Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Prompt Engineering for Low-Resource NLP: A Survey with Urdu-English Applications

Areeba Saeed
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, 59030, Pakistan

Rubaina Nazeer
Department of Software Engineering, National University of Modern Languages, Multan

Syeda Qanitah Naqvi
Department of Software Engineering, National University of Modern Languages, Multan

Muhammad Sajid Maqbool
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, 59030, Pakistan Email: sajid.maqbool@nfciet.edu.pk

Naeem Aslam
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, 59030, Pakistan

Hira Saleem
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, 59030, Pakistan

Muqadas Nadeem
Department of Computer Science, Emerson University, Multan

ABSTRACT

With limited computational resources and labeled data, large language models (LLMs) can be effectively adapted to new tasks through prompt engineering especially in low-resource language environments. Focusing on natural language processing (NLP) in Urdu which is a language with complicated morphology, script variations, code-mixing and limited high-quality datasets, this study investigates prompt engineering strategies. Prompting techniques such as zero-shot and few-shot are methodologically structured and examined especially for Urdu-specific issues including tokenization, linguistic grounding and cross-lingual evaluation. In addition to the survey, a prompt-only experimental investigation is carried out on Urdu abstractive summarization and sentiment classification. Multilingual (mGPT, mBART), generalist instruction-tuned (GPT-4, Gemini, Gemma) models and an Urdu-specific (urT5) model are evaluated using a single prompt-based architecture.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

For cross-lingual summarization, a two-stage pipeline is used involving urT5 generates summaries in Urdu that are then translated into English using mBART for evaluation. Clear task-dependent performance differences between model variants are revealed by quantitative results obtained using common automatic measures. The findings, with Urdu-specific models demonstrating stronger alignment in generation-centric tasks and generalist models remaining more effective for instruction-driven classification, show that prompt engineering is a valuable but non-uniform adaptation method for Urdu NLP, along with providing practical guidance for future low-resource language research.

Keywords: Prompt Engineering, Low-Resource NLP, Urdu-English Translation, Large Language Models, Few-Shot Learning, Multilingual AI

Introduction

In the field of natural language processing (NLP), the emergence of large language models (LLMs) like GPT-3 or ChatGPT help in significant advancement by allowing the generation of intelligible and context-aware text [1]. Prompting plays a significant role in the rapid development of these models by providing structured input to a model using which it generates output which makes this technique a major field known as prompt engineering [2, 6]. By carefully designing input prompts the users are able to refine responses of the model across different tasks including translation and summarization [7, 9].

A growing trend in NLP towards more adaptive and customizable language solutions is represented by the transition from parameter-intensive fine-tuning to prompt interaction. However, underrepresented languages such as Urdu continue to face considerable challenges while all the advancements in NLP have primarily benefited high-resource languages [4, 5].

The unavailability of high-quality, native datasets is a primary challenge because most of the available resources are either translated from English or not sufficient in size and variety [4, 10]. The improvement of models for these languages is also halted by the limited availability of benchmark metrics and evaluation tools [3, 11]. Moreover, major modern technologies typically developed for English have seen restricted application in Urdu-related NLP tasks such as abstractive summarization and cross-lingual transfer learning [8, 9]. The progress is further affected by issues like underrepresentation in multilingual models, systemic biases and shortage of robust pre-trained models [5]. The urgent need for research for the development of standardized datasets, employment of advanced techniques and assurance of comprehensive model evaluation for Urdu and other low-resource languages is emphasized by these limitations [3, 4].

An extensive overview of prompt engineering techniques especially for the Urdu-English language pair, focusing on their application in low-resource languages is provided in this article [6, 7]. The article points out successful techniques and organizes them according to task type and model architecture by reviewing recent studies related to multilingual prompting, model adaptation and evaluation strategies [2]. Specific limitations like prompt instability, dataset scarcity and language-specific biases are also mentioned in this study [2, 5].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Structural organization of the remaining article is as follows: Section 2 provides a foundational overview of prompt engineering and common prompt engineering techniques. A detailed review of recent literature on prompt-based methods for multilingual and low-resource NLP is presented in Section 3. Major prompting techniques including few-shot prompting and chain-of-thought prompting along with their applicability to low-resource scenarios are outlined in Section 4. Section 5 discusses key challenges in low-resource languages especially in Urdu. Existing applications and opportunities in Urdu-English NLP tasks are explored in Section 6. Section 7 summarizes the findings and proposes directions for future research.

Background & Terminology

Prompt Engineering – Definition and Evolution

Large Language Models (LLMs) are guided by structured text instructions referred to as input prompts designed by the process known as prompt engineering to generate context specific and precise outputs. Rather than retraining or modifying parameters of the model, prompt engineering leverages the pre-trained capabilities of LLMs by strategically shaping the input so that it influences the model's behavior across a wide range of tasks [2], [6]. As compared to traditional fine-tuning, this technique has emerged as a flexible and resource-efficient alternative. Conventional NLP models are trained using a process that demands considerable computational resources, time and labeled datasets as these models are first trained on large text datasets and then refined on task-specific datasets [1]. On the other hand, prompt engineering enables the models to adapt to new tasks keeping the model parameters unchanged and controls output generation through customized prompts [7].

