# BanglAssist: A Bengali-English Generative Al Chatbot for Code-Switching and Dialect-Handling in Customer Service

Francesco Kruk fkruk01@ethz.ch ETH Zurich Zurich, Switzerland Savindu Herath sherath@ethz.ch ETH Zurich Zurich, Switzerland Prithwiraj Choudhury pchoudhury@hbs.edu Harvard Business School Boston, MA, USA

#### **ABSTRACT**

In recent years, large language models (LLMs) have demonstrated exponential improvements that promise transformative opportunities across various industries. Their ability to generate human-like text and ensure continuous availability facilitates the creation of interactive service chatbots aimed at enhancing customer experience and streamlining enterprise operations. Despite their potential, LLMs face critical challenges, such as a susceptibility to hallucinations and difficulties handling complex linguistic scenarios, notably code switching and dialectal variations. To address these challenges, this paper describes the design of a multilingual chatbot for Bengali-English customer service interactions utilizing retrievalaugmented generation (RAG) and targeted prompt engineering. This research provides valuable insights for the human-computer interaction (HCI) community, emphasizing the importance of designing systems that accommodate linguistic diversity to benefit both customers and businesses. By addressing the intersection of generative AI and cultural heterogeneity, this late-breaking work inspires future innovations in multilingual and multicultural HCI.

#### **CCS CONCEPTS**

- Human-centered computing → Natural language interfaces;
- $\bullet \ Computing \ methodologies \rightarrow Natural \ language \ generation;$
- Social and professional topics → Cultural characteristics;
- $\bullet$  Information systems  $\rightarrow$  Multilingual and cross-lingual retrieval.

#### **KEYWORDS**

Generative AI, Chatbot, Code Switching, Dialect Handling, Multilingual Retrieval, Cross-Lingual Retrieval, Customer Service

#### **ACM Reference Format:**

Francesco Kruk, Savindu Herath, and Prithwiraj Choudhury. 2025. BanglAssist: A Bengali-English Generative AI Chatbot for Code-Switching and Dialect-Handling in Customer Service. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3706599.3720226

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI EA '25, April 26-May 1, 2025, Yokohama, Japan © 2025 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-1395-8/2025/04. https://doi.org/10.1145/3706599.3720226

#### 1 INTRODUCTION

The rapid evolution of generative artificial intelligence (GenAI) has profoundly impacted the field of human-computer interaction (HCI), particularly the way users interact with natural language processing (NLP) systems [30]. At the heart of this transformation are large language models (LLMs), which demonstrate an unprecedented ability to generate coherent and contextually appropriate text across a wide range of applications through extensive pre-training on diverse datasets [16]. One domain that stands to gain significantly from these advancements is customer service, where personalized, contextualized, accessible, and scalable communication is crucial [27].

However, the deployment of LLMs in customer service faces several substantial challenges. First, LLMs are prone to generating factually incorrect information, a phenomenon commonly referred to as hallucination [17]. Hallucinations occur due to the probabilistic nature of the LLM generation mechanism [17]. Second, these models often struggle with linguistically complex scenarios, which are prevalent in multilingual and culturally diverse contexts [5]. These struggles include the inability to understand specific dialects, such as South Asian Englishes (SAsE) [10], and the combination of different languages within the same utterance, a phenomenon known as code mixing or switching (CSW) [19]. Since multiculturalism and multilingualism are global phenomena [4], these shortcomings represent a serious threat to the democratization of GenAI and the accessibility of this technology for a universal audience. One disadvantaged group in this regard is found in the Bengal region, which includes Bangladesh and the Indian states of West Bengal and Assam. With an approximate population of 300 million, Bengalis are the third-largest ethnic group on the planet [6, 7]. The linguistic interplay used in this region, often referred to as "Banglish" [20], creates unique interaction challenges for GenAI systems, which are known to underperform with languages and dialects less represented in their training dataset [1]. Users in these settings face suboptimal experiences with traditional chatbots, including the chatbot's misinterpretation and reduced usability and diminished trust on the user's side [5]. Addressing these shortcomings requires a user-centric approach that prioritizes the needs and linguistic behaviors of such populations to improve on the two main characteristics that question-answering tools should satisfy, namely answer correctness and answer delivery [30].

To overcome hallucination and linguistic challenges, we developed and evaluated *BanglAssist*: a multilingual GenAI customer service agent designed to enhance user experience in Banglish interactions. By employing retrieval-augmented generation (RAG),

BanglAssist integrates the generative strengths of LLMs with a twostep retrieval pipeline to ground responses in factual and contextually relevant information and deliver correct answers. Frequently asked questions (FAQs) provided by a streaming service provider (referred to hereafter by the pseudonym *MultilingualCo*) serve as the factual basis for the responses of BanglAssist. MultilingualCo belongs to one of Eastern India's largest entertainment companies, and its content library contains more than 800 titles. With roughly 2 million active users and hundreds of requests per week, their customer service deals with requests formulated in English, Bengali, and Banglish on a daily basis. To address MultilingualCo's need for an adaptive and scalable system that can handle CSW and dialectal variations effectively, BanglAssist incorporates role play and few-shot prompting. These ensure that BanglAssist always acts in the name and interest of MultilingualCo and consistently matches the language and script of the user. The guiding principles for its development focus on deterministic tasks and adapting to the user's language, which reduces the user's cognitive effort during the problem resolution and enhances their experience.

