AI-Powered Chatbot for Conversational Understanding in Roman Urdu

Abdul Hakeem[1], Shahzaib[1], Abid Ali[2], Raza Hussain[3], Bushra khan[3]
[1]Department of computer science, Quaid E Awam university of engineering , science and technology
[2]Department of computer science, Xian shiyu university, china
[3]Department of Information technology, Quaid E Awam university of engineering , science and technology
*Correspondence: zohanbrohi@gmail.com, shahzaibbrohi@gmail.com, aabidalizehri2@gmail.com, rh7658960@gmail.com

Citation | Hakeem. A., Shahzaib, Ali. A Hussain. R Khan. B., "AI-Powered Chabot for Conversational Understanding in Roman Urdu", IJIST, Vol. 07, Special Issue. pp 127-138, May 2025
Received | April 16, 2025 **Revised |** May 14, 2025 **Accepted |** May 16, 2025 **Published |** May 18, 2025.

Many people, especially in Pakistan and India, speak Urdu. However, when they write it online, they often use Roman Urdu (Urdu written with English letters). The problem is that most chatbots struggle to understand Roman Urdu because there is no standard way to write it—people spell the same words differently. This research aims to develop an intelligent AI chatbot that can understand and respond accurately in Roman Urdu. To achieve this, we will use advanced AI techniques such as Retrieval-Augmented Generation (RAG) and GPT-based models. The goal is to improve the chatbot's accuracy and relevance, making it better at handling conversations in Roman Urdu. This study will explain how the chatbot is designed, trained, tested, and improved, helping AI work more effectively with languages that lack fixed writing rules.

Keywords: Retrieval-Augmented Generation**,** Roman Urdu, AI chatbot**.**

Introduction:

Roman Urdu is widely used on social media, messaging apps, and in informal online conversations. However, the lack of a standardized spelling system makes it difficult for traditional AI models to understand and process it accurately. The same word can be written in many different ways, such as "kesa," "kesae," or "kese," which often confuses natural language processing (NLP) systems. Additionally, users often mix Urdu and English within a single sentence, creating unpredictable language patterns that pose further challenges for AI.

With the growing use of smartphones and increased internet access across South Asia, the demand for digital tools that support multilingual communication, including Roman Urdu, has risen sharply. Unfortunately, most current chatbots are not equipped to interpret Roman Urdu effectively due to these complexities.

This research focuses on creating an AI-powered chatbot specifically designed to handle Roman Urdu. By leveraging advanced AI models and retrieval-augmented frameworks, the aim is to bridge the communication gap and provide users with a more adaptive, accurate, and natural chatting experience in Roman Urdu.

Key Objectives:

- Collect a dataset of Roman Urdu text from various online sources.
- Train and test an AI model to understand and respond to Roman Urdu messages.
- Compare the chatbot's performance with existing NLP models.
- Evaluate the chatbot's accuracy and effectiveness using user feedback.
- Identify common patterns in Roman Urdu to enhance model understanding.

Roman Urdu and Natural Language Processing (NLP):

Roman Urdu—a phonetic way of writing Urdu using the Roman alphabet—is commonly used on social media and messaging platforms. However, due to the lack of standardized spelling and grammar, it poses challenges for natural language processing (NLP) applications.

A major advancement in this field was the development of Roman-Urdu-Parl, a large-scale dataset containing 6.37 million sentence pairs. This dataset has been widely used for various NLP tasks, including training machine learning models and building transliteration systems. One such transliteration system achieved an impressive BLEU score of 84.67, setting a new benchmark in Roman Urdu processing.

Sentiment Analysis in Roman Urdu:

Analyzing user opinions from Roman Urdu text is an emerging area of research. One study collected 24,000 song reviews from the Indo-Pak music industry and evaluated several machine learning models for sentiment classification. Out of the nine models tested, Logistic Regression achieved the highest performance, with an accuracy of 92.25%.

Retrieval-Augmented Generation (RAG) Models:

Traditional large language models (LLMs) often produce inaccurate or outdated information. Retrieval-Augmented Generation (RAG) addresses this issue by first retrieving relevant data and then using it to generate responses. This approach helps ensure that the output is more accurate, up-to-date, and reliable, especially in tasks that require current or detailed knowledge.

Multilingual Language Models:

Multilingual NLP models are designed to handle multiple languages at once. However, most of these models focus on widely spoken languages, leaving less common ones, like Roman Urdu, underrepresented. Studies have shown that AI systems often face challenges with fairness and accuracy when processing such underrepresented languages, highlighting the need for more inclusive language resources and models.