Specialized models for specific tasks are produced by fine-tuning whereas prompt engineering enables the same general-purpose model to perform multiple tasks just by changing the prompt. This difference is illustrated in

Figure 1. In low-resource settings where labeled data and processing infrastructure are limited, this technique is lightweight and accessible because no gradient updates or back-propagation is needed. Recent work has introduced advanced techniques like role prompting, chain-of-thought and self-consistency prompting which in earlier approaches was just a simple template. Interpretability, reasoning ability and output stability have been improved by these developments.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

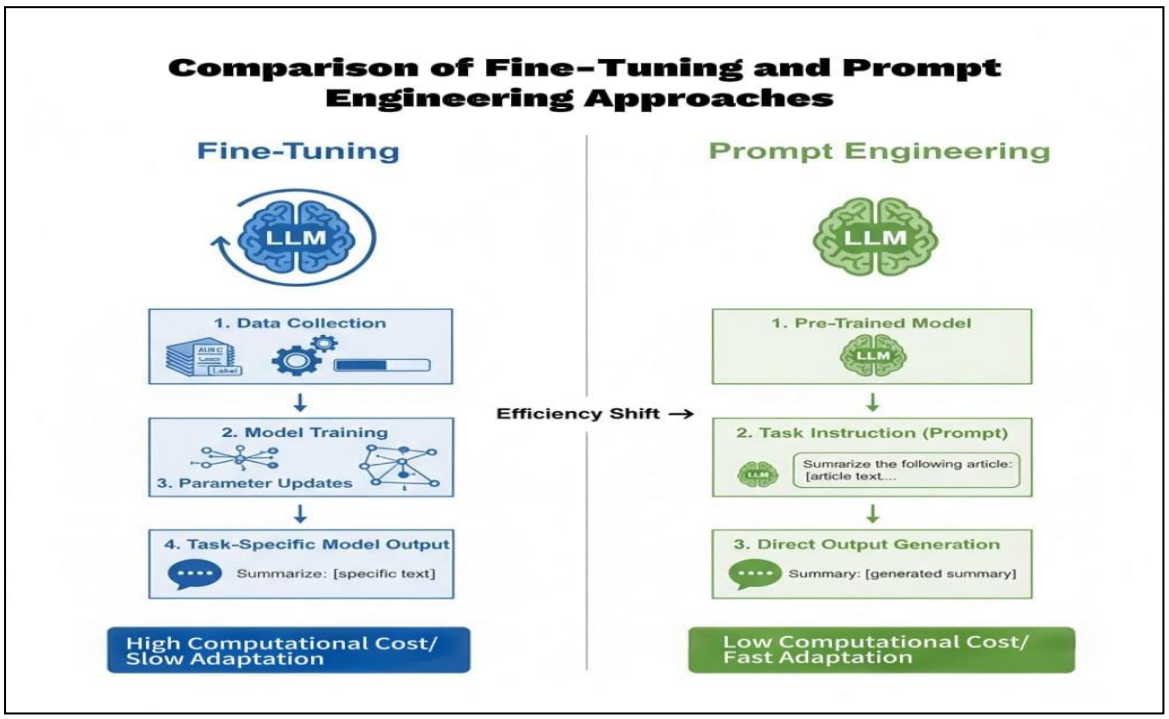


Figure 1. Comparison of fine-tuning and prompt engineering approaches.

Prompting Strategies: Types and Use Cases

Effective prompting strategies are necessary for prompt engineering which will direct LLMs towards accurate and context-aware results. Complexity, input structure and areas of applicability differentiate these strategies [2,7].

Zero-Shot Prompting

Without providing prior examples, **zero-shot prompting** directs the model to complete a task. To interpret the request, the model merely depends on its pre-trained linguistic knowledge [1]. For low-resource languages where labeled data is limited, this approach is particularly advantageous. However, due to insufficient exposure to high-quality Urdu data during the pre-training phase, the effectiveness of the model for Urdu is usually constrained [2, 6].

Few-Shot Prompting

The framework of zero-shot is expanded by **few-shot prompting** by attaching a few examples within the input prompt. The task pattern is illustrated by these examples using which the model can generalize. For low-resource languages, this approach enhances accuracy in classification, summarization and translation tasks [2,9].

Chain-of-Thought (CoT) Prompting

The model is allowed to generate intermediate reasoning steps before finalizing the output in **chain-of-thought prompting**. CoT prompting helps in decomposing complex linguistic patterns into manageable steps thus mainly developed for arithmetic and logical reasoning [2]. This approach enhances structured generation

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

and syntactic understanding for Urdu which is a morphologically rich language.

Role-Based Prompting
Explicit identity or role is assigned to the model in this strategy. For example, "You are an Urdu language instructor. Explain past tense forms." [2, 6].

Self-Consistency Prompting
For a single prompt, the model generates multiple responses and selects the most consistent one. Especially in reasoning tasks, this technique improves reliability and factual accuracy, and also reduces output instability in low-resource settings [2, 11].

Instruction-Based Prompting
The model is guided towards a specific objective using explicit instructions. Ambiguity is reduced by giving clear directions and in multilingual applications, robustness of instruction-tuned models such as InstructGPT and FLAN-T5 is improved [7]. Even with minimal fine-tuning these instructions aligned models perform well across diverse tasks which demonstrate that, for Urdu NLP contexts where high-quality data is limited, these models are ideal [2].

Contextual and Knowledge-Infused Prompting
To enhance comprehension, these advanced techniques incorporate external resources such as bilingual dictionaries or high frequency word lists. Accuracy of translation for morphologically complex or culturally sensitive languages is improved using these techniques and has shown benefits in Urdu NLP tasks [3, 5].

