Enhancing User Message Intent Detection in a Mobile Health Smoking Cessation Intervention: Fine-tuning a Large Language Model with Data Downsampling and Error Correction

Shagoto Rahman, Cornelia (Connie) Pechmann, Ian G Harris

Submitted to: Journal of Medical Internet Research
on: September 02, 2025

Table of Contents

Enhancing User Message Intent Detection in a Mobile Health Smoking Cessation Intervention: Fine-tuning a Large Language Model with Data Downsampling and Error Correction

Shagoto Rahman[1] MS; Cornelia (Connie) Pechmann[2] PhD; Ian G Harris[1] PhD

[1]Department of Computer Science University of California, Irvine Irvine US
[2] The Paul Merage School of Business, SB1-4317 University of California, Irvine Irvine US

Corresponding Author:
Shagoto Rahman MS
Department of Computer Science
University of California, Irvine
University of California, Irvine
Irvine
US

Abstract

Background: Although smoking cessation aids including support groups and nicotine replacement therapy (NRT) can help people quit smoking, quit rates remain low. Emerging mobile health interventions like online support groups can help overcome barriers related to aid accessibility and convenience. Combined with NRT, online support groups hold significant potential but demand continuous, labor-intensive efforts to deliver timely responses that maintain participant engagement. Accurate user intent detection, which is the process of understanding the purpose behind a user's message, can play a critical role by identifying individual needs and consequently providing timely and proper responses. Recent large language model advancements in natural language processing and artificial intelligence (AI) have shown promise. However, these systems often struggle when faced with a large number of intent categories or the complex nature of human language. Uneven data across intent categories—some rare and others dominant—makes it harder for the system to correctly recognize user intent and respond.

Objective: The main goal of this study was to develop an AI tool, especially a large language model that could accurately recognize users' message intents, despite imbalances and complexities in data. In our application, users' message intents were related to a smoking-cessation support-group intervention and utilization of the free nicotine replacement therapy (NRT) provided as part of that intervention.

Methods: We consistently used a state-of-the-art public domain large language model, Llama-3 8B from Meta. First, we used the model off-the-shelf. Second, we fine-tuned it on our annotated domain dataset of 25 intent categories. Third, we downsampled the predominant intent category to reduce bias and fine-tuned the model. Finally, we combined downsampling with corrected annotations to create a cleaned dataset for another round of fine-tuning. This stepwise approach progressively improved classification accuracy by addressing prior limitations.

Results: Without fine-tuning, the large language model achieved an unweighted-average F1-score of 0.29 and a weighted-average F1-score of 0.37, where unweighted treated all categories equally, while weighted emphasized larger ones. Fine-tuning alone achieved unweighted and weighted-average F1-scores of 0.72 and 0.86 respectively. Downsampling plus fine-tuning achieved unweighted and weighted-average F1-scores of 0.80 and 0.85 respectively. Downsampling, fine-tuning and human error correction achieved an unweighted-average F1-score of 86% and a weighted-average F1-score of 90%.

Conclusions: On smoking cessation data, large language models performed poorly without fine-tuning, underscoring the need for domain-specific training. However, with domain-specific fine-tuning, performance has suffered because of the highly imbalanced dataset. Downsampling the majority category before fine-tuning improved results moderately, however left room for further enhancement and raised concerns about potential noise in the dataset. Carefully reviewing the misclassified samples has helped identify annotation inconsistencies, and after correcting these errors and fine-tuning the model on the corrected dataset, best performance has been achieved.

(JMIR Preprints 02/09/2025:83437)

DOI: https://doi.org/10.2196/preprints.83437

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Shagoto Rahman, M.S.
PhD Student
Department of Computer Science,
University of California, Irvine
Irvine, CA, United States
Email: shagotor@uci.edu

Cornelia (Connie) Pechmann, PhD
Professor
The Paul Merage School of Business, SB1-4317
University of California, Irvine
Irvine, CA, United States
Email: cpechman@uci.edu

Ian G Harris, PhD
Professor
Department of Computer Science,
University of California, Irvine
Irvine, CA, United States
Email: iharris@uci.edu

Enhancing User Message Intent Detection in a Mobile Health Smoking Cessation Intervention: Fine-tuning a Large Language Model with Data Downsampling and Error Correction

Abstract

Background: Although smoking cessation aids including support groups and nicotine replacement therapy (NRT) can help people quit smoking, quit rates remain low. Emerging mobile health interventions like online support groups can help overcome barriers related to aid accessibility and convenience. Combined with NRT, online support groups hold significant potential but demand continuous, labor-intensive efforts to deliver timely responses that maintain participant engagement. Accurate user intent detection, which is the process of understanding the purpose behind a user's message, can play a critical role by identifying individual needs and consequently providing timely and proper responses. Recent large language model advancements in natural language processing and artificial intelligence (AI) have shown promise. However, these systems often struggle when faced with a large number of intent categories or the complex nature of human language. Uneven data across intent categories—some rare and others dominant—makes it harder for the system to correctly recognize user intent and respond.