Preliminary evaluations of BanglAssist were performed using quantitative and qualitative studies based on exported customer service logs provided by MultilingualCo. Quantitative analysis was conducted to evaluate BanglAssist's retrieval performance, while qualitative analysis was used to assess its generation performance. The overall answer accuracy of 0.81 attests to the effectiveness of BanglAssist. It correctly answered in the query language (Bengali, English, and Banglish) and script 100% of the time. Our retrieval results also highlight the effectiveness of translating multilingual queries into English before retrieval, even with the application of multilingual models, especially when addressing the complexities of CSW. Finally, our findings underscore the potential for GenAI systems to adapt to complex linguistic environments through appropriate design when paired with high-quality contextual data that focus on information correctness and exhaustiveness rather than form and delivery. An example query and answer provided by BanglAssist can be found in Appendix A.

The main contribution of this study is to advance the current HCI discourse by leveraging NLP technologies tailored to address the needs of communities less represented within the global digital and linguistic landscape. This work inspires the development of personalized, contextualized, accessible, and scalable multilingual and multicultural solutions, thereby enhancing inclusivity and equity in digital interactions.

#### 2 RELATED WORK

#### 2.1 GenAI Chatbots in Customer Service

The recent surge in GenAI tools has garnered significant attention from industry professionals due to their potential to automate tasks and improve human productivity [2]. Among these applications, customer service has become the main focus in businesses, drawing considerable scholarly interest [2, 31]. Unlike rule-based chatbots, which rely on predefined rules and struggle with varying conversation tones or atypical scenarios, generative chatbots can adapt to user sentiments and manage a broader range of requests, thereby improving the likelihood of successful customer interactions [2]. As a prime example of HCI, the application of chatbots in customer

service has prompted a growing body of HCI research exploring new design needs and opportunities emerging from this technology [22].

Recent research studying the design needs of GenAI in customer service applications calls for expanding its use beyond traditional collaborative interactions [30] and creating environments that promote user engagement with customer service chatbots [27]. Shi et al. [2023] suggest extending GenAI applications beyond collaborative tasks to deterministic ones, where models incorporate environmental information without user input. In our solution, this concept is realized through a retrieval pipeline that equips BanglAssist with contextual information from MultilingualCo to deliver relevant responses. Additionally, Rapp et al. [2023] emphasize the importance of creating environments that maintain user engagement with customer service chatbots. Our approach explores this by leveraging adaptive language matching to enhance conversational engagement and effectiveness.

# 2.2 Drawbacks of GenAI Models in Customer Service Applications

LLMs are prone to generating responses that appear accurate but may contain entirely false information, a phenomenon often referred to as hallucination [9]. This occurs because LLMs rely on probability distributions to produce answers. When faced with queries outside the scope of their training data or context, the models generate responses based on the most likely predictions, which, while seemingly plausible, can be entirely incorrect [17]. If left undetected, these inaccuracies can mislead users and erode their trust in the chatbot [17].

Moreover, language understanding is critical for a GenAI chatbot serving multilingual and multicultural customers, as effective communication across diverse linguistic and cultural contexts is fundamental to providing personalized and accessible customer service. However, at the present state of the technology, language understanding in LLMs presents a significant challenge in integrating them into customer service applications such as chatbots. The global adoption of language models (LMs) and their application in international contexts have led to the development of multilingual models capable of handling multiple languages simultaneously. A key objective of these models is to transfer knowledge from high-resource languages, which are well represented in training data, to low-resource languages, which are underrepresented, striving for comparable performance across all languages [8, 37].

Despite these developments, multilingual models continue to struggle with tasks involving CSW and dialects such as SAsE [12, 36, 39]. CSW, which involves the mixing of multiple languages within a sentence, is a prevalent phenomenon [19]. The global rise of English as a *lingua franca* has also contributed to the formation of distinct dialects, including SAsE, spoken in countries such as Afghanistan, Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka [10, 38]. Although some studies have explored the application of LMs to these dialects [12], comprehensive empirical analyses remain scarce. The poor performance of LLMs in handling CSW and dialects poses a significant threat to the democratization of GenAI, limiting its accessibility to a universal audience and marginalizing certain communities. See Figure 2 in Appendix B for a Banglish

example. Addressing these shortcomings is essential to ensure that GenAI technologies are inclusive and equitable for all users.

