden risks of llms: An empirical study on robustness, consistency, and credibility. *Preprint*, arXiv:2305.10235.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Llmeval: A preliminary study on how to evaluate large language models. *Preprint*, arXiv:2312.07398.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

A Implementation Details

A.1 Models

`XLM-Roberta-large` is availabe on Hugging Face[9] under MIT license. `mT5-large` is available on Hugging Face[10] under Apache-2.0 license. `Llama-3-8b` is available on Hugging Face[11] under llama3 license. `GPT-4` is available under proprietary licence. All models used in this paper comply with their respective license.

A.2 Datasets

Urdu IMDB sentiment analysis dataset OBDL license. Urdu NER dataset is availabe under MIT license. The abuse detection dataset by Akhter et al. (2020) and by Amjad et al. (2022), Sarcastic Tweets Dataset (Khan and Najeeb, 2023), fake news dataset by Amjad et al. (2020), Topic Classification dataset, and Universal Dependencies (Nivre et al., 2020) are available under CC BY 4.0.

UQuAD dataset is under CC0-1.0 while UQA (Arif et al., 2024a) is under CC BY 4.0. XSUMUrdu (Munaf et al., 2023) summarization dataset is also under CC BY 4.0 license. Paraphrasing dataset is under MIT license. Dakshina dataset (Roark et al., 2020) for transliteration is under CC BY-SA 4.0 and OPUS-100 (Zhang et al., 2020) (Tiedemann, 2012) for translation is under GPL-3.0 license. UrduAssistant (Khalil, 2023) dataset has a MIT license.

All datasets used in this paper comply with their respective license.

B Model Size and Budget

We fine-tuned `XLM-Roberta-large` for classification tasks which has 550 million parameters. We fine-tuned `mT5-large` for generation tasks which has 1.2 billion parameters. `Llama-3-8b` has 8 billion parameters and is fine-tuned for both generation and classification tasks.

Nvidia A100 80GB, Nvidia A100 40GB and Nvidia RTX 6000Ada 48GB were used for fine-tuning of the models. Infernece was done on Nvidia

[9] https://huggingface.co/FacebookAI/xlm-roberta-large

[10] https://huggingface.co/google/mt5-large

[11] https://huggingface.co/meta-llama/Meta-Llama-3-8B

RTX 6000Ada 48GB and Nvidia RTX 4090 24GB. Total GPU time was approximately 200 hours.

C Human Annotators

There are two human annotators in this study: one is the author of this paper (Computer Science graduate), and the other is a research intern (Computer Science senior). Both annotators are native speakers of Urdu from Pakistan. The research intern was informed about how the data would be used for the evaluation of LLMs for Urdu.

D Dataset Size

This section provides details about the datasets used for the classification and generation tasks evaluated in this study. The test dataset sizes for each task are summarized in the tables below.

D.1 Classification

Table 6 shows the test dataset sizes for classification tasks including Sentiment Analysis, Abuse Detection, Sarcasm Detection, Fake News Detection, Topic Classification, Part-of-Speech (PoS) Tagging, and Named Entity Recognition (NER) Tagging.

Task	**Test Size**
Sentiment Analysis	1000
Abuse Detection	567
Sarcasm Detection	1000
Fake News Detection	130
Topic Classification	1000
PoS Tagging	1000
NER Tagging	1000

Table 6: Test dataset size for each classification task

D.2 Generation

Table 7 shows the test dataset sizes for generation tasks including Summarization, Paraphrasing, Transliteration, English to Urdu Translation, and Urdu to English Translation.

E Reproducibility and Hyperparameter

The table 8 presents the Lora configuration that we use for fine-tuning `Llama`.

Task	**Test Size**
Question Answering	1000
Summarization	568
Paraphrasing	1000
Transliteration	1000
Translation (en-ur)	1000
Translation (ur-en)	1000
AI Assistant	50

Table 7: Test dataset size for each generation task

Rank	16
Alpha	16
Target Modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `down_proj`, `up_proj`, `gate_proj`

Table 8: Lora config for `Llama` fine-tuning

The table 9 presents the training parameters, including the number of epochs and batch sizes, we use for fine-tuning `XLM-R`, `mT5`, and `Llama`.