Objective: The main goal of this study was to develop an AI tool, especially a large language model that could accurately recognize users' message intents, despite imbalances and complexities in data. In our application, users' message intents were related to a smoking-cessation support-group intervention and utilization of the free nicotine replacement therapy (NRT) provided as part of that intervention.

Methods: We consistently used a state-of-the-art public domain large language model, Llama-3 8B from Meta. First, we used the model off-the-shelf. Second, we fine-tuned it on our annotated domain dataset of 25 intent categories. Third, we downsampled the predominant intent category to reduce bias and fine-tuned the model. Finally, we combined downsampling with corrected annotations to create a cleaned dataset for another round of fine-tuning. This stepwise approach progressively improved classification accuracy by addressing prior limitations.

Results: Without fine-tuning, the large language model achieved an unweighted-average F1-score of 0.29 and a weighted-average F1-score of 0.37, where unweighted treated all categories equally, while weighted emphasized larger ones. Fine-tuning alone achieved unweighted and weighted-average F1-scores of 0.72 and 0.86 respectively. Downsampling plus fine-tuning achieved unweighted and weighted-average F1-scores of 0.80 and 0.85 respectively. Downsampling, fine-tuning and human error correction achieved an unweighted-average F1-score of 86% and a weighted-average F1-score of 90%.

Conclusions: On smoking cessation data, large language models performed poorly without fine-tuning, underscoring the need for domain-specific training. However, with domain-specific fine-tuning, performance has suffered because of the highly imbalanced dataset. Downsampling the majority category before fine-tuning improved results moderately, however left room for further enhancement and raised concerns about potential noise in the dataset. Carefully reviewing the misclassified samples has helped identify annotation inconsistencies, and after correcting these errors and fine-tuning the model on the corrected dataset, best performance has been achieved.

Introduction

Although most smokers wish to stop smoking, cessation rates remain low in part because of the insufficient utilization of the cessation aids [1-3]. The use of mobile health interventions to aid people in quitting smoking has become commonplace [4,5] including text messaging programs [6,7], and quit-smoking apps [4,8]. As traditional in-person support groups significantly improve quit rates [9,10], online groups have emerged to help overcome utilization barriers related to accessibility and convenience [11]. Online smoking-cessation support groups offer considerable promise; however, they require ongoing, resource-intensive efforts to provide timely, 24/7 responses that sustain participant engagement. Research has consistently demonstrated that active involvement in online health support groups leads to improved outcomes [12,13] and studies also show that low engagement is often linked to dropouts [14].

Chatbots can enhance user engagement and minimize dropout rates by providing continuous, cost-effective support. A leading determinant of the effectiveness of chatbot systems is accurate intent detection, defined as the capability to precisely and consistently identify and classify each user's underlying goal or purpose based on their input message. Accurate intent detection can improve user engagement by providing the inputs needed to the system to provide personalized, accurate, and timely support. It can also minimize the need for human intervention. With accurate intent detection, a chatbot can engage users by actively responding to messages with relevant content, addressing

their concerns, and delivering evidence-based health information and support. For example, if a user indicates they have failed to quit smoking, the chatbot can identify this particular intent and suggest ways to overcome the challenge.

Previous chatbot studies have primarily focused on one-to-one conversations between chatbot and user, e.g., for mental health or substance abuse screening [15] or interventions [16]. However, we study conversations in online groups, often involving several group members, which have unique properties compared to one-to-one conversations including multi-party dialogue, interruptions, sub-group discussions, and incomplete messaging leading to exceptionally unstructured and noisy data [17]. In addition, research has been done to detect users' message intents in medical settings, e.g., online forums or office visits, but most involved relatively few intent categories which limited their ability to fully capture users' diverse needs [18,19].

Moreover, data imbalance is particularly common in the health domain because sufficient data is rarely available for all intent categories yet ones with fewer samples can be critically important [20,21]. For example, an intent category such as "overdose" may have very few messages, but it represents a critical concern. Most prior studies did not properly ensure that the samples or exemplars were sufficiently balanced across intent categories, possibly leading to suboptimal model performance. Data downsampling, which involves reducing the number of samples from highly dominant majority categories to achieve a more balanced category distribution, can play a vital role. This approach can help the classifying model learn to detect all intent categories comparably, including those that are rare but critically important.

