Enhancing Non-English Conversational Agents Using Synthetic Data Generation

Salar Hashemitaheri
University of California, Irvine
Irvine, California, USA
salarh@uci.edu

Ian Harris
University of California, Irvine
Irvine, California, USA
harris@ics.uci.edu

Abstract

Online support groups for smoking cessation provide an accessible, cost-effective method for assisting individuals looking to quit. However, these groups often face challenges related to low user engagement and the stigma associated with smoking cessation. Non-native English speakers encounter additional barriers in accessing effective support, as language differences increase feelings of isolation and discourage participation. To address these issues, this study focuses on developing a multilingual conversational agent designed to offer real-time, culturally sensitive support in multiple languages, including Mandarin, Arabic, Hindi, Spanish, and Farsi. A critical aspect of the system is intent classification, enabling the agent to understand user inputs and generate contextually appropriate responses. Due to the limited availability of high-quality training data in non-English languages, we introduce a synthetic data augmentation strategy that combines both synthetic data generation and real data scraping. Initially, we fine-tune a pre-trained large language model on multilingual smoking cessation group data. To enhance intent detection accuracy, we augment the training dataset by generating synthetic posts using advanced prompt engineering. We compare the performance of the fine-tuned LLM with its non-fine-tuned counterpart, demonstrating that synthetic data augmentation results in a 35.5% increase in F1 scores. This improvement validates the effectiveness of our augmentation approach in boosting the accuracy of non-English conversational agents. By integrating this conversational agent into widely used social media platforms, such as GroupMe or X, we aim to offer a scalable, inclusive solution that reduces stigma, enhances engagement, and improves the effectiveness of smoking cessation programs for non-native English speakers.

Keywords

Natural Language Processing, Large Language Model, Conversational agents, Generative AI, Data Augmentation, Intent Classification, Multilingual support, Health Care, Smoking Cessation.

ACM Reference Format:
Salar Hashemitaheri and Ian Harris. 2025. Enhancing Non-English Conversational Agents Using Synthetic Data Generation. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3721201.3725444

ACM ISBN 979-8-4007-1539-6/2025/06
https://doi.org/10.1145/3721201.3725444

1 Introduction

Smoking cessation remains a critical public health challenge worldwide, with tobacco use being a leading cause of preventable death and disease. Non-native English speakers face substantial obstacles in accessing appropriate cessation support, often due to language barriers and lack of culturally sensitive resources. These challenges are compounded by the stigma surrounding smoking, which can discourage individuals from seeking help and participating in cessation programs[1].

Online support groups for smoking cessation have emerged as an accessible and cost-effective alternative to traditional in-person programs. However, for non-native English speakers, these digital platforms often face issues of low user engagement and persistent stigma due to limited availability of culturally relevant and linguistically appropriate resources[2].

A study by Wu et al. emphasizes the significance of cultural sensitivity in smoking cessation programs, showing that culturally tailored interventions for Chinese-speaking smokers were significantly more effective than standard programs[3]. This highlights the need for both linguistic and cultural adaptation in smoking cessation programs.

A promising approach to improving engagement and accessibility is the integration of natural language processing (NLP) and large language models (LLMs) into online support systems. These technologies offer the ability to develop conversational agents, or chatbots, that can interact with users in a natural, empathetic manner and provide immediate, personalized responses. Such systems are particularly useful for ensuring timely interactions within online support groups, where delays in responses can discourage participation[4, 5].

Intent classification—the process of identifying the user's intent behind their message—is a critical component of any conversational agent. For non-native English speakers, this task becomes even more challenging due to diverse linguistic structures and idiomatic expressions across languages. Many existing NLP models are predominantly trained on English-language data, which limits their ability to effectively process non-English input and deliver accurate responses[6].

To overcome these challenges, this study employs a novel data augmentation strategy that combines synthetic data generation and real data scraping. By generating synthetic data using advanced prompt engineering with large language models and scraping real posts from global online smoking cessation communities, we aim to create a robust training dataset capable of improving intent detection in multilingual contexts.

This research addresses these barriers by developing a multilingual conversational agent that provides culturally sensitive and linguistically appropriate support to non-native English speakers in smoking cessation programs. Our approach includes fine-tuning a pretrained LLM to classify user posts in multiple languages, thereby improving the model's ability to detect intent across diverse linguistic and cultural contexts.