While prior research has evaluated LLMs and embedding models on existing datasets in the context of CSW [36, 39], our study aims to provide fresh insights by examining the performance of state-of-the-art models in real-world applications like customer service. Specifically, we focus on Banglish, a blend of Bengali and English, which exemplifies the complexities of CSW [20].

### 2.3 Improvement Techniques for GenAI Models

While predictive AI models require hyperparameter tuning to adapt their performance to specific downstream tasks [11], LLMs have enabled the use of instructions formulated in natural language, commonly known as prompts, to achieve the same goal [3, 32]. To address hallucinations and poor performance when dealing with CSW and dialects, two workarounds have been proposed: 1) manipulating the prompt to provide task-specific instructions and relevant context [18, 28] and 2) fine-tuning the pre-trained model on a task-specific dataset [40].

Some of the most common prompt engineering techniques include few-shot prompting and chain-of-thought (CoT) prompting, in which the user instructs the LLM to execute a task by providing examples of how it should do so and "a coherent series of intermediate reasoning steps that lead to the final answer for a problem," respectively [34]. In our solution, we applied role-play prompting [14] to ensure that BanglAssist acts as MultilingualCo's customer service agent and few-shot prompting to improve its ability to match the user's language and script. RAG has also emerged as a prompting technique to provide the LLM with additional, task-specific information and reduce the risk of hallucinations [18]. BanglAssist presents a two-step retrieval pipeline with initial vector search and subsequent reranking to provide the LLM with the most relevant context based on the user query.

While prompting can be a very effective and efficient technique, fine-tuning can improve a model's performance by adding a training step to pre-trained models on task-specific datasets [40]. In the case of language-specific tasks, this could imply training the models on datasets containing question-answer pairs in said languages or sentences with the same meaning but written in different languages or scripts. While fine-tuning can be very powerful and have a significant impact, it also requires notable amounts of high-quality data [33]. Because of a lack of such data for Banglish sentences, we chose to focus on prompt engineering in this work.

BanglAssist builds on a wide range of literature and research to leverage recent advancements in GenAI and prompt engineering to provide factually grounded answers to complex multilingual queries, ensuring appropriate delivery by matching their language and script. By doing so, our work addresses key limitations in current customer service chatbots, thereby contributing to the democratization and broader adoption of GenAI within the field.

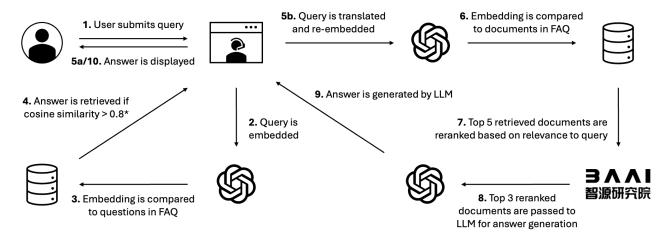
#### 3 ARCHITECTURE AND DESIGN

BanglAssist is designed to handle multilingual customer service requests in complex linguistic scenarios, particularly Banglish. The LLM *GPT-40* by OpenAI [25](hereafter referred to as "GPT") was chosen as the backbone for BanglAssist because of its outstanding

linguistic understanding. To ensure the correctness of the answers, BanglAssist employs a RAG pipeline composed of an initial retrieval of five relevant documents based on cosine similarity and reranking these documents to retrieve the three most relevant ones. The exact prompt used can be found in Appendix C. To further reduce hallucinations and ensure strict adherence to instructions, GPT's temperature was set to 0, resulting in optimal language and script alignment. The user interface (UI) was developed with the *Streamlit* open-source library which facilitates rapid prototyping in Python [13]. Figure 4 in Appendix D showcases several UI screenshots.

Before submitting their query, users are presented with three FAQs randomly sampled from a list provided by MultilingualCo (see Figure 4.a). This dataset includes question-answer pairs in English and Bengali that cover multiple topics concerning MultilingualCo's platform and services. It is important to note that the English and Bengali pairs are not exact translations of each other and partially cover different questions. Specifically, 61% are written in English while the remaining 39% are written in Bengali. These question-answer pairs also serve as the fact base to enhance GPT's context during the answer generation. Upon selection of a question through the interface, the corresponding answer is displayed in the interface without requiring model inference (see Figure 4.b). If users do not find the three questions relevant, they can submit their own requests. Figure 1 provides an overview of BanglAssist's pipeline to process these requests.