Large language models can deliver state-of-the-art performance in recognizing message intents, because they are trained on vast amounts of data. As a result, these models can grasp the deeper meaning of text and often replicate human-like understanding. However, for highly specific classification tasks, large language models must be fine-tuned on domain-specific datasets where humans have annotated the message intents for this fine-tuning phase. Since data annotation is primarily done by humans, it often contains errors due to message ambiguity, human misunderstanding or carelessness or bias, or other factors. Manually reviewing the entire dataset to find human annotation errors is costly and time-consuming. Instead, focusing only on the data points where the model and human annotators disagree on the user's message intent can save time and effort while effectively identifying annotation issues.

In this study, we demonstrate common challenges and solutions using large language models for mobile health interventions. We implement a state-of-the-art, public domain, large language model called Llama-3 8B from Meta to detect the intent in user messages in an online smoking-cessation intervention. Our study method consists of four parts. Part one involves intent detection using the Llama-3 8B large language model off-the-shelf, without fine-tuning. Part two fine-tunes the model on a domain-specific dataset which human annotators annotated with what they believed to be the intent of each user message. Part three combines fine-tuning with downsampling of the majority (predominant) intent category which was found to bias the model and weaken its performance at intent recognition. Part four adds correction of human error in intent annotation, adding this to fine-tuning and downsampling. The rest of this paper is organized as follows: We elaborate on prior work, discuss our methods, report our experimental results, summarize the principal findings and limitations, and state future research directions and a conclusion.

Prior Work

Conversational Agents in Healthcare

Conversational healthcare agents have become important resources for providing quick and easily accessible medical assistance to patients. Recent research suggests that by harnessing advancements in natural language processing, they can provide personalized patient support and alleviate the excessive workload of many healthcare professionals. Babu et al. created a BERT-based medical conversational agent to provide responses to people seeing medical advice or appointments or information about specific medical conditions [22]. The BERT (Bidirectional Encoder Representations from Transformers) model was fine-tuned on public databases like MIMIC-III, Pubmed, and a Covid-19 dataset totaling 11,000 samples of medical questions. The conversational agent integrated these extracted datasets with the ongoing conversation context and produced relevant and personalized feedback. When compared to baseline models such as LSTM (Long Short-Term Memory), SVM (Support Vector Machine), and Bi-LSTM (Bi-directional Long Short-Term Memory), the chatbot demonstrated superior performance, attaining 98% accuracy along with high precision, recall, and F1-scores combining precision and accuracy. However, the system demanded high computation power and suffered from data bias and interpretability issues.

Yu et al. developed a conversational agent by combining a fine-tuned DialoGPT model and ChatGPT to assist mental health conversations [23]. The authors sought to address current model limitations in delivering proper and safe therapeutic responses due their lack of training on specialized data. So, authors fine-tuned the DialoGPT model on a domain specific dataset focused on Cognitive Behavioral Therapy (CBT) and integrated it with ChatGPT's dynamic conversational capabilities. This hybrid conversational agent system was compared to conventional rule-based models and large language models on metrics such as BLEU (Bilingual Evaluation Understudy) which assesses the match between the model output and human-written text. Inputs from mental health professionals and patients were also considered. The results demonstrated that the hybrid system outperformed others in terms of conversational quality, relevance, and emotional resonance.

Zamani et al. sought to address the limitations of rule-based chatbots as mental health conversational agents including lack of response creativity and low user acceptability [24]. They also sought to address the risks of large language models in terms of generating harmful texts, and the limitations of data imbalance. The authors evaluated the impact of novel text augmentation techniques to improve the message intent detection. Their study demonstrated that increasing the training dataset through augmentation could boost model accuracy by contributing to dataset diversity. The study incorporated Easy Data Augmentation (EDA), synonym replacement utilizing WordNet, embedding-based replacement, and context-aware substitutions. Model performance improved markedly through data augmentation, with up to a 26.4% increment in the average intent detection score.

Large Language Models for Intent Detection in Healthcare

Grasping the speaker's message intent is essential for facilitating smooth and meaningful human interaction. With recent advancements in natural language processing and deep learning, various recent studies have demonstrated success in detecting the subtle intents behind spoken expressions. Aftab et al. developed an intent classification method for healthcare conversational agents by incorporating Bayesian Long Short-Term Memory Networks (LSTMs) [25]. They used an open-source dataset containing 6,661 text utterances corresponding to 25 unique medical symptom-related intents with a relatively balanced distribution of intent categories. Their method attained an impressive 99.4% accuracy in recognizing users' message intent on the test dataset. Moreover, the

authors used Shannon entropy from several random model runs to measure uncertainty, which helped the system distinguish between familiar and unfamiliar inputs. Also to estimate model uncertainty, they used Monte Carlo dropout during inference. As a result, the chatbot handled uncertain predictions more appropriately, allowing it to generate safer, less harmful responses.