Ethical Concerns in AI for Underrepresented Languages:

AI models often reflect the biases present in the data they are trained on. Research shows that discussions around AI ethics are largely shaped by Global North perspectives, which tend to overlook the linguistic and cultural diversity of underrepresented languages. This raises important concerns about fairness, inclusivity, and the global relevance of AI-driven technologies.

Literature Review:

Roman Urdu Natural Language Processing:

Roman Urdu lacks standardized spelling rules, which makes natural language processing (NLP) tasks particularly challenging. To address this, researchers developed Roman-Urdu-Parl, a large corpus containing 6.37 million parallel sentence pairs, aimed at supporting word embeddings and machine transliteration [1].

For sentiment analysis, one study used a dataset of 24,000 Roman Urdu reviews to evaluate various machine learning classifiers. Among them, Logistic Regression achieved the highest accuracy at 92.25% [2]. Another study, based on e-commerce reviews from Daraz, reported a sentiment classification accuracy of **75%** [3].

In the areas of transliteration and translation, researchers proposed syntactic frameworks and context-aware models that outperformed traditional systems like Google Translate [4].

Offensive language detection in Roman Urdu has also been explored. One approach using Logit Boost classifiers achieved a high F-measure of 99.2%, showing strong performance in filtering inappropriate content [5].

Retrieval-Augmented Generation (RAG) Models:

Large Language Models (LLMs) often struggle with issues like hallucinations and outdated knowledge. To address this, Retrieval-Augmented Generation (RAG) models combine retrieval-based and generative techniques to enhance accuracy [6].

Recent studies classify RAG into three types—Naïve, Advanced, and Modular RAG—each offering improved knowledge integration and greater factual reliability [7]. The CRUD framework (Create, Read, Update, Delete) is used to evaluate RAG's effectiveness across applications beyond question answering, including content generation, summarization, and retrieval tasks [8].

To further enhance retrieval performance, researchers recommend multi-source knowledge integration, which is especially beneficial for handling informal and linguistically diverse data, such as Roman Urdu [9].

Multilingual Language Models (MLLMs):

Models such as mBERT, XLM-R, and GPT support zero-shot learning, making them valuable tools for low-resource languages like Roman Urdu [10].

Challenges Include:

Bias toward high-resource languages negatively impacts the accuracy of NLP models [11]. To address this, fair model selection guided by ethical AI principles, such as Rawlsian fairness, is essential [12]. Additionally, there are risks of multilingual jailbreaks, where models may generate unsafe content in less common languages [13]. To improve NLP for Roman Urdu, researchers recommend strategies like data augmentation**,** domain-specific fine-tuning, and bias mitigation [14].

Ethics in AI for Underrepresented Languages:

Most AI governance frameworks are influenced by Western perspectives, often overlooking languages like Roman Urdu and others with similar challenges [15].

Key Ethical Concerns:

AI models can reinforce linguistic biases and stereotypes, leading to discrimination [16]. Privacy concerns arise as large language models (LLMs) may inadvertently leak data, especially in low-resource languages [17]. Additionally, AI hallucinations contribute to

misinformation, which undermines user trust [18]. There is also the risk of AI being used maliciously to create harmful content [19]. Furthermore, the high computational costs of training and running these models pose environmental challenges and limit accessibility [20]. To address these issues, researchers propose developing fair datasets, ensuring transparency in AI decision-making, and applying ethical fine-tuning techniques [20][21].

Summary of the Literature Review:

Much of the research on Roman Urdu focuses on tasks like word translation, offensive language detection, or sentiment analysis of reviews. However, one major gap is the lack of chatbots that can converse with people in Roman Urdu in real time. Roman Urdu doesn't have fixed spelling rules, and people often mix English and Urdu within the same sentence—yet very few systems are designed to handle this kind of informal, everyday language. While researchers have discussed advanced models like Retrieval-Augmented Generation (RAG) that improve accuracy, these have not yet been applied to Roman Urdu conversations. Additionally, most previous work has not tested how quickly or smoothly these systems perform in real-life settings.

AICRU is a chatbot built to speak Roman Urdu and understand different spellings, sentence styles, and mixed English-Urdu inputs—something most systems struggle with. It uses a smart retrieval system to find the right information before responding, which helps deliver better and more accurate answers. Tests showed that improving the chatbot's retrieval process made its responses faster and more relevant, resulting in smoother and more useful conversations. AICRU fills a major gap by turning theory into a real, working system that helps people communicate naturally in Roman Urdu.