After receiving the user query (written in English, Bengali, or Banglish) through the UI (1), the query is embedded using textembedding-3-large by OpenAI [24](2). The query embedding is then compared with the embedding of the questions included in the FAQ list (3). The latter embeddings are computed before inference using text-embedding-3-large again. A hit is achieved if the cosine similarity between the query and the FAQ question embeddings is greater than 0.8 (4). This value was determined empirically to ensure that only queries deemed equivalent to questions in the FAQ list result in a match. When a match occurs, the answer to the question with the highest similarity to the user query is directly displayed in the UI (5a), minimizing inference costs (screenshot shown in Figure 4.b). If there are no hits, the user query is translated into English using GPT and embedded again using text-embedding-3-large (5b). Because queries can be formulated in various languages (multilingual case) or even using different languages in the same sentence (CSW case), the query translation to English can be seen as a normalization step to ensure accurate and effective retrieval of relevant context for the subsequent LLM answer generation. This step addresses the inability of embedding and reranker models to correctly encode Banglish text (see Figure 2). The embedded (English) query is then compared to the embeddings of a combined string including both the questions and answers for each entry in the FAQ list (6). Through cosine similarity, five relevant documents are retrieved from the FAQ list depending on their similarity to the (English) query (7). These are passed to the reranker model bge-reranker-v2-m3 created by the Beijing Academy of Artificial Intelligence (BAAI) [23], which identifies the three most relevant question-answer pairs among the selection (8). This context is passed to GPT for the generation of the final answer (9), which is displayed to the user through the interface (10)(screenshot shown in Figure 4.c).



<sup>\*</sup>If cosine similarity > 0.8, continue with 5a. Else, continue with 5b.

Figure 1: Pipeline of the multilingual customer service chatbot BanglAssist

#### 4 EVALUATION AND RESULTS

To evaluate BanglAssist's performance before deploying it, we used a dataset composed of prior customer service data provided by MultilingualCo. For this preliminary evaluation, we selected 20 queries, of which six were written in Bengali using Bengali script, nine in Bengali using English script (representing the Banglish scenario), and five predominantly in English (representing the SAsE scenario). A more detailed breakdown of the evaluation queries can be found in Appendix E.

### 4.1 Retrieval Evaluation

First, we quantitatively evaluated the retrieval of context by measuring the context's cosine similarity to the query, followed by the reranking of the retrieved context based on its relevance to the query. The evaluation was conducted using three metrics: <code>precision@k</code>, <code>recall@k</code>, and <code>MRR@k</code> (see Appendix F for the formulas of each metric). Specifically, <code>k</code> was set to 5 to measure the retrieval results before reranking and to 3 to measure the reranker's effectiveness. Because all queries were translated to English before retrieval, the evaluation covers all queries indistinctively of their original language.

Table 1 summarizes the retrieval evaluation results. The first section includes the results before reranking with k=5 to isolate the first retrieval performance while the second section includes the results both before and after reranking for k=3 to assess the impact of reranking. To determine the number of relevant FAQ entries for each query, GPT was used because of its advanced language understanding and the volume of the task. This analysis resulted in five queries without relevant context which were therefore only considered in the generation evaluation. Moreover, one query was answered through direct retrieval and was hence also excluded from the evaluation. It is important to note that the number of relevant documents per query varied from more than 5 (highest

Table 1: Overall average retrieval scores before reranking (k = 5 and k = 3) and after reranking (k = 3)

Step	Precision@k	Recall@k	MRR@k
k = 5: Retrieval per	formance evaluat	ion	
Before reranking	0.57	0.42	0.85
k = 3: Reranking pe	erformance evalu	ation	
Before reranking	0.74	0.35	0.85
After reranking	0.69	0.35	0.86

was 25) to fewer than 3 (lowest was one), making it impossible for some queries to achieve maximum precision@k and recall@k scores.

#### 4.2 Generation Evaluation

Second, we qualitatively evaluated the answer-generation capability of BanglAssist. We first assessed whether the answers matched the language and script of the questions (Lang. match) and then how many answers were retrieved based on the initial FAQ check rather than being generated (Gen. rate). Finally, we computed the accuracy of the answers relative to the provided context, evaluating the extent of deviation and potential hallucination. To illustrate this process, we assigned a score of 1 to answers directly retrieved from the FAQ list, while responses that lacked any reference to the context were assigned a score of 0. For queries lacking relevant context, we deemed answers accurate if they informed the user of the inability to provide an answer, requested additional context, or offered to connect the user with a human operator, as outlined in the prompt instructions. To determine the exact score, we applied a penalty of 0.2 for answers containing information not present

Table 2: Average answer generation scores per language

Language	Lang. Match	Gen. Rate	Accuracy
Bengali	1.00	0.83	0.92
English	1.00	1.00	0.68
Banglish	1.00	1.00	0.84

in the context and a penalty of 0.3 for answers that deviated from the instructions in the prompt. However, responses that diverged from MultilingualCo's answers but adhered to the provided context were still classified as accurate. Table 2 summarizes our generation evaluation results.