Emerging Strategy: Prompt Tuning and Contextual Calibration
Soft prompt embeddings which are trainable vectors that guide the generation of the model are introduced in **prompt tuning** while actual parameters of the model are kept frozen. To improve consistency, **contextual calibration** adjusts statistical distributions of responses [2]. A frequent issue in low-resource environments in which small variation in prompt majorly alters output known as prompt sensitivity is reduced by both of these methods [11].

Common Prompt-Based Models in NLP
The success of prompt engineering depends heavily on the adaptability of underlying LLMs. Recent years have seen the development of numerous models—ranging from general-purpose to multilingual and fine-tuned variants—that form the basis for prompting research [2].

GPT Series: GPT-3, GPT-3.5, GPT-4, GPT-4-Turbo
OpenAI's GPT-3 and GPT-4 series set the benchmark for zero-shot and few-shot learning across summarization, reasoning, and translation tasks [1].

mT5 and urT5
The mT5 model supports more than 100 languages, including Urdu, enabling prompt-based multilingual learning. urT5, an Urdu-specific derivative, achieves strong

Online ISSN 3007-3197
Print ISSN 3007-3189

performance on summarization even with limited data [11].

XLM-R and Serengeti
These multilingual transformers support cross-lingual understanding and have been applied to low-resource African and South-Asian languages [3],[5].

LLaMA and LLaMA-FT
Developed by Meta, LLaMA models are efficient open-source alternatives for research in low-resource contexts [4].

Mistral 7B and Gemma
Both lightweight open-source models have shown promising results in prompt-based translation experiments [3].

NLLB (No Language Left Behind)
Meta's No Language Left Behind project offers state-of-the-art translation for over 200 languages and integrates smoothly with prompt-driven pipelines [5].

Flan-T5 and PaLM
Instruction-tuned models like Flan-T5 and PaLM excel in few-shot reasoning tasks and provide generalizable performance for low-resource NLP [2].

Multilingual GPT Models (mGPT, XGLM, bLOOM)
mGPT, XGLM, and bLOOM extend GPT-style architecture to multilingual settings and are particularly effective for Urdu translation and summarization [9].

BART and T5 (Classic Architectures)
Classic encoder-decoder models such as BART and T5 remain benchmarks for text-to-text generation tasks [10].

Prompt Design Principles & Challenges
The effectiveness of prompt engineering depends not only on model selection but also on how clearly and coherently prompts are constructed. The quality and stability of generated outputs are directly affected by the wording, organization and contextual richness of the prompts. Key design principles and challenges regarding low-resource environments like Urdu are outlined in this section.

Significance of Prompt Formulation
The interpretation and response of LLM to a task are determined by the phrasing of a prompt. Accurate results can be generated by using clear and precise instructions whereas ambiguous phrasing results in off-topic outputs [1, 2]. As linguistic and cultural nuances may not be well represented in training data, clarity is necessary for multilingual contexts [3].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Significance of Prompt Formatting

Structuring the prompts using bullet points, templates, or numbered lists guide the model more effectively. Few-shot examples arranged in consistent formats improve comprehension [7].

Significance of Context Inclusion

Providing contextual background or task-specific information helps in maintaining continuity and enhancing reliability in responses [2, 3].

Related Work

With a specific focus on Urdu, this section reviews previous research on prompt engineering for multilingual and low-resource natural language processing (NLP). The literature is organized according to the prompting taxonomy introduced in Section 2, describing how different strategies have been utilized for Urdu related NLP tasks like summarization, classification and translation.

Prompt Strategies Applied to Low-Resource and Urdu NLP Tasks

In low-resource settings, zero-shot and few-shot prompting (Section 2.2.1-2.2.2) have been widely explored as a lightweight alternative to fine-tuning. According to studies, due to limited exposure of Urdu language during pre-training, zero-shot prompting often generates inconsistent results even though it enables tasks execution without labeled data [1, 6]. On the other hand, few-shot prompting, by providing minimal task examples within the prompt, has been shown to improve output reliability in tasks like Urdu sentiment classification and translation [2, 9].

For reasoning-intensive tasks, Chain-of-Thought (CoT) (Section 2.2.3) has primarily been evaluated. Recent work suggests that CoT prompting which was originally introduced for mathematical and logical reasoning, particularly when combined with instruction-tuned models, can support structured generation and syntactic clarity in morphologically rich languages such as Urdu [2].

In multilingual prompting research, instruction-based prompting and instruction-tuned models (Section 2.2.6) play a central role. As compared to non-instruction-tuned models, models like FLAN-T5 and InstructGPT have shown improved coordination with input task instructions and higher cross-lingual robustness which make them suitable baselines for low-resource languages [7].

By learning soft prompt representations or normalizing response distributions, approaches that are currently emerging such as prompt tuning and contextual calibration (Section 2.2.8), aim to decrease prompt sensitivity and stabilize outputs. Recent surveys highlight their potential applicability in multilingual and low-resource settings even though these methods have primarily been evaluated on high-resource languages [2, 11].