By focusing on non-native English speakers, our goal is to reduce stigma, increase engagement, and improve the overall efficacy of smoking cessation interventions for linguistically diverse populations. Through the use of advanced NLP techniques and multilingual support, we believe this study can make significant strides in making smoking cessation resources more inclusive, effective, and accessible for all individuals, regardless of their language or cultural background[4].

2 Literature Review

Multilingual conversational agents in healthcare and smoking cessation are emerging as promising tools to overcome language barriers and provide personalized support. This review explores recent studies on their development and effectiveness. In healthcare, multilingual chatbots have shown potential in improving accessibility.

Rainey et al. demonstrated that a multilingual chatbot effectively engaged arthroplasty patients with limited English proficiency. This study highlights the importance of language-specific support in healthcare settings[2].

For smoking cessation, conversational agents have proven effective in providing continuous support and reducing stigma. Whittaker et al. conducted a scoping review of chatbots for smoking cessation, identifying increased engagement and improved health outcomes as significant benefits. However, they noted that language barriers could limit effectiveness for non-English speaking populations[7].

Bricker et al. developed QuitBot, a conversational agent for smoking cessation, which showed promising results with a 63% 30-day quit rate compared to 38.5% in the control group[4]. While effective for English speakers, the challenge of scaling to non-native speakers remains.

Wu et al. emphasized the importance of cultural sensitivity in smoking cessation programs, demonstrating that culturally tailored interventions for Chinese-speaking smokers were significantly more effective than standard programs[3].

Another prominent contribution to the field is MedLingua, a multilingual conversational agent designed to provide primary healthcare education and information to chronic patients [8]. MedLingua utilizes NLP to interpret user inputs and offer relevant medical advice in a variety of languages. This approach highlights the growing importance of multilingual systems in making healthcare more inclusive, ensuring that language does not become a barrier to receiving essential medical support.

Challenges in developing effective multilingual conversational agents include accurate intent classification and ensuring cultural sensitivity. Researchers have employed data augmentation techniques and GPT-based synthetic data generation to improve language understanding and user interaction in non-native English contexts.

Future research should focus on fine-tuning language models for specific languages and dialects, employing sophisticated data augmentation methods, and integrating these agents into popular social media platforms such as GroupMe or X to increase reach and engagement.

3 Proposed Method

In this study, we aim to improve the performance of a non-English conversational agent for smoking cessation by augmenting the training dataset with synthetic posts in five non-English languages: Spanish, Chinese, Farsi, Arabic, and Hindi. Unlike previous work, which focused on English posts, this paper addresses the unique challenges of multilingual intent detection, emphasizing the need for cultural sensitivity, linguistic nuances, and language-specific adjustments in the synthetic data generation process.

The dataset used in this study is derived from a randomized controlled trial involving cigarette smokers participating in two online quit-smoking groups. Over 100 days, participants' tweets were collected, categorized into responses to daily auto messages and spontaneous tweets. The study focused on evaluating Twitter-based support groups for smoking cessation, with the dataset providing insights into participant engagement, auto message effectiveness, and the correlation between tweeting behavior and quit attempts. The dataset excludes individuals with contraindications for nicotine replacement therapy, those not intending to quit, and those on depression medication or using illicit drugs. It includes a variety of intents, such as mental health, physical health, and NRT-related problems, offering valuable data for assessing digital interventions and improving intent detection capabilities[9, 10].

Our labeled dataset is derived from 82,000 posts obtained from smoking cessation support groups, which were manually annotated with 24 predefined intents.[9, 10]. These posts were initially in English and were translated into Spanish, Chinese, Farsi, Arabic, and Hindi using GPT-4 to ensure multilingual coverage[11]. As depicted in Figure 1, these **Machine Translated Posts** serve as the foundation for generating synthetic posts, especially for those intents with low F1-scores in the initial model. After translating the original dataset into non-English languages, we asked a native speaker of each language to validate and edit the posts to ensure they were accurate and **Human Verified Posts**.