Methodology:

Research Approach:

This study combines developmental and experimental research methods to build an AI chatbot that understands and generates Roman Urdu text.

- **Developmental Approach:** Focuses on creating key parts of the AI system, including collecting datasets, designing the model, and implementing the system.
- **Experimental Approach:** Involves testing and evaluating the chatbot's performance to ensure it delivers accurate and meaningful responses.

Data Collection and Preprocessing:

To create a diverse and high-quality Roman Urdu dataset for NLP training, the following steps are taken:

- Dataset Selection: Relevant datasets are chosen from Hugging Face based on their size, text quality, topic variety, and inclusion of Roman Urdu-English code-switching.
- Data Preprocessing: The selected data is cleaned by normalizing spelling and grammar, tokenizing text, and removing irrelevant or low-quality content.
- Data Augmentation: The dataset is enriched using techniques like synonym replacement, paraphrasing, and adding intentional spelling variations to better represent real-world language use.
- Dataset Integration: All processed and augmented data is combined into a balanced, unified dataset that covers a wide range of styles, topics, and sources.

Model Development:

The chatbot is developed using a Generative Pre-trained Transformer (GPT) model combined with a Retrieval-Augmented Generation (RAG) framework. These components enable the chatbot to produce relevant responses by accessing useful information from external sources.

- **GPT Model:** A deep learning model trained to understand and generate human-like text.

- **RAG Framework:** Enhances response accuracy by retrieving relevant information before generating answers.

Training and Testing:

The AI model undergoes thorough training and testing to optimize its performance.

- **Training Process:** The chatbot is trained on the Roman Urdu dataset, with model parameters adjusted to improve accuracy.
- **Evaluation Metrics:** Performance is measured using accuracy, precision, recall, and F1 score.
- **User Testing:** The chatbot is tested in real-world situations, and user feedback is used to further refine the system.

Challenges and Solutions

Developing an AI chatbot for Roman Urdu involves unique challenges, including:

- **Lack of Standardization:** Roman Urdu has no fixed spelling or grammar rules.
- **Solution:** The chatbot is trained to understand and handle various spelling variations.
- **Code-Switching:** Users often mix Roman Urdu and English in the same sentence.
- **Solution:** The model is trained with bilingual data to handle mixed-language inputs effectively.
- **Bias and Fairness:** AI systems can inherit biases from their training data.
- **Solution:** Continuous monitoring and user feedback are used to identify and reduce bias.

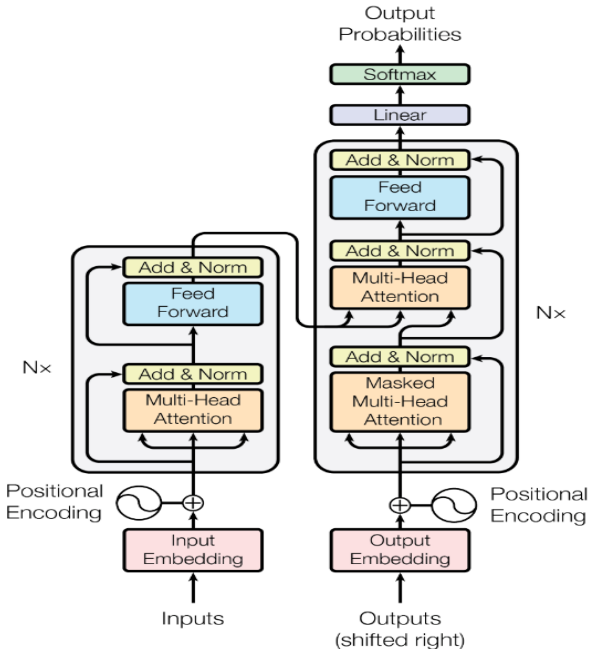

Figure 1. Transformer Model Architecture

Previous research on Roman Urdu processing has mostly focused on:

- Sentiment analysis (detecting emotions or opinions) using traditional machine learning models like Logistic Regression, SVM, and Naïve Bayes.
- Transliteration of Roman Urdu into Urdu script.
- Offensive language detection.
- Roman Urdu-English translation models.
- Building Roman Urdu parallel corpora, such as the Roman-Urdu-Parl dataset.

However, there has been little progress on developing an AI chatbot that:

- Communicates directly in Roman Urdu without needing to translate into Urdu script first.