Table 9: Training Parameters for Different Models. E represents number of epochs and B represents the batch size.

	`XLM-R`		`mT5`		`Llama`	
Task	E	B	E	B	E	B
Sentiment Analysis	1	16	-	-	4	1
Abuse Detection	4	16	-	-	3	8
Sarcasm Detection	4	16	-	-	3	8
Fake News Detection	6	8	-	-	3	1
Topic Classification	4	32	-	-	3	8
POS Tagging	7	16	-	-	1	4
NER Tagging	10	16	-	-	3	4
Question Answering	-	-	4	8	1	1
Summarization	-	-	1	4	2	1
Paraphrasing	-	-	3	3	4	4
Transliteration	-	-	5	8	2	4
Translation-en-ur	-	-	2	8	1	8
Translation	-	-	3	8	1	8
AI Assistant	-	-	2	8	2	8

F Human and LLM Evaluation

Figure 12 presents the rank distribution for all models across all generation tasks, as evaluated by two human evaluators: `GPT-4-Turbo`, `Llama-3-8b`, and `Claude 3.5 Sonnet`.

G Prompts and Evaluation Criteria

Table 10 contains the CoT prompts used for the classification tasks. The prompts for classification without CoT are the same, except that the reasoning is not generated, and only the label is required in the output. Table 11 contains the prompts used for the generation tasks. Table 12 presents the criteria given to human evaluators, `GPT` and `Llama` for assessing the quality of outputs for summarization, paraphrasing, transliteration, English to Urdu translation, and Urdu to English translation. For each evaluation, the LLMs are asked to reason the answer to improve the scoring of the outputs.

Figure 12: Rank counts of models for all generation tasks.

Classification Task	CoT Prompt

Sentiment Analysis	`You are an Urdu sentiment classifier. The input text should be labeled according to its sentiment. The label list is: ['positive', 'negative']`
	`Use Chain-of-Thought (cot) to reason your answer.`
	`ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY:`
	`{"cot": ..., "label": ...}`

Abuse Detection	`You are an Urdu abuse detector. The input text should be labeled according to whether it is abusive or not. The label list is: ['abusive', 'not abusive']`
	`Use Chain-of-Thought (cot) to reason your answer.`
	`ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY:`
	`{"cot": ..., "label": ...}`

Classification Task	CoT Prompt
Sarcasm Detection	You are an Urdu sarcasm detector. The input text should be labeled according to whether it is sarcastic or not. The label list is: ['sarcastic', 'not sarcastic'] Use Chain-of-Thought (cot) to reason your answer. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"cot": ..., "label": ...}
Fake News Detection	You are an Urdu fake news detector. The input text should be labeled according to whether it is fake news or not. The label list is: ['fake news', 'not fake news'] Use Chain-of-Thought (cot) to reason your answer. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"cot": ..., "label": ...}
Topic Classification	You are an Urdu topic classifier. The input text should be assign a label from the label list: ['business', 'entertainment', 'health', 'politics', 'science', 'sports', 'world', 'other'] Use Chain-of-Thought (cot) to reason your answer. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"cot": ..., "label": ...}
PoS Tagging	You are an Urdu part-of-speech tagger. The word wrapped in <hl> tag should be assigned a PoS tag from the label list: ['noun', 'punctuation mark', 'adposition', 'number', 'symbol', 'subordinating conjunction', 'adjective', 'particle', 'determiner', 'coordinating conjunction', 'proper noun', 'pronoun', 'other', 'adverb', 'interjection', 'verb', 'auxiliary verb'] Use Chain-of-Thought (cot) to reason your answer. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"cot": ..., "label": ...}
NER Tagging	You are an Urdu named entity recognizer. The word wrapped in <hl> tag should be assigned a NER tag from the label list: ['time', 'person', 'organization', 'number', 'location', 'designation', 'date', 'other'] Use Chain-of-Thought (cot) to reason your answer. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"cot": ..., "label": ...}

Table 10: The table presents the CoT prompts used for classification tasks.