#### 5 DISCUSSION

## 5.1 Implications of the Retrieval Evaluation

Based on the retrieval results in Table 1, we can conclude that the reranker slightly improves MRR@3 from 0.85 to 0.86, indicating a retrieval improvement. However, the decrease in precision@3 from 0.74 to 0.69 seems to counter this, suggesting a worse performance after reranking. It is therefore important to consider what these two metrics indicate and which one is more relevant for our application. While a higher MRR indicates that the first relevant document in the selection is being correctly ranked higher by the reranker, a lower precision implies that less relevant documents are being selected after reranking. This means that reranking tends to improve results for queries that can be answered by a single document but likely worsens the performance when dealing with queries that require multiple inputs to be resolved. This implies that companies with an extensive knowledge base at their disposal will likely benefit from reranking results before generation while companies with limited context could suffer from it and lose valuable information to correctly answer user requests.

Finally, breaking down the retrieval results by query language does not show a significant difference between queries written in English and queries written in Banglish (for the breakdown, see Appendix G). While these results do not conclusively suggest that the tool can handle certain languages better than others (after all, the queries were translated to English, so from BanglAssist's perspective, they were all written in the same language), it can be concluded that our solution does not perform worse on particularly complex linguistic scenarios such as Banglish than it does for English. This is an important insight, as it shows the potential of BanglAssist to process any kind of language through the simple act of "normalizing" all queries by translating them into English.

However, adding a translation step through LLMs may also introduce new biases and risks associated with the loss of the original meaning [29]. Although we did not notice any particular negative impact caused by the translation step in our results, a thorough evaluation of the translation step could help mitigate the risks associated with it and prevent it from affecting the user experience [21, 35].

#### 5.2 Implications of the Generation Evaluation

Out of the 20 queries we examined, only one led to an answer directly retrieved from the FAQ list and not generated by GPT. Because of a cosine similarity slightly greater than 0.8, this proves that the check works as intended and that the threshold of 0.8 was chosen correctly to ensure that only questions considered equivalent to the ones included in the dataset result in a hit. On a larger volume of queries, this could lead to significant cost savings, since embedding and retrieval are much cheaper than LLM inference (almost 100 times cheaper, according to OpenAI's pricing website [26]). The trade-off with this cost reduction is the reduced personalization and adaptation of the answer to the exact user query.

In our evaluation, BanglAssist consistently matched the language (Bengali or English) and script (Bengali or Roman) used in the customer query. In comparison, MultilinugalCo's answers were always exclusively written in English. This proves the effectiveness of our solution in creating a more personalized, adaptive, and inclusive customer service environment that can potentially lead to higher user engagement and satisfaction.

Finally, we observed a relatively high answer accuracy, especially for Bengali and Banglish queries. This indicates the effectiveness of BanglAssist in creating an accessible and contextualized environment for multicultural and multilingual users. While the answer adherence to the retrieved context was high, we noticed that in some cases, GPT suggested contacting the customer support email address or phone number rather than suggesting connecting the user with a human operator as instructed. This was observed especially in answers for which the context contained these instructions, which likely overwrote the prompt. Lastly, in some answers, GPT added a generic salutation in the end, including a placeholder for the customer service operator's name. While the reason for this behavior cannot be assessed as it is likely related to GPT's training, it highlights the weakness of this model to act as an independent agent and its tendency to create template-like content for users.

#### 5.3 Limitations and Future Work

Our analyses highlight the potential of BanglAssist in dealing with complex customer service queries, as well as its limitations that provide opportunities for further improvements. While the reranking results are promising, the latency caused by this step could outweigh the benefits. A future version of the tool might exclude this step, making the tool faster and computationally cheaper. We also acknowledge the limited amount of context at our disposal to inform BanglAssist's responses. Our generation and retrieval results potentially indicate the need for higher-quality context, covering more topics. We also conclude from our qualitative analysis that said context should only focus on the information quality and not the delivery, as it might interfere with BanglAssist's prompt. Finally, we are aware of the limitation of our evaluation method given the volume and variety of processed queries. We therefore plan on deploying a refined version of BanglAssist to test its performance in a real-world setting.

#### 6 CONCLUSION

In this paper, we have explored the development and evaluation of a multilingual chatbot, BanglAssist, designed to address complex linguistic challenges such as CSW and dialectal variations in Bengali-English customer service interactions. By leveraging a retrieval pipeline that aligns generated responses with contextual information from a curated FAQ database (RAG) and targeted prompt engineering, we demonstrated the chatbot's capability to provide contextually accurate and linguistically adaptive responses. The quantitative and qualitative evaluation episodes provided insights into the system's precision, recall, relevance, and reliability, contributing to a broader understanding of GenAI applications in real-world scenarios. This research offers valuable contributions to the HCI community, emphasizing the critical need for inclusive design in AI-driven customer service solutions to accommodate linguistic diversity, which would improve user engagement and accessibility in multilingual and multicultural settings. By addressing the shortcomings of existing models, our work not only advances the state of chatbot technology but also promotes the democratization of GenAI, ensuring it serves a wider, more diverse audience. Future work could extend these findings by exploring additional linguistic contexts and further refining retrieval and generation mechanisms to enhance performance across a broader spectrum of languages and applications.