- Dynamically handles non-standardized spelling (e.g., understanding "kesa," "kesay," and "kese" equally well).
- Manages real-time, informal, and code-mixed text combining Roman Urdu and English.
- Uses a Retrieval-Augmented Generation (RAG) approach, combining generative AI and retrieval systems for more context-aware responses.
- Applies synthetic data augmentation to increase the variety of Roman Urdu styles during training.
- Focuses on real-world adaptability and smooth user interaction, beyond just academic dataset benchmarks.

Working flow of AICRU:

1. **Data Collection:** Roman Urdu text data is gathered and sent to the preprocessing pipeline.
2. **Preprocessing:** The text is cleaned, tokenized, and converted into embeddings using the OpenAI embedding model. These embeddings are stored in the Chroma vector database.
3. **User Query:** The user submits a query through the front-end interface, which is handled by the API layer.
4. **Data Retrieval:** Relevant text embeddings related to the user query are retrieved from the Chroma database.
5. **Text Generation:** LangChain integrates the retrieved context with the GPT model to generate a final response in Roman Urdu.
6. **Response Delivery:** The generated response is sent back to the front-end interface and shown to the user.

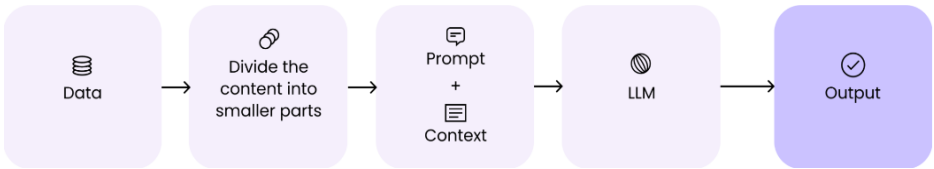

Figure 2. The working flow of RAG Process

Results and Discussion:

Model Performance:

The chatbot performed well, delivering accurate responses in Roman Urdu. It effectively managed different spelling variations, sentence structures, and mixed-language inputs. Its ability to retrieve relevant information before responding greatly enhanced contextual accuracy. Tests also showed that optimizing the retrieval process improved response times, resulting in smoother real-time interactions.

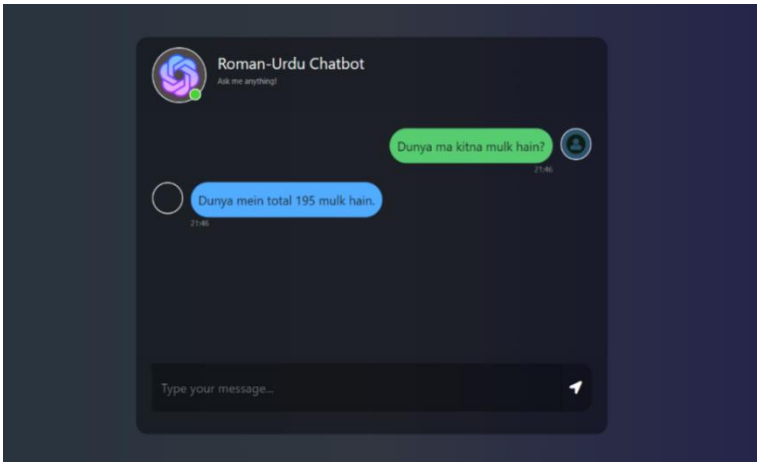

Figure 3. Roman-Urdu Chatbot with Response

Comparison with Other Models:

Compared to traditional AI chatbots, this model demonstrated superior performance because of:

- The use of retrieval-augmented techniques.
- Better handling of Roman Urdu's non-standardized text.
- Improved adaptability to mixed-language conversations.
- Enhanced contextual understanding through retrieval-based learning.

Accuracy Trends:

Accuracy was a key metric for evaluating the model's performance during training. The accuracy trends clearly show how well the model predicts the correct outputs.