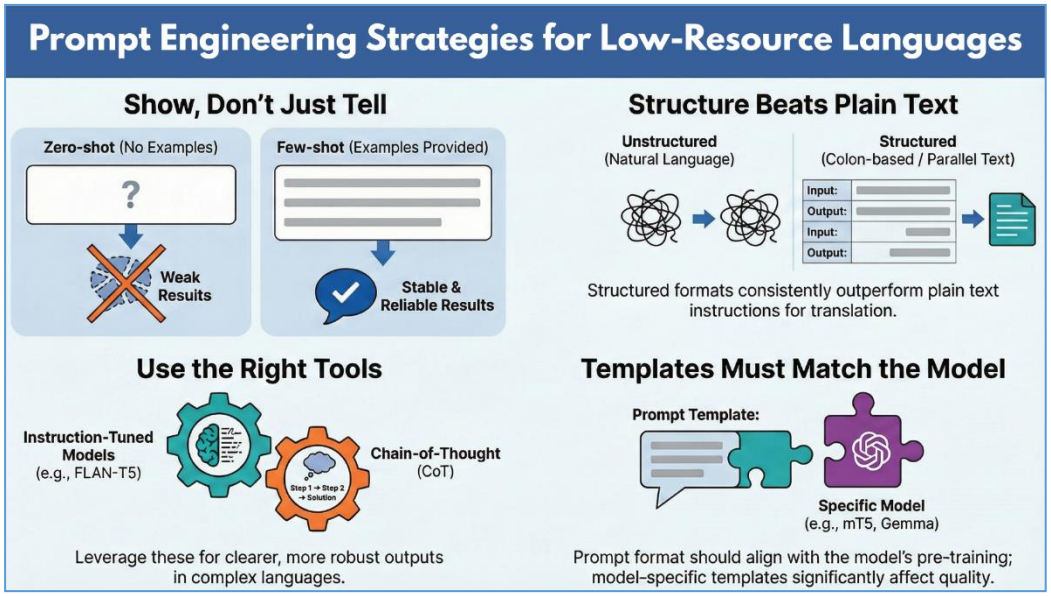
Prompt-Based Translation and Summarization in Low-Resource Languages

For low-resource languages including Urdu and other underrepresented languages from South Asia and Africa, prompt-based translation and summarization have got increasing attention. According to existing studies, prompt structure and formatting considerably influence performance in cross-lingual tasks [3]. Structured in-context

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



templates (e.g., colon-based or parallel text formats) especially when combined with lexical cues such as bilingual dictionaries, consistently perform better than unstructured natural language instructions [5].

More stable outputs have been reportedly generated by multi-example (few-shot) prompts as compared to single-example settings in translation tasks. Whereas, from generalist LLMs, consistency in summarization outputs can be improved by carefully constrained decoding parameters (e.g., fixed temperature) [2, 11]. Strong language integrity is obtained by Urdu specialized models when operating monolingually whereas fluent but sometimes shallow summaries are generated in zero-shot settings by general-purpose instruction-tuned models for Urdu summarization [10].

Also, outputs have been affected by model-specific prompt templates. The importance of aligning prompt format with model pre-training objectives and tokenizer behavior is emphasized by the studies which involve multilingual architectures like Serengeti, Gemma and mT5 [3]. A conceptual overview of prompting techniques commonly used in existing literature for translation and summarization in low-resource scenarios is provided in **Figure 2**.

Figure 2. Conceptual overview of prompting techniques for translation and summarization tasks

Reported Performance Trends of LLMs on Urdu Tasks

Depending on **model type**, **task complexity** and **degree of adaptation**, the performance of large language models (LLMs) varies for Urdu tasks according to existing literature. In generative tasks like summarization and translation, the performance of generalist models, especially instruction-tuned GPT-style architectures, often substantially benefitted from few-shot prompting [2, 8]. On the other hand, for analytical tasks like sentiment classification and information retrieval, specialist and multilingual models which are fine-tuned on Urdu or related languages tend to show strong performance [5].

However, multiple limitations are common across studies. Evaluations are commonly

Online ISSN 3007-3197
Print ISSN 3007-3189

carried out on relatively small or translated datasets and prompt sensitivity remains a recurring challenge due to which minor changes in phrasing cause notable variations in output quality [2]. Cultural authenticity is limited and lexical artifacts are introduced because the majority of widely used Urdu benchmarks like UquAD, UQA and IMDB Urdu Sentiments are generated from translated sources [4].

Collectively, complementary strengths between generalist and specialist models are highlighted by existing research as: flexibility and fluency are offered by generalist LLMs under prompt-based interactions, whereas stronger linguistic grounding is provided by specialist models when sufficient adaptation data is available. For Urdu and low-resource languages, these characteristics encourage further research into hybrid strategies that integrate prompt engineering with lightweight fine-tuning and culturally grounded evaluation.

Low-Resource NLP Challenges

Unique challenges are demonstrated during the development of natural language processing (NLP) systems for low-resource languages that are not much prominent in high-resource settings. The scarcity of training data, inadequate linguistic tools, inconsistent evaluation metrics and instability in the performance of large language models (LLMs) are some limitations in low-resource languages such as Urdu. These limitations directly affect the effective application of prompt engineering techniques in multilingual and culturally diverse environments [2, 4].

Training Data Limitations

The extensive and diverse datasets are mostly unavailable for low-resource languages whereas for high-resource languages massive datasets are available. Datasets which are specific to any domain, machine translated and some small academic compilation, are currently available for Urdu [3]. Weak model generalization and poor semantic comprehension are because of this limited coverage. Urdu morphology and cultural nuance are misinterpreted by the models which are mainly pre-trained on English text as a result downstream performance is decreased. Synthetic data creation and cross-lingual dataset expansion has been expected to resolve these limitations [5].