#### REFERENCES

- Christina Alexandris. 2024. GenAI and Socially Responsible AI in Natural Language Processing Applications: A Linguistic Perspective. In Proceedings of the AAAI Symposium Series, Vol. 3. 330–337.
- [2] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. Generative AI at work. Technical Report. National Bureau of Economic Research.
- [3] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXiv preprint arXiv:2310.14735 (2023). Retrieved from https://arxiv.org/abs/2310.14735.
- [4] Sylvia Xiaohua Chen, Verónica Benet-Martínez, and Michael Harris Bond. 2008. Bicultural Identity, bilingualism, and psychological adjustment in multicultural societies: immigration-based and globalization-based acculturation. *Journal of personality* 76, 4 (2008), 803–838.
- [5] Yunjae J Choi, Minha Lee, and Sangsu Lee. 2023. Toward a Multilingual Conversational Agent: Challenges and Expectations of Code-mixing Multilingual Users. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–17.
- [6] Central Intelligence Agency (CIA). 2025. Field Listing Ethnic groups. Retrieved January 18, 2025 from https://www.cia.gov/the-world-factbook/field/ethnic-groups/.
- groups/.

  [7] Wikipedia contributors. 2025. Bengalis Wikipedia, The Free Encyclopedia. Retrieved January 18, 2025 from https://en.wikipedia.org/w/index.php?title=Bengalis&oldid=1268353351.
- [8] Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. arXiv preprint arXiv:2107.00676 (2021). Retrieved from https://arxiv.org/abs/2107.00676.
- [9] Carla Ferraro, Vlad Demsar, Sean Sands, Mariluz Restrepo, and Colin Campbell. 2024. The paradoxes of generative AI-enabled customer service: A guide for managers. Business Horizons (2024).
- [10] Ravinder Gargesh. 2019. South Asian Englishes. The handbook of world Englishes (2019), 105–134.
- [11] Savindu Herath Pathirannehelage, Yash Raj Shrestha, and Georg von Krogh. 2024. Design principles for artificial intelligence-augmented decision making: An action design research study. European Journal of Information Systems (2024), 1–23.
- [12] Faye Holt, William Held, and Diyi Yang. 2024. Perceptions of Language Technology Failures from South Asian English Speakers. In Findings of the Association for Computational Linguistics ACL 2024. 4067–4081.
- [13] Snowflake Inc. 2025. Streamlit A faster way to build and share data apps. Retrieved January 20, 2025 from https://streamlit.io/.

- [14] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 4099–4113.
- [15] Dennis Kurzon. 2010. Romanisation of Bengali and other Indian scripts. Journal of the Royal Asiatic Society 20, 1 (2010), 61–74.
- [16] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a humanai collaborative writing dataset for exploring language model capabilities. In Proceedings of the 2022 CHI conference on human factors in computing systems.
- [17] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. Hill: A hallucination identifier for large language models. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–13.
- [18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020), 9459–9474.
- [19] Auromita Mitra and Indranil Dutta. 2023. Mixed language processing increases cross-language phonetic transfer in Bengali–English bilinguals. *Bilingualism: Language and Cognition* 26, 5 (2023), 896–909.
- [20] Massrura Mostafa and Marium Jamila. 2012. From English to Banglish: Loanwords as opportunities and barriers?: Is English swamping the Bangla language? English Today 28, 2 (2012), 26–31.
- [21] Michael Muller, Anna Kantosalo, Mary Lou Maher, Charles Patrick Martin, and Greg Walsh. 2024. GenAICHI 2024: Generative AI and HCI at CHI 2024. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 1–7.
- [22] Luminița Nicolescu and Monica Teodora Tudorache. 2022. Human-computer interaction in customer service: the experience with AI chatbots—a systematic literature review. *Electronics* 11, 10 (2022), 1579.
- [23] Beijing Academy of Artificial Intelligence. 2025. BAAI/bge-reranker-v2-m3. Retrieved January 20, 2025 from https://huggingface.co/BAAI/bge-reranker-v2-m3.
- [24] OpenAI. 2025. Embeddings. Retrieved January 20, 2025 from https://platform.openai.com/docs/models#embeddings.
- [25] OpenAl. 2025. GPT-4o. Retrieved January 20, 2025 from https://platform.openai.com/docs/models#gpt-4o.
- [26] OpenAI. 2025. Pricing. Retrieved January 18, 2025 from https://openai.com/api/pricing/.
- [27] Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simeoni. 2023. Collaborating with a text-based chatbot: An exploration of realworld collaboration strategies enacted during human-chatbot interactions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1-17
- [28] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927 (2024). Retrieved from https://arxiv.org/abs/2402.07927.
  [29] Katie Seaborn and Yeongdae Kim. 2023. "I'm" Lost in Translation: Pronoun
- [29] Katie Seaborn and Yeongdae Kim. 2023. "I'm" Lost in Translation: Pronoun Missteps in Crowdsourced Data Sets. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. 1–6.
- [30] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-Centric Survey and Taxonomy of Human-Generative-AI Interactions. arXiv preprint arXiv:2310.07127 (2023). Retrieved from https://arxiv.org/abs/2310.07127.
- [31] Alex Singla, Alexander Sukharevsky, Lareina Yee, and Michael Chui. 2024. The state of ai in early 2024: Gen ai adoption spikes and starts to generate value. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
- [32] Fouad Trad and Ali Chehab. 2024. Prompt engineering or fine-tuning? a case study on phishing detection with large language models. Machine Learning and Knowledge Extraction 6, 1 (2024), 367–384.
- [33] Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes. In Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track). 236–249.
- [34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35 (2022), 24824–24837.
- [35] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 214–229.
- [36] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in codeswitching?. In Proceedings of the Fifth Workshop on Computational Approaches to