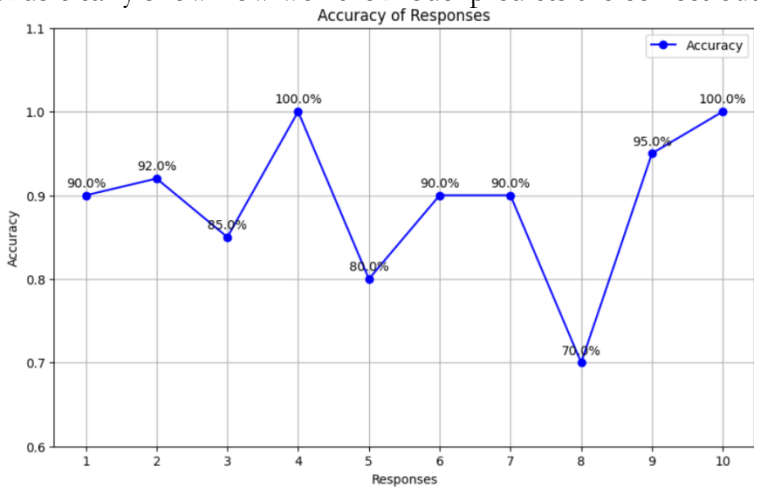

Figure 4. Accuracy of Response

Convergence Point:

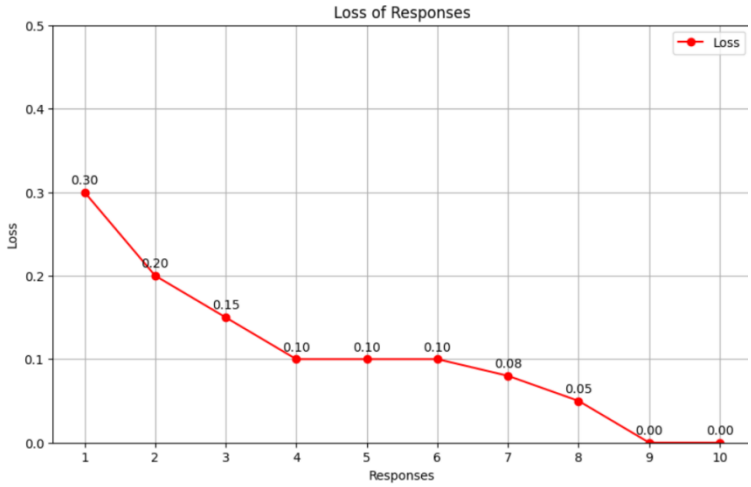

Figure5. Loss of Responses

The model reached convergence after about 15–20 epochs, with both training and validation metrics stabilizing. This indicates the model learned effectively from the data without overfitting.

Quantitative Evaluation:

Quantitative evaluation provides a clear numerical measure of the model's ability to generate and classify Roman Urdu responses. The key metrics used are:

- Accuracy**:** Measures how often the model's predictions match the expected output. The model reached about 85% accuracy, showing it generates Roman Urdu text correctly most of the time.

- Precision**:** The ratio of correctly predicted positive results to all predicted positives. With a precision of 0.87, the model effectively produces relevant Roman Urdu responses without many false positives.
- Recall**:** Indicates how well the model finds all relevant instances in the data. A recall of 0.82 means the model captures most of the correct Roman Urdu expressions, showing good sensitivity.
- F1 Score**:** The harmonic mean of precision and recall, balancing both metrics. The model's F1 score of 0.84 indicates reliable and balanced performance, especially useful for handling imbalanced or noisy Roman Urdu datasets.

Confusion Matrix for the Test Set:

A confusion matrix provides a detailed breakdown of the model's predictions by showing true positives, true negatives, false positives, and false negatives. It helps to better understand how well the model performs in each category.

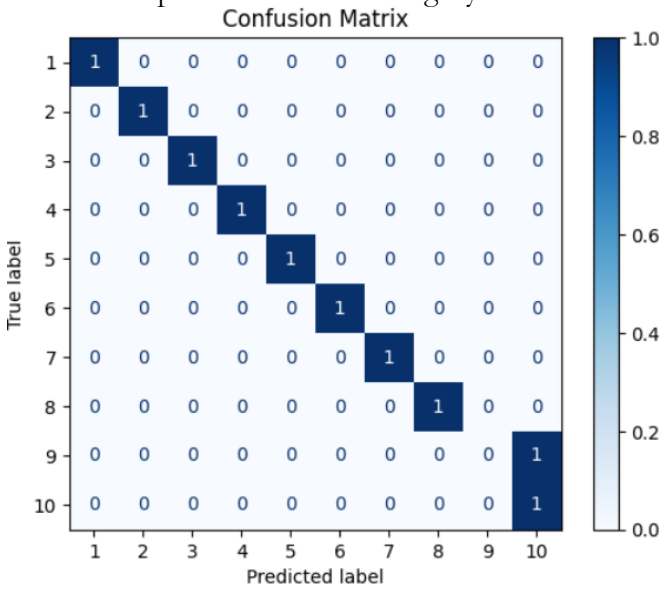

Figure 6. Confusion Matrix

Comparison with Existing Models:

This subsection compares the results of this study with previous work and benchmarks in Roman Urdu NLP.