Tokenization and Linguistic Representation

For morphologically rich languages like Urdu which depend on complex word forms and inflectional patterns, effective tokenization is necessary. Byte Pair Encoding (BPE), SentencePiece and many more widely used tokenizers are utilized for high-resource languages and struggle in accurately segmenting Urdu text [2]. Distorted meanings and inconsistent embeddings are caused by this. For improving model fluency and representational quality, development of Urdu-specific tokenizers and integration of morphological analyzers remain essential.

Prompt Sensitivity and Stability

Large language models are often highly sensitive to slight variations in wording, punctuation or formatting of prompts. This sensitivity is increased in low-resource contexts because of limited language exposure during pre-training [2]. Two prompts that appear to be semantically identical may generate dramatically different outputs

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

for Urdu. This volatility may partially reduce by using calibration techniques, prompt standardization and self-consistency but cannot be completely eliminated [11].

4.4 Evaluation and Benchmarking Challenges

For measuring progress in low-resource NLP, reliable evaluation is essential. However, as compared to English datasets such as GLUE or XSUM, Urdu lacks standardized large scale benchmarks. Instead of natively authorized currently available Urdu datasets such as UQuAD, UQA, IMDB Urdu Sentiments are frequently translated which cause lexical artifacts and cultural inconsistencies. Therefore, capability of real-world models may not be accurately reflected by available performance metrics. Cross-model comparability would be enhanced by the development of culturally grounded benchmarks and standardized scoring systems.

Cross-Lingual Transfer Limitations

Knowledge sharing from high-resource languages to low-resource ones is enabled by cross-lingual transfer. However, due to weak embedding alignments and cultural mismatch between source and target languages, the effectiveness of this transfer is restricted [5]. Additionally, Perso-Arabic script of Urdu and rich morphology create hurdles in transfer from Latin-based dataset. These inconsistencies can be eliminated by including adapter-based fine-tuning, multilingual embeddings and domain-specific pre-training [3].

Cultural and Contextual Bias

Cultural and contextual misrepresentation frequently occurs even when LLMs perform well synthetically. Linguistically correct but culturally inappropriate results may be generated by the prompts which include idiomatic or religious expressions [4]. Training dataset dominated by Western text sources is the root cause of this bias. Such biases can be reduced by building balanced, culturally representative datasets and applying fairness-aware prompt tuning.

Experimental Setup

To evaluate the effectiveness of **prompt-only adaptation** for Urdu-English summarization and sentiment classification, the experimental setup utilized in this article is described in this section. Consistent with recent prompt engineering studies, this framework enables a fair comparison of generalist, multilingual and Urdu-specific language models in low-resource settings [2, 7].

Models and Prompt-Only Evaluation Framework

To cover a broad spectrum of architectural design and linguistic specialization, six representative large language models (LLMs) were selected:

Generalist proprietary models: Instruction-tuned models, GPT-4, Gemini 2.5 pro and Gemma are utilized for cross-lingual task execution using zero-shot and few-shot prompts.

Multilingual models: mGPT and mBART which are trained on multilingual datasets with varying degrees of Urdu exposure.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Urdu-specialized model: urT5 which is known for strong morphological and semantic representation of Urdu language.

From the monolingual nature of urT5 a key methodological consideration arises. A **two-stage pipeline** is used to evaluate urT5 for cross-lingual summarization, since it is not capable of directly generating English summaries: first urT5 generates Urdu abstractive summaries which are subsequently translated into English using mBART in second stage. A realistic deployment scenario for monolingual low-resource models in cross-lingual application is reflected by this composition pipeline (urT5 → mBART) which allows direct comparison with generalist English-output LLMs [5].

Datasets

To evaluate summarization and sentiment classification, two publicly available Urdu datasets were utilized.

UATS-100K Abstractive Summarization Dataset.

The UATS-100K, which is a subset of the larger UATS-23 benchmark corpus was used for summarization experiments [16]. The dataset is suitable for abstractive summarization as each summary corresponds to a human-written news headline.

The 100,000-sample subset is commonly used in existing studies to deal with computational constraints. In this article, **100 articles** were randomly sampled from this subset to use for prompt-based evaluation across all models. This subset allows controlled and consistent comparison of prompt-only behavior between different model families.

IMDb Urdu Reviews Sentiment Classification Dataset

The IMDb Urdu Reviews dataset, which consists of 50,000 machine-translated Urdu movie reviews derived from the original English IMDb corpus, was used to conduct sentiment classification experiments [8]. Each review is labeled with binary sentiment class (positive or negative). **100 reviews** were selected for evaluation for consistency with the summarization setup.

Data Preprocessing

Standardized preprocessing pipelines were applied to both datasets to ensure textual consistency and minimize noise.

Summarization Dataset Preprocessing.