Zhang et al. developed Conco-ERNIE, an innovative model designed to enhance user intent detection in complex medical queries by permitting single intent, multi-intent, and implicit intent recognition [26]. The model used a special feature extractor called ERNIE with an added layer to help it understand and label important parts of medical questions. The authors incorporated the Apriori algorithm to find patterns in how medical concepts related to each other, allowing it to better detect intents when users' uttered meanings were unclear. Moreover, they introduced an attention-based method that combined the meaning of words in a query with the patterns of related medical concepts. This helped the model focus better on important information and ignore irrelevant details. The authors tested their method on a dataset of real medical questions and it showed clear improvements in detecting multiple and hidden user intents. Overall, the model achieved a score of 89.4% in test data, much higher than other popular methods. The authors then added their model Conco-ERNIE to a real healthcare chatbot linked to a large medical knowledge base to assist real users.

Sakurai et al. studied large language models for intent detection in conversations involving donation requests [27]. The study examined conversations where the responder expressed criticism of donations or sought to make donors feel guilty. The authors primarily studied GPT-4 in this context and found it often misidentified critical, sarcastic, or negative affective responses as motivational or neutral. GPT-4 was tested on both real and artificially created conversations and its predictions about intent were compared to human annotated intents. Humans were able to detect critical and other negative messages, but GPT-4 often misunderstood them as encouraging or motivational. Overall, this study found that a state-of-the-art large language model struggled to detect subtle intents in negative speech.

In this paper, we demonstrate that a current state-of-the-art large language model cannot adequately detect subtle intents expressed by participants in a smoking cessation intervention, e.g., the model struggles to decipher substantively different user messages about the intervention-provided nicotine replacement therapy (NRT). We then illustrate the specific steps that can be taken, using existing analytical methods, to improve the model's intent detection. We show vast improvements by fine-tuning the off-the-shelf model using a domain-specific dataset with human annotations of 25 message intents. We show additional substantial improvements by downsampling (reducing excess examples of) an overly dominant intent category and correcting human annotation errors identified by examining model-human disagreements in intent classification. In sum, we show that large language models can perform very poorly in domain-specific contexts like ours but can be massively improved with current analytical approaches.

Methods

Overview

Figure 1 presents our method for training a large language model for user message intent detection. It involved four stages: (1) Large language model without fine-tuning, where the model predicted intents without any domain-specific training dataset solely using its own knowledge; (2) Large language model with fine-tuning, where the model was fine-tuned on an intent-annotated dataset to improve performance; (3) Large language model with fine-tuning and downsampling, which

addressed category imbalance by reducing the sample size of the majority (dominant) category before fine-tuning; and (4) Large language model with fine-tuning, downsampling, and human annotation error correction, where model-human disagreements were reviewed for potential human annotation errors, corrected if necessary, and used to retrain the model. This stepwise approach allowed for progressive enhancement in classification accuracy, with each stage addressing specific limitations of the previous one.