Generation Task	Prompt
Summarization	You are a text summarizer. Your task is to summarize the given Pakistani Urdu text in 1 to 2 sentences. The summary should be in Urdu. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"summary": ...}

Generation Task	Prompt
Paraphrasing	You are a text paraphraser. Your task is to paraphrase the given Pakistani Urdu text. The paraphrased text should be in Urdu. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"paraphrase": ...}
Transliteration	You are a machine transliterator. Your task is to transliterate the given Pakistani Urdu text to Roman Urdu. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"transliteration": ...}
Translation (en-ur)	You are a machine translator. Your task is to translate the given English text to Pakistani Urdu. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"translation": ...}
Translation (ur-en)	You are a machine translator. Your task is to translate the given Pakistani Urdu text to English. ALWAYS RETURN JSON OBJECT IN FOLLOWING FORMAT ONLY: {"translation": ...}

Table 11: The table presents the prompts used for generation tasks.

Task	Prompt/Criteria
Summarization	You are a Pakistani Urdu language expert tasked with evaluating the quality of the summary produced by the summarization model. Score the given output with respect to the given input on a continuous score from 0 to 100 based on the following criteria: 1. Includes main points and key information from the original text 2. No grammatical errors 3. Conveys information in a brief manner 4. Gives correct information based on the original text Think step by step and use reasoning. ALWAYS RETURN JSON OBJECT IN THE FOLLOWING FORMAT ONLY: {"reasoning": ..., "score": ...}
Paraphrasing	You are a Pakistani Urdu language expert tasked with evaluating the quality of paraphrased text produced by the paraphrasing model. Score the given output with respect to the given input on a continuous score from 0 to 100 based on the following criteria: 1. Retains the original meaning and key ideas 2. No grammatical errors 3. Use of different words and phrases than the original text Think step by step and use reasoning. ALWAYS RETURN JSON OBJECT IN THE FOLLOWING FORMAT ONLY: {"reasoning": ..., "score": ...}

Task	Prompt/Criteria
Transliteration	You are a Pakistani Urdu language expert tasked with evaluating the quality of transliterated text produced by the transliteration model. Score the given output with respect to the given input on a continuous score from 0 to 100 based on the following criteria: 1. Correctness of words (keeping in mind that different words can have different spelling) 2. Proper capitalization of words Think step by step and use reasoning. ALWAYS RETURN JSON OBJECT IN THE FOLLOWING FORMAT ONLY: {"reasoning": ..., "score": ...}
Translation (en-ur)	You are a Pakistani Urdu language expert tasked with evaluating the quality of translated text produced by the translation model. Score the given output with respect to the given input on a continuous score from 0 to 100 based on the following criteria: 1. Conveys the meaning of the original text without omissions or additions 2. No grammatical errors 3. Retains the style and tone of the original text Think step by step and use reasoning. ALWAYS RETURN JSON OBJECT IN THE FOLLOWING FORMAT ONLY: {"reasoning": ..., "score": ...}
Translation (ur-en)	You are a Pakistani Urdu language expert tasked with evaluating the quality of translated text produced by the translation model. Score the given output with respect to the given input on a continuous score from 0 to 100 based on the following criteria: 1. Conveys the meaning of the original text without omissions or additions 2. No grammatical errors 3. Retains the style and tone of the original text Think step by step and use reasoning. ALWAYS RETURN JSON OBJECT IN THE FOLLOWING FORMAT ONLY: {"reasoning": ..., "score": ...}
AI Assistant	Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 0 to 100. ALWAYS RETURN JSON OBJECT IN THE FOLLOWING FORMAT ONLY: {"reasoning": ..., "score": ...}

Table 12: The table presents the criteria used to evaluate the outputs of the generation task in the form of an LLM prompt.