- Linguistic Code-Switching. 142-153.
- [37] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In Proceedings of the 1st Workshop on Multilingual Representation Learning. 1–15.
- [38] Jie Zeng and Jianbu Yang. 2024. English language hegemony: retrospect and prospect. Humanities and Social Sciences Communications 11, 1 (2024), 1–9.
- [39] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 12567–12582.
- [40] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019). Retrieved from https://arxiv.org/abs/1909.08593.

#### A EXAMPLE QUERY AND ANSWERS

Banglish Query: Pariya movie ta dekhte dekhte akta time er por load hocche na apnader server e jokhon ato problem tale server thik na kore backend service ta na dekhe cinema gulo streaming e deben na ok?

**Generated answer:** Pariya movie ta dekhte dekhte load hocche na, eta khub oshubhidar. Apnar device e kono somossa hote pare. Apni niche deya troubleshooting step gulo follow kore dekhte paren:

- (1) Apnar phone er settings e giye app cache clear korun:
  - Home screen theke Settings e jaan.
  - Apps ba Applications select korun.
  - MultilingualCo select korun.
  - Storage e giye Clear Data ba Clear Storage select korun, tarpor OK korun.
- (2) Ekti different internet connection try korun:
  - Onno kono Wi-Fi network e connect korun.
  - Apnar device ke bondhu ba porshider barite niye giye stream korun
  - Jodi cellular data network use korte paren, ta hole seta diye MultilingualCo connect kore dekhen.

Jodi ei step gulo follow korar poro somossa thake, tahole amra apnake ekjon human operator er sathe jogajog korte sahajjo korte pari.

MultilingualCo's answer: Hello, Greetings of the day. Thank you for contacting us. Really sorry for the trouble caused. Please let us know which device you are using so that we can assist you as soon as possible. Please mention the device brand and model too. If you are trying to play from a phone, please mention it's Android version too. Also, if you are trying to play from your laptop/pc, share us which browser you are using and also please share us which browser and which browser version you are using. Awaiting your response. Thanks and Regards, Team MultilingualCo

# B EXAMPLE SENTENCE MISSCLASSIFICATION BASED ON BANGLISH EMBEDDINGS

Figure 2 shows the 3D principal component analysis representation of a known multilingual embedding model processing two sentences, each written in English, Bengali, and Banglish (represented by the Bengali sentence written in Roman script [15]). As shown in the figure, while the model correctly clusters the Bengali and English sentences, it is not able to correctly map the Banglish sentences. This shows the flaw that many multilingual (embedding)

models suffer from, namely the underperformance for linguistic cases they were not trained on [1].

#### C BANGLASSIST PROMPT

```
prompt = f"""

<>>[INST] You are a professional and empathetic customer service assistant for MultilingualCo, a leading Bengali streaming service provider. Your primary goal is to deliver accurate, clear, and helpful responses to customer questions while maintaining a conversational and respectful tone. Use the provided context to answer the question.\n\n

# Guidelines:\n

- Always respond in the same script and language as the user question, regardless of the language or script used in the provided context.\n

- Analyze the script of the user question (e.g., Bengali, Roman, or others) and ensure your answer matches it exactly.\n

- Examples:\n

- User Question: "Eta dekhache. Ki korbo ekhon?" (Roman script)\n

- Context: "বিসালিক বিধিনিয়েম্বর কারণে ক্ষেত্রকটি নিষিষ্ট ছান থেকে অলু কিছু সামগ্রী দেখা

| সীমাবন্ধ করা ক্ষেত্রভা (Bengali script)\n

- Answer: "Apin je shamogriti khunzhchen, ta apnar deshe

| upolobdho noy." (Roman script)\n

2. Act as if the provided context is your own inherent knowledge. Do not reveal or imply that the information comes from external inputs.

3. If the provided context does not allow you to answer confidently, be transparent about it. Instead of suggesting email communication, offer to directly connect the customer with a human operator for further assistance.

4. **Conversation Flow**:\n

- If this is the first interaction (no chat history), warmly greet the customer and address them by name if available.\n

- If this is the first interaction, skip the greeting and respond directly to the query.\n\n

# Inputs:\n

- User Question: {question}\n

- Context: {reranked_context} [/INST] </s>
```