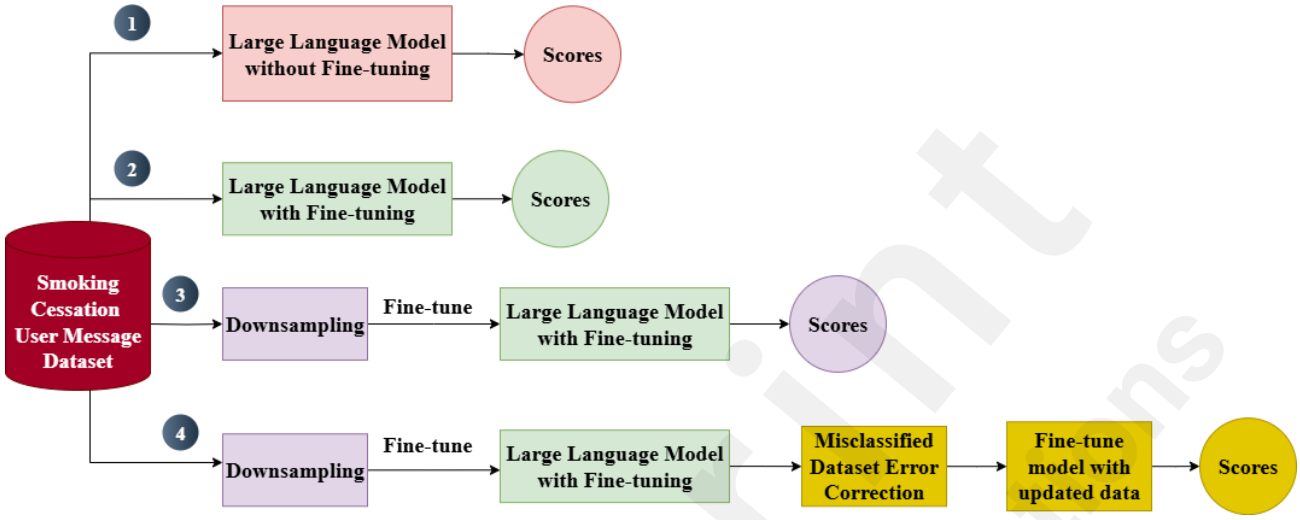


Figure 1. Overview of Methodology

Dataset

For this study, we utilized the smoking-cessation user messages from the Tweet2Quit dataset [17]. Tweet2Quit was a social media-based online intervention to assist people with quitting smoking. Participants were assigned to private online support groups of 20 members and they engaged in peer support discussions about smoking cessation. The dataset contained over 82,000 messages gathered from 45 support groups participating in two clinical trials conducted between 2012-2014 and 2016-2019. Each support group remained active for three months and generated approximately 1,822 messages on average. The messages were created by individuals actively working to quit smoking and then to stay quit, and as a result, contained domain rich, real-world interactions regarding quitting and the free nicotine replacement therapy or NRT (nicotine patches, gum and lozenges) provided as a quitting aid.

A team of 25 trained research assistants annotated the intent behind each message, yielding a comprehensive and detailed annotated corpus. A total of 25 intent categories were identified while annotating the messages. The topics were clinically related (e.g., nrt_howtouse, nrt_overdose, nrt_mouthirritation), reflected social support for quitting (e.g., greetings, support), or off-topic (e.g., small talk, weather, news). Table 1 presents the 25 message intent categories identified in the dataset, along with their definitions and representative examples. Each message was independently annotated by two research assistants, with any disagreements resolved by a third adjudicator with the greatest domain expertise. The annotation demonstrated high reliability, with a Cohens' Kappa score of 0.93 overall, and agreement between specific annotators ranging from 87.3% to 97.2% across intent categories. The "off-topic" intent category was used to indicate irrelevant user messages unrelated to

smoking cessation or support group bonding. It alone accounted for 56% of the dataset which indicated significant category imbalance.

Table 1. Message Intent Category Definitions and Examples

Intent Category	Definition	Examples
cigsmell	Complains about cigarette smell	1. Cigarette smoke & cigarette smokers stink to me now. 2. I can smell people smoking a mile away now. It's gross to me now.
cravings	Discusses craving/wanting/ temptation/urge/trigger to smoke or how to fight/avoid craving	1. It's hard because all I want to do is reach over and get a cigarette and smoke. 2. my urge is in the evening watching TV I usually occupy my mind with crochet that's all I really have
ecigs	States something about ecigs, vape, or liquid or juice	1. Is anyone going to try vaping to help quit? The non-nicotine juice of course. I've been debating. 2. I do like the Vape pen
fail	Admits to smoking a cigarette	1. I have to admit I slipped today and smoked a cigarette 2. mornings I'm a chain smoker
greetings	Says a greeting, e.g., hi	1. Good afternoon! How's everyone doing today? 2. Good morning
health	States something about relation between smoking and health	1. My breathing is much better and I don't never want to smoke a cigarette again. 2. To be able to breathe better
off-topic	Discusses other topic e.g., small talk, weather, news	1. Now i gotta go check mine!!! 2. I skipped breakfast and went right for lunch
nrt_dontwork	States NRT doesn’t work, don’t like, no longer needed	1. Since the gum and patches don't seem to work on my boring days I have struggled 2. This gum is a no go for me. I called the smoke shop in my town.
nrt_dreams	States they either sleep with the patch or try not to do so; often discusses dreams	1. I'm sleeping with patch on. Have the craziest most vivid dreams. 2. Yes! Very lucid dreams. I had to stop sleeping with them on.
nrt_howtouse	Asks question or gives info about how to use NRT in general eg dose	1. How often should I use the lozenges? 2. So your not suppose to constantly chew the gum right? Suppose to just let it lay in your gums once broken up right?
nrt_itworks	States reason why NRT used, e.g., it works, stops craving	1. I'm using 1 patch a day and chewing the gum as needed! Both help immensely with cravings! 2. I have stepped down to the 7mg patch!!!! Yea!!!!!
nrt_mouthirritation	States NRT gum/lozenge causes bad taste, throat	1. I toss the gum when it starts burning my mouth or throat.