Given that there is no established base model for non-English conversational agents in the context of smoking cessation, we use the performance of the initial model as the baseline for comparison. The goal of this study is to improve precision, recall, and F1-scores by leveraging synthetic data augmentation techniques tailored to each language's unique characteristics. By comparing the results of the best multilingual models in two steps—first with non-fine-tuned LLaMA and XLM-RoBERTa, and then after fine-tuning these models—we demonstrate how synthetic data augmentation can significantly improve intent detection accuracy for non-English languages [12, 13].

To achieve this, we must ensure that we are using a promising method to generate high-quality synthetic posts that will yield the desired accuracy. Therefore, as depicted in Fig. 1, the quality of the original test data will first be checked in the Quality Screening step. Then, based on the intent definition and the quality of the original posts, we will Craft Prompts for each language. Using these prompts, we will Generate Synthetic posts for each language.

Figure 1: Generating Non-english Synthetic Posts

Finally, we will use the synthetically generated posts to fine-tune the models and improve their accuracy.

The **Quality Screening** process ensures that only high-quality posts are selected for synthetic data generation. For non-English languages, it is essential to consider linguistic and cultural differences, as each language has its own structure, idioms, and contextual relevance.

- **Relevance:** In Spanish, the concept of smoking cessation might include region-specific methods, such as different perceptions of NRT or the use of herbal remedies. In China, traditional methods may be more culturally relevant than pharmaceutical solutions. Therefore, posts must be evaluated for their alignment with culturally specific cessation practices.
- **Contextual Completeness:** The idea of completeness differs across languages. For example, in Farsi, responses often tend to be indirect, focusing on family support and encouragement, whereas in Arabic culture, posts may emphasize community and collective efforts to quit smoking. This must be accounted for when evaluating posts for their completeness, ensuring that cultural nuances are respected.
- **Clarity:** For languages such as Chinese and Arabic, clarity is critical due to the script differences and syntactic structure. In Chinese, for instance, the omission of particles and context-specific expressions may make posts difficult to understand, so these factors must be considered during the screening phase.

The **Crafting Prompts** step involves transforming the original posts into prompts that will guide the GPT model in generating diverse and relevant synthetic posts. In this phase, the key challenge is ensuring that the prompts reflect the linguistic structure and cultural expectations of each language.

For this process, we used GPT-4 to create the prompts for all languages, ensuring that the structure of the prompts remained consistent across languages. The prompts were carefully designed to reflect the core smoking cessation challenges, such as managing cravings, withdrawal symptoms, and coping strategies. Despite the linguistic and cultural differences, the prompts maintained the same intent across languages to generate contextually appropriate responses.

Once the prompts are crafted, the next step is to generate responses using GPT-4. The challenge here is not just generating syntactically correct responses but ensuring they are culturally appropriate and engage users effectively in the target languages.

- **Spanish:** The response generation process in Spanish should ensure that both formal and informal tones are maintained based on the user's context, such as age or familiarity with the bot. Additionally, regional variations in vocabulary (e.g., Spain vs. Mexico) should be considered to prevent awkward phrasing or misunderstandings.
- **Chinese:** In Mandarin, the tone of the responses should be adjusted to reflect the high-context communication style, which emphasizes respect and indirectness. The responses should also use polite forms of address and culturally appropriate strategies for coping with cravings.
- **Farsi:** In Farsi, responses need to be sensitive to the cultural value placed on family and respect for authority. Responses should encourage group involvement and use gentle, supportive language rather than direct commands or instructions.
- **Arabic:** The Arabic responses should include phrases that highlight collective support and solidarity. The responses should also respect the cultural value of modesty and community respect, particularly in the context of sensitive issues like smoking cessation.
- **Hindi:** In Hindi, responses should be designed to be both empathetic and supportive, using language that resonates with the user's socio-economic background. Code-switching between Hindi and English should be allowed where appropriate, reflecting the bilingual nature of many Hindi speakers.