The following steps were applied to the UATS-100K subset:

Duplicate removal: Identical article-summary pairs were removed

Short summary filtering: Articles with summaries shorter than five characters were discarded

Unicode normalization: Using NFC normalization, all Urdu text was normalized to standardize equivalent character forms

Noise removal: Using regular expressions, HTML tags, URLs and control characters were removed

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Script filtering: To retain Urdu-dominant samples, a script-ratio filter was applied
Tokenization: Manual tokenization was not performed, instead model-specific SentencePiece tokenizers were used to ensure compatibility with transformer-based models
The natural morphological and orthographic structure of Urdu is preserved by using this preprocessing while maintaining consistency across models.
5.3.2 Sentiment Classification Dataset Preprocessing.
Following preprocessing pipelines were used for IMDb Urdu Reviews subset:
Unicode normalization using NFC
Noise removal including HTML tags, URLs and excessive whitespace
Code mixing handling: English words were kept due to the frequent code-mixing in Urdu digital text. However, samples dominated by non-Urdu text were eliminated
Tokenizer consistency: To ensure consistent encoding across tasks, the same SentencePiece tokenizer was used.

Prompt Design and Task Specification
In prompt-only adaptation settings, the design of prompt plays a central role in controlling model behavior. Prompts were carefully customized according to architectural and linguistic characteristics of each model by following best practices outlined in prior prompt-engineering research [2].
Two prompting strategies were used throughout:
Zero-shot prompting, to assess baseline task understanding.
Few-shot prompting, to provide structured guidance using minimal examples.

urT5 Prompt Design (Urdu Abstractive Summarization).
Prompts were designed completely in idiomatic Urdu, avoiding English meta-instructions as urT5 is specifically trained on Urdu text. Analytically, urT5 responds best to direct task verbs such as خلاصہ بنائیں (“write a summary”).

Zero-shot prompt:
مندرجہ ذیل خبر کا جامع اور مختصر خلاصہ تحریر کریں:
{ARTICLE_TEXT}
خلاصہ:

Few-shot prompt:
مثال:
خبر: پاکستان نے نئی معاشی پالیسی کا اعلان کر دیا۔
خلاصہ: نئی معاشی پالیسی کا مقصد ملکی ترقی اور سرمایہ کاری کو فروغ دینا ہے۔

اب آپ اس خبر کا خلاصہ بنائیں:
{ARTICLE_TEXT}
خلاصہ:

mBART Prompt Design (Urdu → English Translation).
To guide translation, mBART needs explicit language tags. For the urT5 → mBART pipeline, the Urdu summaries generated by urT5 were translated using the following

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

prompt:
<ur> {URDU_SUMMARY} </ur>
Translate to English:

mGPT Prompt Design (Urdu Sentiment Classification).
mGPT was prompted using English task instructions combined with Urdu input text for sentiment classification, which demonstrated the most consistent behavior.

Zero-shot prompt:
Read the following review and classify its sentiment as Positive or Negative.
Review:
{REVIEW_TEXT}
Answer only: Positive or Negative.

Few-shot prompt:
Review: یہ فلم بہت شاندار تھی، میں نے بہت لطف اٹھایا۔
Sentiment: Positive
Review: کہانی بہت کمزور تھی اور ایکٹنگ بھی متاثر کن نہیں تھی۔
Sentiment: Negative

Now classify the following review:
{REVIEW_TEXT}
Sentiment:

Proprietary Baseline Prompts (GPT-4, Gemini, Gemma)
Instruction-tuned generalist models were evaluated using standardized English prompts:

Summarization:
Summarize the following Urdu news article and provide the final summary in English.
Ensure the summary is concise and captures the key events accurately.
Article:
{ARTICLE_TEXT}
English Summary:

Sentiment classification:
Determine whether the sentiment of the following Urdu review is Positive or Negative.
Review:
{REVIEW_TEXT}
Sentiment:

Evaluation Metrics
ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and BERTScore-F1 were used to evaluate the performance for cross-lingual summarization.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Accuracy and F1-score were used as evaluation metrics for sentiment classification, following standard practice in binary sentiment analysis [8].

Results and Analysis

Under a **prompt-only evaluation setting**, the quantitative results for cross-lingual summarization and Urdu sentiment classification are presented in this section.

Cross-Lingual Summarization Results

Using ROUGE-1, ROUGE-2, ROUGE-L and BERTScore-F1, cross-lingual summarization performance was evaluated on a subset of 100 articles from the UATS-100K dataset. To directly compare urT5 with models that generate English summaries natively. Its evaluation was conducted using the urT5 → mBART translation pipeline. **Table 1** summarizes the quantitative results across all evaluated model variants.

Table 1. Cross-Lingual Summarization Performance Across Model Variants

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore-F1**
Gemma	0.0915	0.0193	0.0759	0.8302
GPT-4	0.1055	0.0248	0.0853	0.8315
Gemini 2.5 Pro	0.1028	0.0222	0.0783	0.8293
mBART (zero-shot)	0.1006	0.0319	0.0859	0.8231
mBART (few-shot)	0.0506	0.0055	0.0451	0.8409
urT5 → mBART (prompt-only)	**0.2103**	**0.0666**	**0.1680**	**0.8496**

Analysis

Multiple consistent patterns are shown by the results. First, relatively same ROUGE and BERTScore values are achieved by generalist instruction-tuned LLMs (GPT-4, Gemini, Gemma) which demonstrate their strong zero-shot abstraction ability even for Urdu input. However, their limited lexical overlap with reference summaries indicates weak alignment with Urdu specific discourse structure.

On the other hand, significantly higher ROUGE scores are offered by the urT5 → mBART pipeline across all variants. This improvement is because of the strong Urdu-native semantic and morphological foundation of urT5 which enables it to generate summaries that closely align with the format and content of the reference headlines before translation. Effective cross-lingual evaluation is enabled by the subsequent mBART translation which preserves much of this semantic alignment.