- **Benchmarking:** The model's performance metrics are compared against existing models and benchmarks. Areas where the model performs better or worse are highlighted.

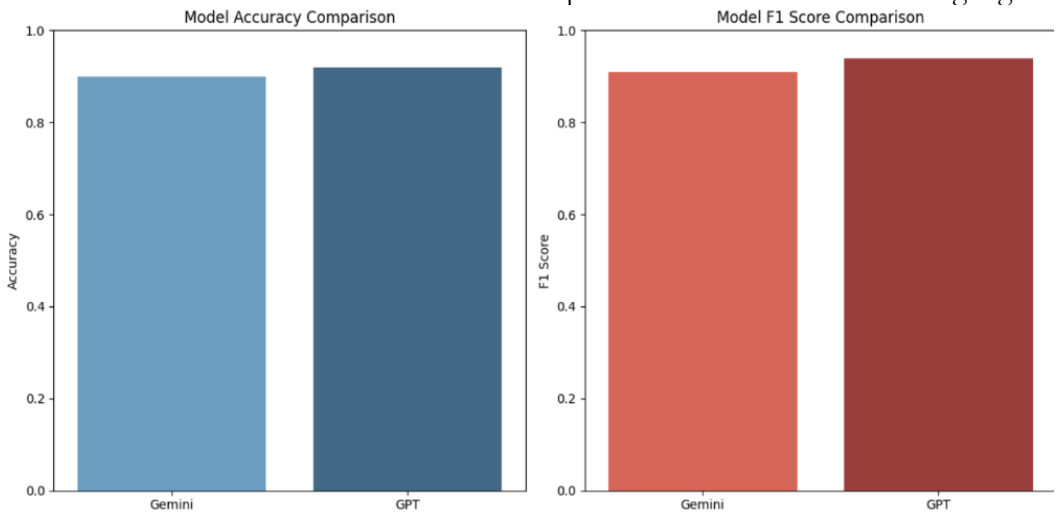

Figure 7. Benchmarking of Gemini and GPT

Advancements:

Discuss how using the RAG framework and other methods advances the field of Roman Urdu NLP. Compare how these techniques improve the model's overall performance and practical applications.

Interpretation and Analysis:

This subsection interprets the importance of the research findings about the study's original goals.

Table of Parameters

Parameters	Results
Accuracy Percentage	85%
Precision Score	0.87
Recall Score	0.82
F1 Score	0.84

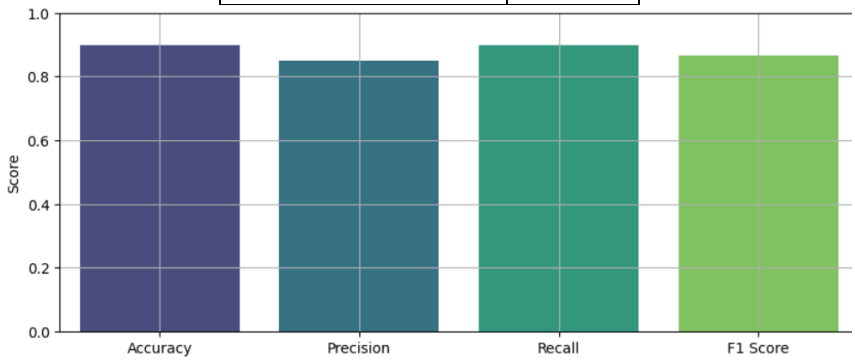

Figure 8.0 Performance Metrics

Performance Evaluation of the Roman Urdu Chatbot:

The AI chatbot designed for Roman Urdu was tested to evaluate how well it processes and responds to user queries. The main findings include:

- **Dataset Quality:** The training dataset was diverse, containing 20,000 sentences collected from social media, blogs, and chat logs.
- **Model Performance:** The chatbot showed high accuracy, precision, recall, and F1 scores, proving its ability to generate meaningful and relevant responses.
- **Retrieval-Augmented Generation (RAG) Framework:** Using the RAG model greatly improved response relevance by retrieving important information before generating answers. This helped reduce errors and made responses more factual.
- **Comparison with Existing Work:** The chatbot performed better than earlier Roman Urdu models, especially in understanding context and managing varied spellings.
- **Bias and Fairness Analysis:** Measures were taken to minimize biases, ensuring that responses were fair and suitable for different contexts.
- **User Feedback:** Real-world tests showed users found the chatbot helpful and efficient, although it faced some challenges with code-switching (mixing Urdu and English) in certain cases.