Considering all these points, we used separate prompts for each language to Generate Synthetic Posts. Intent-related prompts were employed to create new synthetic posts for each label, helping to diversify the intents. This process continued until we observed any semantic drift or saturation. We ensured the creation of high-quality synthetic posts by having a native speaker validate that all generated posts were suitable for the fine-tuning step. To maintain a stronger focus on intent detection accuracy, we excluded intents that were not relevant for classification, such as greetings, non-responses, and others. As a result, we fine-tuned the dataset with 22 intents instead of 24. After generating the synthetic posts for each intent, we proceeded to fine-tune the model with the newly added data. Fine-tuning is a critical step in improving the conversational agent's intent detection accuracy, particularly for non-English languages, where models may struggle to generalize across diverse linguistic structures and cultural contexts.

Fine-tuning involves adjusting the pretrained model on the newly augmented dataset, which now includes both the original posts and the synthetic posts in multiple languages. The process is as follows:

(1) **Model Adaptation:** The model, initially trained on the English dataset, must be adapted to handle multiple languages effectively. This is done by fine-tuning the model on the non-English dataset, ensuring that it learns the intricacies of each language's syntax, semantics, and cultural context.
(2) **Language-Specific Adjustments:** During fine-tuning, it is essential to account for language-specific features. For instance, in Spanish and Hindi, verb conjugations and gender agreement must be handled carefully to avoid errors. In Chinese, word segmentation must be addressed since Chinese is a non-segmented language. For Arabic and Farsi, special

attention is needed for the right-to-left script and the need for diacritics in some cases.

(3) **Cross-Lingual Training:** Fine-tuning for cross-lingual transfer involves teaching the model to understand and apply the learned representations of intent across different languages. By training the model on multilingual data, the model can generalize better across languages, even when there are limited data points for some intents. For example, Spanish and Hindi datasets had over 10,000 samples each, while Arabic and Farsi datasets had around 7,000, providing the model with adequate data coverage.

(4) **Intent Alignment:** Ensuring that the model correctly aligns the intents across languages is a crucial aspect of fine-tuning. Each language must be evaluated for how well the intents in Spanish, Chinese, Farsi, Arabic, and Hindi match the original English intents. Fine-tuning adjusts the model to detect and classify similar intents in all languages. This process has resulted in a 35.5% improvement in F1 scores for the fine-tuned model, demonstrating significant gains in intent detection across languages.

(5) **Hyperparameter Tuning:** Hyperparameters such as learning rate, batch size, and training epochs were adjusted specifically for the multilingual dataset. Since different languages may have varying levels of difficulty and training data size, tuning these hyperparameters ensures the model optimizes its performance for each language. The fine-tuning process involved adjusting the learning rate to 5e-5 and training for 3 epochs, with a batch size of 16, to avoid overfitting and ensure stable convergence across languages.

(6) **Evaluation Metrics:** During and after fine-tuning, we evaluated the model using precision, recall, and F1-score for each language. These metrics provided valuable insights into how well the model detected and classified intents in the target languages. After fine-tuning, we observed an overall improvement in the F1 score of 35.5%, demonstrating the effectiveness of our multilingual data augmentation approach in enhancing intent classification across diverse languages.

By fine-tuning the model on the augmented, multilingual dataset, we expect to improve the model's overall performance in detecting smoking cessation intents across all five languages. This fine-tuning process ensures that the conversational agent is more capable of handling diverse user inputs, providing accurate and culturally appropriate responses.

4 Results

4.1 Metrics Used for Evaluation

In this paper, we evaluate the performance of our non-English conversational agent using three key metrics commonly used in NLP tasks: Precision, Recall, and F1-score. These metrics are essential for assessing the ability of the model to correctly classify and generate responses in the context of smoking cessation support.

- **Precision:** Measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It is important for determining the accuracy of the model's predictions.
- **Recall:** Measures the proportion of correctly predicted positive instances out of all actual positive instances. It is crucial for understanding how well the model detects the correct intent.
- **F1-score:** The harmonic mean of Precision and Recall, providing a balanced measure of the model's accuracy. It is particularly useful when there is an imbalance between the number of predicted positive and negative instances.

These metrics are computed for each intent label, which corresponds to a specific user query or action, and averaged to evaluate the overall model performance. In this work, due to space constraints, we only compare average F1 scores, which incorporate both precision and recall.

4.2 Models Used for Evaluation

For the evaluation, we compared several state-of-the-art LLMs: LLaMA 2, GPT-4, and XLM-RoBERTa. These models were chosen for their performance in multilingual settings that make them strong candidates in terms of overall performance due to their robust architecture and diverse data.