Interestingly, independent summarization performance of mBART is not always improved by few-shot prompting. This indicates that task generalization may already be saturated in zero-shot settings for multilingual encoder-decoder models and rather than performance improvement, unrefined few-shot examples may result in stylistic inconsistency.

The evaluation subset is intentionally small, and results reflect relative prompt-level behavior rather than statistically robust performance claims. However, the findings

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

show that language-specific pretraining is important in low-resource summarization tasks [5,10].

6.2 Sentiment Classification Results

Using Accuracy and F1-score, sentiment classification was evaluated on 100 reviews from the IMDb Urdu Reviews dataset. **Table 2** presents the quantitative results across evaluated models.

Table 2. Sentiment Classification Accuracy and F1-Score

Model	Accuracy	F1-Score
mGPT (Zero-Shot)	–	–
mGPT (Few-Shot)	0.5833	0.5404
GPT-4 (Zero-Shot)	0.8600	0.8574
Gemini 2.5 Pro	0.8900	0.8896
Gemma	0.6400	–

Under both zero-shot and few-shot prompting, urT5 failed to generate valid sentiment labels because of its pre-training objective as a generative summarization model rather than a classification model.

Analysis

Generalist LLMs considerably perform better than multilingual prompt-only models for sentiment classification. Strong instruction-following capability and exposure to large-scale labeled sentiment dataset during pre-training of GPT-4 and Gemini are reflected by their high Accuracy and F1-scores.

Moderate performance is exhibited by mGPT under few-shot prompting, but it struggles in zero-shot settings that indicates its limited task abstraction without explicit examples. This behavior aligns with prior findings on multilingual LLMs in classification tasks [3].

Sentiment classification depends more on semantic polarity detection rather than deep morphological generation as compared to summarization, which seems to favor instruction-tuned generalist models over language specific generators. This difference highlights the task-dependent nature of prompt-only adaptation in low-resource contexts.

Cross-Task Comparative Insights

Table 3 displays a qualitative comparison of generalist LLMs and prompt-only Urdu-grounded pipelines to integrate findings across both tasks.

Table 3. Qualitative Comparison of Generalist and Prompt-Only Urdu NLP Pipelines

Dimension	Generalist LLMs	Prompt-Only Urdu Models

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 1 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Semantic abstraction	High	Moderate
Lexical and structural alignment	Moderate	High
Urdu linguistic grounding	Limited	Strong
Resource efficiency	Low	Very high
Best application scenario	Zero-shot deployment	Prototyping and low-resource adaptation

Interpretation

Not a single model variant universally dominates across tasks. While Urdu-specific models demonstrate clear benefits in generation-heavy tasks that need fine-grained linguistic alignment, generalist LLMs perform well in instruction-driven classification tasks. In low-resource languages, prompt-only adaptation emerges as a potential strategy for rapid prototyping and exploratory research, especially when integrated with carefully customized prompts and cross-lingual pipelines.

6.4 Limitations and Future Research Directions

While analyzing the results presented in this study, several limitations should be taken into consideration. First, the experimental evaluation is carried out on small subsets of publicly available datasets and limited to a prompt-only adaptation setting. The performance potential of hybrid or fine-tuned systems is not completely reflected by this framework. However, it enables controlled comparisons across heterogeneous model families. Moreover, the semantic adequacy, cultural appropriateness and stylistic fidelity might not completely reflect as the evaluation mainly relies on automatic metrics.

To overcome current limitations, future research should integrate prompt engineering with supporting techniques:

Prompt Optimization and Calibration: Instability can be reduced and task consistency can be improved by designing context-aware and dynamically adaptable prompts for Urdu-English or bilingual tasks.

Synthetic Data Generation: High-quality Urdu-English parallel dataset can be generated by utilizing advanced models like GPT 3.5 which may help in reducing data scarcity and enhancing model diversity [7].

Lightweight Fine-Tuning (LoRA/ QLoRA): While keeping computational cost minimum, effective fine-tuning techniques can strengthen the cross-lingual alignment of specialist models.

Hybrid Prompt + Fine-Tuning Framework: Balanced, contextually accurate systems can be obtained by combining the fluency of generalist LLMs with domain-specific precision of specialist models like urT5 and mBART.

Online ISSN 3007-3197
Print ISSN 3007-3189

Human-Centered Evaluation: Fairness and authenticity in generated Urdu text can be assured by combining native speaker feedback and cultural evaluation with automated metrics.

Handling Code-Switching: Including mixed Urdu-English input to expand the evaluation will test the adaptability of the model in naturally bilingual settings and reflect real-world usage.

Supplementary Materials: No supplementary materials were provided.

Funding: This research received no external funding.

Data Availability Statement: All publicly available datasets (UATS-100K, IMDb Urdu Reviews) and models are used (GPT 3.5, mGPT, mBART, urT5) and all experimental prompts are included within the article. No new datasets were generated in this study.

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

References

Abid, K., Aslam, N., Fuzail, M., Maqbool, M. S., & Sajid, K. (2023). An efficient deep learning approach for prediction of student performance using neural network. VFAST Transactions on Software Engineering, 11(4), 67-79.

Arif, Samee, et al. "Generalists vs. specialists: Evaluating large language models for urdu." arXiv preprint arXiv:2407.04459 (2024).

Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). UNDERSTANDING URBAN EXPANSION THROUGH MULTI-TEMPORAL SATELLITE DATA ANALYSIS. Kashf Journal of Multidisciplinary Research, 2(09), 252-273.