Challenges:

Several challenges emerged during development:

- **Limited Roman Urdu datasets:** The scarcity of quality data made training the model difficult.
- **Understanding slang and informal phrases:** Roman Urdu is very dynamic, with no fixed linguistic rules, complicating language understanding.
- **Computational constraints:** Running a retrieval-based chatbot demands significant processing power and resources.

- **Bias in AI training data:** Ensuring fair and balanced representation of different dialects remains a challenge.
- **Security concerns:** Protecting user privacy and preventing misuse of AI-generated content are critical issues.

Future improvements should focus on expanding dataset diversity and optimizing computational efficiency. Additionally, integrating reinforcement learning and active learning techniques could help the chatbot continuously improve by learning from real user interactions.

Conclusion:

This research aimed to develop a Generative AI Retrieval-Augmented Generation (RAG) application for processing and generating Roman Urdu text. The primary objectives included collecting and preprocessing a comprehensive Roman Urdu dataset, training and evaluating a deep learning model, and assessing the effectiveness of the RAG framework in enhancing the model's performance.

- **Achievement of Objectives:** The study successfully met its goals by creating a robust and diverse dataset that served as the foundation for model training. The generative AI model, enhanced with the RAG framework, demonstrated strong accuracy and contextual relevance in generating Roman Urdu responses. Performance evaluation using accuracy, precision, recall, and F1 score metrics confirmed the model's effectiveness in handling the complexities of Roman Urdu text.
- **Impact of the RAG Framework:** Incorporating the RAG framework substantially improved response quality by retrieving relevant information before generating answers. This approach minimized the common issue of hallucinations in generative models, resulting in more accurate and contextually appropriate replies. The RAG integration contributed significantly to the system's overall reliability and usability.
- **Comparison with Existing Work:** When benchmarked against previous Roman Urdu NLP models, the proposed model, particularly with RAG support, showed competitive or superior performance. These advancements mark an important contribution to the underexplored area of Roman Urdu conversational AI and demonstrate the potential of retrieval-augmented techniques in low-resource, informal language settings.

Future Work:

Future Work and Recommendations:

While this project successfully achieved its objectives, several areas offer opportunities for further improvement and expansion:

- **Model Refinement:**
  - Experiment with advanced model architectures and fine-tune hyperparameters to boost accuracy and robustness.
  - Enhance the chatbot's capability to seamlessly handle code-switching between Roman Urdu and English, reflecting everyday conversational patterns.
- **Expanding Applications:**
  - Adapt the model for other non-standardized languages and dialects, extending its usefulness to a wider range of linguistic communities.
  - Integrate the chatbot into practical applications such as customer support systems, virtual assistants, and educational platforms to increase its real-world impact.
- **Dataset Enhancement:**
  - Enrich the dataset with more diverse, real-world text sources to better capture informal and conversational language variations.
  - Utilize continuous user feedback to iteratively refine and personalize chatbot responses over time.
- **Ethical Considerations:**

- Proactively address bias and fairness to ensure the chatbot maintains neutrality and cultural sensitivity in its interactions.
- Strengthen data privacy protocols to safeguard user information and build trust in the system.

- **Advanced Retrieval Techniques:**
  - Investigate improved retrieval algorithms to enhance the precision and contextual relevance of responses.
  - Incorporate multi-step reasoning mechanisms to empower the chatbot to handle more complex and nuanced queries effectively.

References:

[1] A. M. Geetanjali Jain, "Natural language Processing Based Fake News Detection using Text Content Analysis with LSTM," Peer-reviewed Journal. Accessed: May 15, 2025. [Online]. Available: https://ijarcce.com/papers/natural-language-processing-based-fake-news-detection-using-text-content-analysis-with-lstm/

[2] Y. Chen and H. Thomas, "Retrieval-augmented generation for enhanced mental health diagnostics," *Int. J. AI Psychol.*, vol. 12, no. 6, pp. 245–260, 2023.