Figure 3: Prompt used to generate BanglAssist's replies

#### D BANGLASSIST INTERFACE SCREENSHOTS

Figure 4 shows three screenshots of BanglAssist's interface implementation in Streamlit: (a) the home screen of the chatbot, showing three FAQs the user can choose from; (b) a chatbot answer to an FAQ, printed from the FAQ database; (c) a chatbot answer generated through *GPT-40* based on the user question and the context retrieved from the FAQ database.

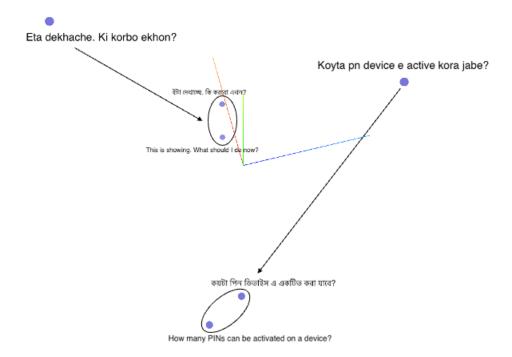


Figure 2: 3D principal component analysis representation of two sentences, embedded in 3 different linguistic variations

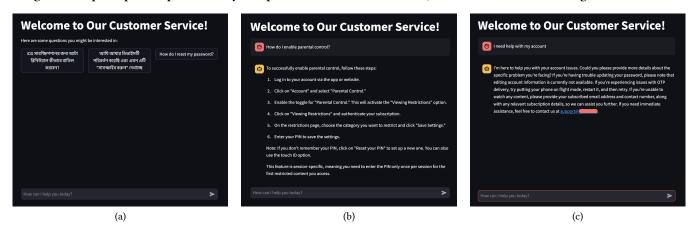


Figure 4: Screenshots of the chatbot implementation in Streamlit: (a) Home screen of the chatbot, showing three FAQs; (b) Answer printed from the FAQ database; (c) Answer generated through *GPT-40* based on the user question and the context retrieved from the FAQ database

# E BANGLASSIST EVALUATION QUERIES BREAKDOWN

Table 3: Language and topic of each query used to evaluate BanglAssist

Query	Language	Topic
1	Bengali	Unable to watch specific episode
2	Bengali	Discount not applied to total amount
3	Bengali	Request to add specific content
4	Bengali	Poor video quality while streaming
5	Bengali	Account idle despite subscription
6	Banglish	Content streamed on illegal websites
7	Bengali	Black screen while playing video
8	English	Issue downloading specific content
9	Banglish	Too many devices connected to account
10	English	Cancel subscription via Google Play
11	Banglish	Problem receiving OTP to login
12	Banglish	Auto-pay showing wrong renewal date
13	English	Problem receiving OTP on new phone
14	Banglish	Wrong period showing for offer
15	English	Dissatisfaction with released content
16	Banglish	Problem accessing platform
17	Banglish	Issue while streaming content
18	Banglish	Issue while streaming specific content
19	English	No video while streaming content
20	Banglish	Content leaked via Telegram

### F EQUATIONS

In the following, |Q| is the total number of queries, MRR indicates the Mean Reciprocal Rank, and  $Rank_i$  is the position of the first relevant document for query i.

Precision@k = 
$$\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\text{Number of relevant docs in top } k}{k}$$
 (1)

Recall@k = 
$$\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\text{Number of relevant docs in top } k}{\text{Total number of relevant docs}}$$
 (2)

MRR@k = 
$$\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{Rank}_i}$$
, where  $\text{Rank}_i \le k$  (3)

# G BANGLASSIST EVALUATION BREAKDOWN BY LANGUAGE

Table 4: Average retrieval scores per language before reranking (k = 5 and k = 3) and after reranking (k = 3)

Language	Precision@k	Recall@k	MRR@k
k = 5: Retrie ( <b>Before rer</b> e	val performance e <b>anking)</b>	valuation	
Bengali	0.67	0.50	1.00
English	0.64	0.39	0.80
Banglish	0.47	0.40	0.81
	king performance		
	king performance l after reranking		
(Before and Bengali	d after reranking 0.89	g) 0.43	1.00
(Before and Bengali English	l after rerankin	0.43 0.30	<b>1.00</b> 0.80
(Before and Bengali	d after reranking 0.89	g) 0.43	
(Before and Bengali English	d after reranking 0.89 0.73	0.43 0.30	0.80
Bengali English Banglish	0.89 0.73 0.67	0.43 0.30 0.35	0.80 0.81