	gum irritation or burning or spicy sensation	2. don't like the gum I'm having a hard time chewing it. Burns my mouth a little
nrt_nauseous	States NRT makes them nauseous or gag or sickly	1. Don't like the gum makes me nauseous. 2. using the #14 patch. Not using gum it makes me sick.
nrt_od	States NRT is too strong, feels like overdose	1. I am going to move to the step 2 patch. The 21mg patches are making me sweat and feel bad 2. I had to move from the 21mg early cause it was just too much nicotine!
nrt_skinirritation	States NRT patch causes skin irritation	1. Anybody else's skin being irritated by the patches? 2. I get itchy from mine too.
nrt_stickissue	States NRT patch won't stick	1. I know that with sweat the patches don't stay on I have problems with mine everyday 2. My patches won't stay on
quitdate	States something about quit date	1. Hey my quitDate is today 2. My last pack of smokes. Friday is the day!
savingmoney	States money they spend per pack or how much they save by quitting	1. I'm putting away the money I save into an account and then can do something fun! 2. all the money spent to damage health
scared	States being scared/nervous/ anxious/ fearful about quitting or sometimes NRT	1. I am really nervous about quitting but ready 2. I've not tried the gum yet. I'm scared of it lol.
smokefree	States success in being smoke free	1. I'm still smokefree starting week 2. YAY 2. I made it thru without smoking. Starting Day 2. Yeah.
smokingless	Has not yet quit only able to cut down on cigs smoked OR starting to quit but still smokes some	1. Can’t say I've completely quit. But 4 or 5 a day is less than 15 2. Proud moment: only 1 cigarette today. Never thought I'd get this far.
stress	States feeling stress	1. I think stress has been tough but I'm not giving up 2. Last week work was really stressful had a lot going on at once
support	Supports other participants	1. Good for you for not giving up. You got this!!! 2. you're doing great ! wooohooo!!!
tiredness	States feeling tired due to unspecified reasons or quitting smoking, not NRT	1. I have been so tired lately. I'm wondering if it is connected to quitting smoking. 2. So very tired...
weightgain	States gaining weight or eating more or anything about weight or exercise	1. Ever since I quit I eat more so I'm going to start eating healthier and exercise 2. I have gained 7 pounds!

The dataset was divided into three mutually exclusive and collective exhaustive subsets: training, evaluation, and test. The test dataset consisted of user messages from groups 30-36. From the remaining groups out of the total 45, 75% of the messages were used for training and 25% for evaluation. The training dataset was used to fine-tune the large language model, allowing it to learn patterns and features from the annotated examples. During training, the model iterated over the dataset multiple times (epochs), improving its performance by minimizing errors. The evaluation (validation) dataset was used to assess the performance of different versions of the large language model based on its training. Since it was unseen by the model during training, it served as an unbiased measure to select the best-performing model versions. The test dataset was reserved for reporting the final classification results. Also unseen by the model, it was used to evaluate the generalizability and robustness of the trained model, providing an estimate of how well the model would perform when deployed in practical settings. Table 2 summarizes the distribution of annotated user messages across the three datasets.

Table 2.Distributions in the Datasets of Categories of Users' Message Intents

Intent Category	**Train Dataset**	**Evaluation Dataset**	**Test Dataset**
cigsmell	328	109	76
cravings	1446	482	267
ecigs	205	68	73
fail	774	258	152
greetings	1828	610	490
health	847	283	265
off-topic	26093	8698	9664
nrt_dontwork	307	102	72
nrt_dreams	258	86	54
nrt_howtouse	354	118	105
nrt_itworks	615	205	135
nrt_mouthirritation	140	46	21
nrt_nauseous	122	41	23
nrt_od	81	27	21
nrt_skinirritation	155	51	35
nrt_stickissue	188	63	83
quitdate	1279	427	318
savingmoney	455	152	90
scared	293	98	83
smokefree	3072	1024	818
smokingless	285	95	33
stress	508	169	148
support	6330	2110	1983
tiredness	153	51	43
weightgain	531	177	167
Total	**46647**	**15550**	**15219**

Large Language Model without Fine-tuning

The first stage of our methodology was classifying the users' message intents using a large language model without any fine-tuning. In other words, the model was tasked with classifying text into predefined domain-specific message intent categories without any exposure to messages annotated with these intent categories in a training dataset. As a common example, a large language model might be prompted to classify the intent of a user message stating, "'I am very sad" as either positive or negative in sentiment, and it might correctly identify the intent as "negative", despite not having been trained on the specific dataset. Since large language models are trained on vast and diverse text corpora, they have acquired an understanding of syntax, semantics, grammar, and sentiment, making them well-suited for general-purpose language tasks such as intent detection, sentiment analysis, and topic summarization.