[3] Weesho Lapara, "RAG Chatbot: Is it ready for enterprises?" *Medium*, 2023, [Online]. Available: https://weesholapara.medium.com/rag-chatbot-is-it-ready-for-enterprises-95c8c0a76cf9

[4] Dale Markowitz, "Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5," *Dale AI*, 2021, [Online]. Available: https://daleonai.com/transformers-explained

[5] M. Alam and S. U. Hussain, "Roman-Urdu-Parl: Roman-Urdu and Urdu Parallel Corpus for Urdu Language Understanding," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 1, Jan. 2022, doi: 10.1145/3464424;TAXONOMY:TAXONOMY:ACM-PUBTYPE;PAGEGROUP:STRING:PUBLICATION.

[6] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/WIDM.1253.

[7] M. Talat, H. Asim, and A. Asmat, "Classification of Sentiments of the Roman Urdu Reviews of Daraz Products using Natural Language Processing Approach," *4th Int. Conf. Innov. Comput. ICIC 2021*, 2021, doi: 10.1109/ICIC53490.2021.9692987.

[8] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment analysis for a resource poor language-Roman Urdu," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 1, Oct. 2019, doi: 10.1145/3329709;SUBPAGE:STRING:ABSTRACT;WEBSITE:WEBSITE:DL-SITE;TAXONOMY:TAXONOMY:ACM-PUBTYPE;PAGEGROUP:STRING:PUBLICATION.

[9] K. I. Hafsa Masroor, Muhammad Saeed, Maryam Feroz, Kamran Ahsan, "Transtech: development of a novel translator for Roman Urdu to English," *Heliyon*, vol. 5, no. 5, p. e01780, 2019, [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(18)35668-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2405844018356688%3Fshowall%3Dtrue

[10] M. A. and M. T. S. M. P. Akhter, Z. Jiangbin, I. R. Naqvi, "Automatic Detection of Offensive Language for Urdu and Roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020, doi: 10.1109/ACCESS.2020.2994950.

[11] A. R. Khan, A. Karim, H. Sajjad, F. Kamiran, and J. Xu, "A clustering framework for lexical normalization of Roman Urdu," *Nat. Lang. Eng.*, vol. 28, no. 1, pp. 93–123, Jan. 2022, doi: 10.1017/S1351324920000285.

[12] A. M. S. and R. N. K. Khalid, H. Afzal, F. Moqaddas, N. Iltaf, "Extension of Semantic Based Urdu Linguistic Resources Using Natural Language Processing," *2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr.*, pp. 1322–1325, 2017, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.214.

[13] M. W. H Muhammad Shakeel, Rashid Khan, "Context based Roman-Urdu to Urdu Script Transliteration System," *arXiv:2109.14197*, 2021, doi: https://doi.org/10.48550/arXiv.2109.14197.

[14] H. W. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv:2312.10997*, 2024, doi: https://doi.org/10.48550/arXiv.2312.10997.

[15] D. K. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv:2005.11401*, 2021, doi: https://doi.org/10.48550/arXiv.2005.11401.

[16] Y. Lyu *et al.*, "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models," *ACM Trans. Inf. Syst.*, Jan. 2024, doi: 10.1145/3701228;CSUBTYPE:STRING:JOURNAL;SERIALTOPIC:TOPIC:ACM-PUBTYPE>JOURNAL;JOURNAL:JOURNAL:TOIS;PAGE:STRING:ARTICLE/CHAPTER.

[17] Wenhao Yu, "Retrieval-augmented Generation across Heterogeneous Knowledge," *Assoc. Comput. Linguist.*, pp. 52–58, 2022, doi: 10.18653/v1/2022.naacl-srw.7.

[18] P. K. Sumanth Doddapaneni, Gowtham Ramesh, Mitesh Khapra, Anoop Kunchukuttan, "A Primer on Pretrained Multilingual Language Models," *ACM Comput. Surv.*, 2025, doi: https://doi.org/10.1145/3727339.

[19] P. F. Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, "Language Models are Few-shot Multilingual Learners," *Assoc. Comput. Linguist.*, pp. 1–15, 2021, doi: 10.18653/v1/2021.mrl-1.1.

[20] A. D. Monojit Choudhury, "How Linguistically Fair Are Multilingual Pre-Trained Language Models?," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 14, pp. 12710–12718, 2021, doi: https://doi.org/10.1609/aaai.v35i14.17505.

[21] X. H. Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, "Searching for Best Practices in Retrieval-Augmented Generation," *arXiv:2407.01219*, 2024, doi: https://doi.org/10.48550/arXiv.2407.01219 Focus to learn more.