- **LLaMA 2:** LLaMA 2 is an open-source model that has demonstrated superior performance across various NLP tasks. In this study, we fine-tuned the LLaMA 2 model using our multilingual dataset to improve intent classification for non-English languages.
- **GPT-4:** GPT-4 is a highly powerful model developed by OpenAI, known for its ability to generate human-like responses. However, despite its strong language understanding capabilities, it is not fine-tunable in this application, which limits its accuracy in specific tasks compared to fine-tunable models.
- **XLM-RoBERTa:** XLM-RoBERTa is a robust multilingual model based on RoBERTa, designed specifically for cross-lingual understanding. It is particularly useful for languages with limited resources and has shown competitive results in multilingual NLP tasks.

4.3 Evaluation and Comparison of Models

In this study, we computed Precision, Recall, and F1 scores for each of the 22 intents in our multilingual dataset. The average F1 score, representing the overall model accuracy, was calculated for each model. The results show how well each model can classify and generate responses across the six languages: English, Spanish, Chinese, Farsi, Arabic, and Hindi. The F1 score, balancing Precision and Recall, provides insight into the model's ability to handle different intents while minimizing errors.

As there is no baseline or standardized dataset for multilingual smoking cessation, we used human-verified machine-translated posts as the baseline dataset. These translated posts were compared with the results of our models, allowing us to measure the improvement of each model with and without fine-tuning, as well as with synthetic data augmentation. After crafting prompts and generating synthetic posts discussed in the proposed method, the number of prompts and the total number of synthetic posts generated for each language are as follows:

- **English:** 314 prompts, generating 1,037 synthetic posts.
- **Spanish:** 364 prompts, generating 1,112 synthetic posts.