When a large language model is used for a classification task without fine-tuning, it is guided through prompt to emulate the task, for example, in the classification task example above, a suitable prompt might be "Classify the sentiment of the following text into one of the two categories {positive, negative}: I am very sad" which is presented to the large language model. Then the model uses its implicit learning and reasoning to perform the task. The model works by predicting the most probable next word in the sequence. For instance, in this example, the language model may evaluate the prompt by considering potential sentiment categories such as "positive" and "negative". Drawing on its prior knowledge and contextual understanding, it may assign a higher probability to "negative" based on the input text and subsequently predict "negative" as the final classification. The main advantage of this approach is that it does not require annotation of users' message intents. The only requirement is to provide the test dataset to the large language model, which then performs the classification based solely on its prior knowledge. This approach saves time and computation resources.

However, the concern is that the domain-specific task may involve text that is not part of the dataset used for pretraining the model. For example, in the domain of smoking cessation, the intent behind the user messages can often be highly domain-specific as health-intervention communications tend to be private. A user message such as "I used gum today and wanted to smoke two hours later", might be interpreted by a language model as indicating a cessation failure. However, within the context of our smoking-cessation dataset, this user message corresponds to a more specific intent, "nrt_dontwork", which signifies that the user perceives the nicotine replacement therapy (NRT) as ineffective.

Thus, knowledge of the domain becomes crucial for getting higher accuracy with large language models. When a model lacks domain-specific knowledge, its performance on related tasks tends to decline. Moreover, in the absence of domain-specific knowledge, large language models may introduce biases or errors stemming from the datasets on which they were originally trained, possibly causing hallucinations. Furthermore, large language models without domain knowledge tend to struggle with classification tasks involving numerous similar intent categories, such as our categories of "nrt_dontwork," and "nrt_itworks", which refer to NRT being ineffective or effective respectively. While these intents have completely different meanings, large language models that lack fine-tuning often cannot distinguish between them.

In this study, we assessed the performance of a large language model without fine-tuning at completing the task we required for our mobile-health application: Recognizing the 25 main intents of users' messages related to smoking cessation which included numerous nuanced messages about using the free NRT. We utilized the state-of-the-art but off-the-shelf Llama-3 8B by Meta. We supplied this model with descriptions of all 25 of our intent categories, as defined in Table 1, to facilitate its classification process and ensure that its predictions were constrained to these predefined

intent categories.

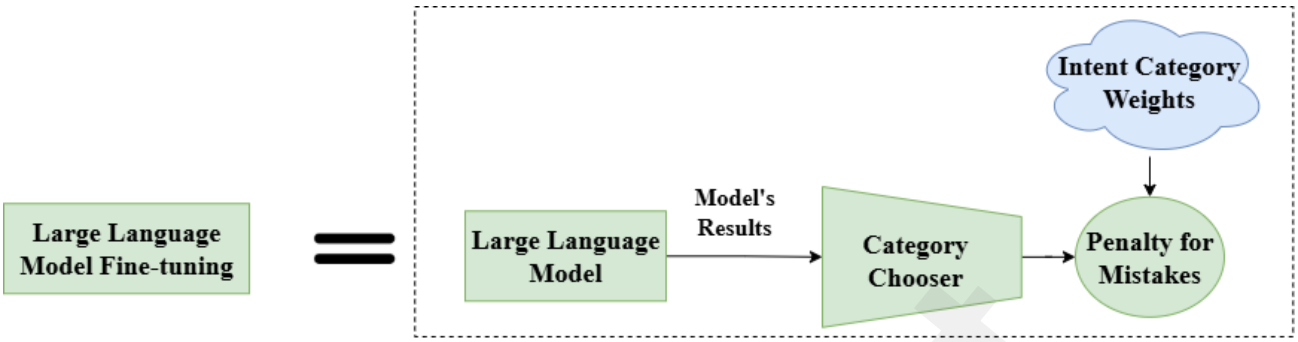


Figure 2. Overview of Large Language Model Fine-tuning

Large Language Model with Fine-tuning

The next stage of our methodology was fine-tuning the same large language model, Llama-3 8B, using
our domain-specific dataset with user messages about smoking cessation. Fine-tuning a large language model refers to further training the model on one or more datasets from the targeted domain, the goal being to transform a general-purpose model into one specialized for the domain for improved performance. Domain specific knowledge is often crucial for large language models because the datasets used in the initial basic training often lack the detailed context relevant to the specific task. Fine-tuning enables the models to learn domain-relevant nuances, semantics, and reasoning.