Basit, A., Hanif, I., Maqbool, M. S., Qayyum, W., Hasnain, M. A., & Nazeer, R. (2023). Cross-Lingual Information Retrieval in a Hybrid Query Model for Optimality. Journal of Computing & Biomedical Informatics, 5(01), 130-141.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901, 2020.

Chen, Banghao, et al. "Unleashing the potential of prompt engineering for large language models." Patterns (2025).

Faisal, F., & Yousaf, U. (2024). LEGAL-UQA: A Low-Resource Urdu-English Dataset for Legal Question Answering. arXiv preprint arXiv:2410.13013.

Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment analysis of omicron tweets by using machine learning models. VFAST Transactions on Software Engineering, 11(1), 67-75.

Furqan, M., Khaja, R. B., & Habeeb, R. (2024). ERUPD--English to Roman Urdu Parallel Dataset. arXiv preprint arXiv:2412.17562.

Hasnain, M. A., Ali, S., Malik, H., Irfan, M., & Maqbool, M. S. (2023). Deep learning-based classification of dental disease using x-rays. Journal of

Online ISSN 3007-3197
Print ISSN 3007-3189

Computing & Biomedical Informatics, 5(01), 82-95.

Hassan, F., Saleem, S., Javed, K., Asim, M. N., Rehman, A., & Dengel, A. (2025). Enhanced Urdu Intent Detection with Large Language Models and Prototype-Informed Predictive Pipelines. arXiv preprint arXiv:2505.07857.

Javed, A., Saeed, U., and Nisar, H., "Abstractive Text Summarization for the Urdu Language: Data and Methods," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 13, no. 2, 2022.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," Proceedings of ACL, pp. 6282–6293, 2020.

Kamran, M., Ali, S., and Hussain, S., "Towards Improving NLP for Urdu and Regional Languages," IEEE Access, vol. 10, pp. 4572–4589, 2022.

Kanwal, F., Abid, M. K., Maqbool, M. S., Aslam, N., & Fuzail, M. (2023). Optimized classification of cardiovascular disease using machine learning paradigms. VFAST Transactions on Software Engineering, 11(2), 140-148.

Kazi, S., Rahim, M., & Khoja, S. A. (2025, January). Crossing Language Boundaries: Evaluation of Large Language Models on Urdu-English Question Answering. In Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (pp. 141-151).

Khoboko, Pitso Walter, Vukosi Marivate, and Joseph Sefara. "Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models." Machine Learning with Applications 20 (2025): 100649.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G., "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," ACM Computing Surveys, vol. 55, no. 9, 2023.

Malik, F., Fuzail, M., Aslam, N., Sarwar, R., Abid, K., Maqbool, M. S., & Yousaf, A. (2024). A Hybrid Machine Learning Model to Predict Sentiment Analysis on X. Journal of Computing & Biomedical Informatics, 6(02), 64-79.

Maqbool, M. S., Fatima, N., Nazeer, R., Aslam, N., Abbas, F., Sumra, U., & Nadeem, M. (2025). A HYBRID DATASET-BASED ENSEMBLE STRATEGY FOR EFFICIENT BREAST CANCER DETECTION. Kashf Journal of Multidisciplinary Research, 2(12), 39-57.

Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2023). Optimized feature extraction and cross-lingual text reuse detection using ensemble machine learning models. Journal of Computing & Biomedical Informatics, 5(01), 26-40.

Munaf, Mubashir, et al. "Low resource summarization using pre-trained language models." ACM Transactions on Asian and Low-Resource Language Information Processing 23.10 (2024): 1-19.

Muneer, I., Waheed, N., Ashraf, M. A., & Adeel Nawab, R. M. (2025). Sentential Cross-lingual Paraphrase Detection for English-Urdu Language Pair. The European Journal on Artificial Intelligence, 38(3), 309-329.

Rafiqee, M. M., Qaiser, Z. H., Fuzail, M., Aslam, N., & Maqbool, M. S. (2023). Implementation of efficient deep fake detection technique on videos dataset using deep learning method. Journal of Computing & Biomedical

Online ISSN 3007-3197
Print ISSN 3007-3189

Informatics, 5(01), 345-357.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T., “Transfer Learning in Natural Language Processing,” Proceedings of NAACL: Tutorials, pp. 15–18, 2021.

Sharif, M., Abbas, Z., Yi, J., & Liu, C. (2024). From Statistical Methods to Pre-Trained Models; A Survey on Automatic Speech Recognition for Resource Scarce Urdu Language. arXiv preprint arXiv:2411.14493.

Shliazhko, D., Nikoulina, V., and Serdyukov, P., “mGPT: Few-Shot Learners Go Multilingual,” arXiv preprint, arXiv:2301.07822, 2023.

Tuz Zuhra, Fatima, and Khalid Saleem. "Towards Development of New Language Resource for Urdu: The Large Vocabulary Word Embeddings." ACM Transactions on Asian and Low-Resource Language Information Processing 24.8 (2025): 1-14.

Zain, M., Hussain, N., Qasim, A., Mehak, G., Ahmad, F., Sidorov, G., & Gelbukh, A. (2025). RU-OLD: A Comprehensive Analysis of Offensive Language Detection in Roman Urdu Using Hybrid Machine Learning, Deep Learning, and Transformer Models. Algorithms, 18(7), 396.