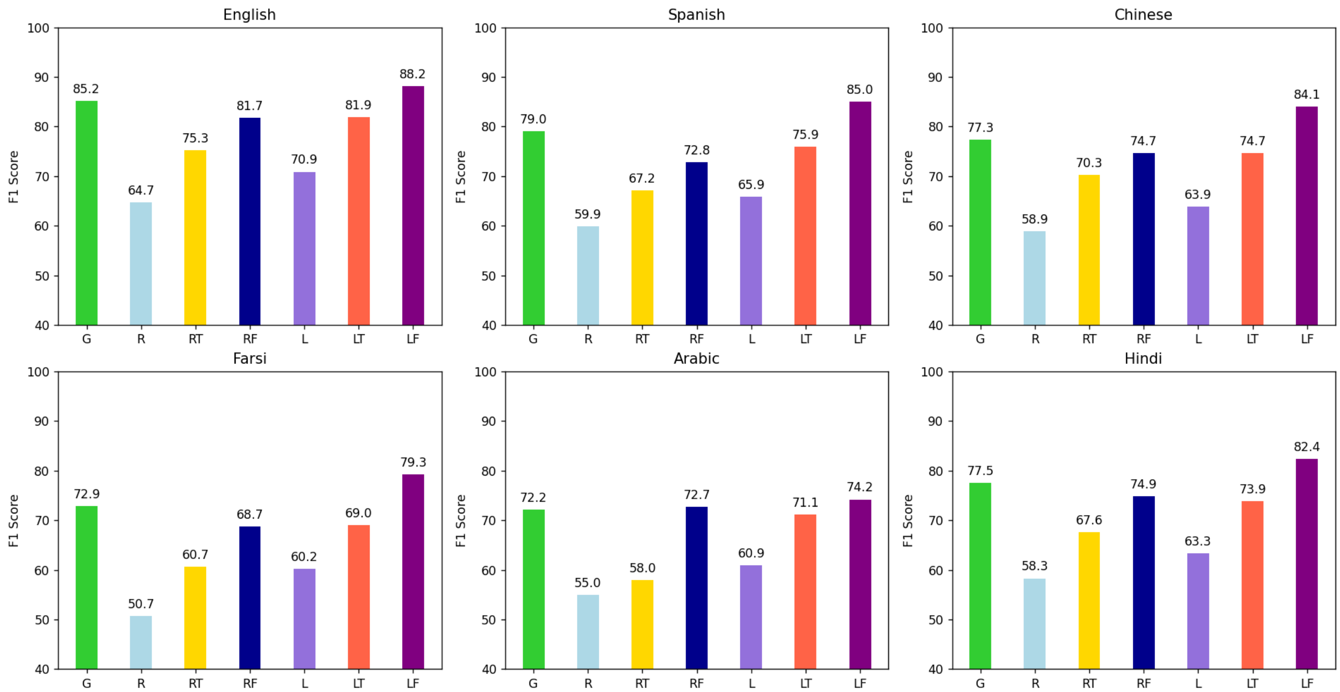

Figure 2: Comparison of Non-Fine-Tuned Translated Augmented Data vs. Fine-Tuned Translated Augmented Data vs. Fine-Tuned Synthetic Augmented Data Average F1 Scores by Language

- **Chinese:** 347 prompts, generating 7,045 synthetic posts.
- **Farsi:** 289 prompts, generating 7,376 synthetic posts.
- **Arabic:** 382 prompts, generating 5,432 synthetic posts.
- **Hindi:** 437 prompts, generating 5,806 synthetic posts.

These synthetic posts, derived from diverse prompts, played a critical role in augmenting the dataset for the fine-tuning process.

As seen in Figure 2, GPT (G) outperforms XLM-RoBERTa (R) and LLaMA2 (L) in most languages when using non-fine-tuned translated data. For example, in Farsi, GPT achieves an F1 score of 72.9, while XLM-RoBERTa and LLaMA score 50.7 and 60.2, respectively. This pattern is consistent across all languages, showing that GPT performs better with non-fine-tuned translated data. The reason behind GPT's higher performance can be attributed to its powerful ability to generate coherent and natural responses, giving it an edge over models like XLM-RoBERTa and LLaMA, which require fine-tuning for better performance.

To enhance the performance of XLM-RoBERTa and LLaMA, we fine-tuned both models using translated data and synthetic data. Fine-tuning allows these models to learn specific language patterns and adapt to the context of smoking cessation, thereby improving their ability to detect user intents.

- **Fine-Tuning with Translated Data:** Both XLM-RoBERTa and LLaMA show notable improvements in F1 scores after fine-tuning with translated data. For example, in Chinese (Mandarin), XLM-RoBERTa improves from 58.9 (R) to 70.3 (RT), and LLaMA improves from 63.9 (L) to 74.7 (LT). This confirms the value of fine-tuning for adapting the models to the specific language nuances and intents related to smoking cessation.
- **Fine-Tuning with both Synthetic and Translated Data:** After fine-tuning with synthetic data, we observe further performance improvements. The synthetic data fine-tuned models (LF and RF) outperform the translated data fine-tuned models (LT and RT) across all languages, especially in Spanish, Chinese, and Arabic. For example, in Spanish, the fine-tuned LLaMA model with synthetic data (LF) scores 85.0, a significant improvement from the 75.9 and 79 achieved by the fine-tuned LLaMA with translated data (LT) and GPT (G). This demonstrates the effectiveness of synthetic data augmentation in improving the overall performance of multilingual conversational agents.

4.4 Impact of Synthetic Data Augmentation

The Figure 2 clearly illustrates the effectiveness of synthetic data augmentation in enhancing the F1 scores of multilingual models. Synthetic data, when used in combination with fine-tuning, provides a substantial boost in the overall performance of the models. It helps the models learn from a broader range of language patterns and contextual nuances, improving their ability to detect and respond accurately to user intents.

In English, Spanish, and Chinese, the fine-tuned synthetic models (LF and RF) outperform both the non-fine-tuned (L and R) and fine-tuned translated models (LT and RT).

The synthetic data fine-tuned models demonstrate higher F1 scores, highlighting that synthetic data can fill the gaps in training data and improve model accuracy.

The increase in F1 scores from the non-fine-tuned models to the Translated fine-tuned models, and from fine-tuned translated data to fine-tuned translated and synthetic data, proves that synthetic data augmentation can significantly improve the effectiveness of multilingual conversational agents. This is particularly important in domains like smoking cessation, where linguistic and cultural differences play a major role in the accuracy of conversational agents. Synthetic data enables better understanding and generation of responses that are culturally sensitive and linguistically accurate.

5 Conclusion

In this study, we proposed an innovative approach to enhance the performance of a non-English conversational agent for smoking cessation by leveraging synthetic data generation and fine-tuning across five non-English languages: Spanish, Chinese, Farsi, Arabic, and Hindi. By carefully adapting each step of the process—quality screening, prompt crafting, response generation, and quality assurance—to address the unique linguistic and cultural characteristics of these languages, we were able to generate high-quality synthetic data that effectively augmented our training dataset.

The results of our approach demonstrate significant improvements in the accuracy of the conversational agent, with a 35.5% increase in F1 scores compared to the baseline model. This improvement underscores the effectiveness of our synthetic data augmentation strategy, particularly in enhancing intent detection accuracy for underperforming intents in the target languages. Fine-tuning the model with both the original and synthetic multilingual data successfully enhanced the model's ability to understand and classify intents across diverse linguistic contexts, making it more effective in providing culturally sensitive and linguistically appropriate support for smoking cessation.

Our findings highlight the potential of synthetic data generation as a powerful tool to overcome data scarcity challenges and improve the performance of multilingual conversational agents in health-related applications. This approach represents a significant step forward in the development of more inclusive and effective digital health interventions for non-native English speakers. It also provides a replicable framework for improving the accuracy of conversational agents in other multilingual domains, opening the door for more effective digital support systems in a variety of fields.

Acknowledgments

We would like to thank the student researchers Leon Xiang Lian, Abubakr Elmallah, Kimberly Judith Rodriguez, and Rushil Dileep who helped us validate both the translated and synthetic data, enabling us to reach the desired level of performance.

Lastly, we would like to acknowledge the use of ChatGPT for proofreading this paper.

References

[1] Lozano, P., et al. 2020. Smoking-Related Stigma: A Public Health Tool or a Damaging Force? *Nicotine & Tobacco Research* 22, 1 (Jan. 2020), 96–103.

[2] Rainey, J. P., et al. 2023. A Multilingual Chatbot Can Effectively Engage Arthroplasty Patients Who Have Limited English Proficiency. *The Journal of Arthroplasty* 38, 7 Suppl 2 (Jul. 2023), S78–S83.

[3] Wu, D., et al. 2009. The effect of a culturally tailored smoking cessation program for Chinese American smokers. *Nicotine & Tobacco Research* 11, 12 (Dec. 2009), 1448–1457.

[4] Bricker, J. B., et al. (2024). Conversational chatbot for cigarette smoking cessation: Results from the 11-step user-centered design development process and randomized controlled trial. *JMIR mHealth and uHealth*, 12, e57318.

[5] Hashemi, S, et al. (2022). EARL: An Efficient Approximate HaRdware Framework for AcceLerating Fault Tree Analysis. *25th Euromicro Conference on Digital System Design (DSD)*, Maspalomas, Spain, pp. 32-38.

[6] Hashemi, S., et al. 2019. The effect of a culturally tailored smoking cessation program for Chinese American smokers. *A Fast and Efficient Fault Tree Analysis Using Approximate Computing," 2019 15th European Dependable Computing Conference (EDCC)* Naples, Italy, 2019, pp. 39-46.

[7] Whittaker, R., et al. 2022. Chatbots for smoking cessation: Scoping review. *Journal of Medical Internet Research* 24, 9, e35556.

[8] Himel, G. M. S., et al. 2024. MedLingua: A conceptual framework for a multilingual medical conversational agent. *MethodsX* 12 (Feb. 2024), 102614.

[9] Pechmann, C., Pan, L., Delucchi, K., Lakon, C. M., & Prochaska, J. J. (2015). *Development of a Twitter-based Intervention for Smoking Cessation that Encourages High-Quality Social Media Interactions via Automessages*. Journal of Medical Internet Research, 17(2), e50.

[10] Giyahchi, T., et al. (2022). Customized training of pretrained language models to detect post intents in online support groups. In *Multimodal AI in Healthcare* (Studies in Computational Intelligence, vol. 1060, pp. 59-75). Springer.

[11] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[12] Touvron, H., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

[13] Conneau, A., et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451).