For example, during fine-tuning, the model might be exposed to user messages about NRT gum aiding or not aiding with then smoking cessation, for instance, “I used gum today and wanted to smoke two hours later.” Fine-tuning allows the model to learn nuanced distinctions and accurately differentiate between semantically similar user message intents. For instance, a model without fine tuning might incorrectly predict the message intent here as “fail” (failure at smoking cessation). However, with fine-tuning on the domain-specific dataset, the model should learn that the user is talking about their NRT gum failing to help them with nicotine cravings, and therefore the correct intent is “nrt_dontwork.”

On the other hand, fully training a large language model on a domain-specific dataset risks the model forgetting its previously acquired general knowledge. Consequently, the model might forget critical knowledge related to language comprehension, reasoning, and generalizing across tasks. Therefore, instead of completely retraining the model, we undertake fine-tuning because it preserves and often freezes the original model weights (i.e., the models’ prior knowledge) while learning new weights (new knowledge) that is built upon the existing foundation. So, this approach enables the model to retain its previously acquired knowledge while simultaneously incorporating new insights from the domain-specific dataset.

Therefore, we fine-tune the Llama-3 8B large language model on our smoking-cessation dataset. Meta’s Llama-3 8B when used in causal mode is designed to predict the next word in a sequence sequentially. When used in this mode, intent detection requires specialized prompting, for example, “Input: I am feeling very sleepy. So the intent of this user message is:”. This prompt is sent to the

model which then predicts the next probable word as “tiredness”. This prompt can be formulated in various ways; however, the approach has several limitations. It is highly dependent on the prompt design and it operates indirectly through generation, which reduces efficiency, makes end-to-end training more challenging, and necessitates careful prompt engineering and output parsing. Often, the model’s output may fall outside the predefined intent categories; for instance, in the previous example where the main intent is “tiredness”, the model might generate a related but nonfocal intent such as “drained.”

For classifying users’ intents from their messages, we use a version of the Llama-3 model that is specifically adapted for classification rather than using the standard causal mode designed for language generation. Instead of predicting the next word in a sentence, this version of the model learns to assign entire messages to intent categories. To achieve this, we added a lightweight classification layer which works as the category chooser producing the final model results. This layer enables the model to directly select the most appropriate category for each input message. Importantly, this approach allows us to begin to address the issue of dataset imbalance, where some intent categories are more common than others. By assigning more weight to underrepresented categories during fine-tuning, the model becomes more attentive to less frequent but clinically important message types.

Additionally, this approach gives us a probability score for each category, which helps us understand how confident the model is in its predictions. This is valuable for downstream decision-making and human oversight. Figure 2 provides an overview of this fine-tuning process, showing that the model’s outputs feed into the category chooser, which then predicts the most appropriate intent category. During fine-tuning, we assign importance to underrepresented categories when the model makes mistakes. This is done by adding category weights to the penalty for mistakes, helping the model pay more attention to rare but clinically important intent categories. We fine-tune the Llama-3 8B model in this way to improve its ability to detect nuanced user intents in smoking-cessation conversations, e.g., related to utilization of the free nicotine replacement therapy (NRT) provided as a cessation aid.

Large Language Model with Fine-tuning and Downsampling

The third part of our method was downsampling the most frequent message intent category and then fine-tuning the model on the updated dataset. Dataset imbalance is a typical problem in classification tasks, meaning that the number of instances (samples or examples) often varies quite substantially across the different focal categories. For example, in our dataset, the “off-topic” intent category contained 44,445 samples, whereas the “nrt_od” category contained only 129 samples. In a multi-category classification task, if a particular category has a much higher number of samples than the other categories, the model being trained tends to be biased toward choosing the category with the higher number of samples. In other words, in the training dataset the model sees many more samples of this majority category, and ends up favoring it in its predictions. Thus, the minority categories are underpredicted.

For our application which is to accurately predict the intent of user messages in smoking-cessation support groups, both majority and minority intent categories require accurate predictions. But, if we train the model with all our data, the model will tend to be more accurate in predicting the highly dominant “off-topic” category and less accurate in predicting the minority categories with fewer samples. Additionally, the “off-topic” intent category refers to messages considered irrelevant or uninformative for smoking cessation or support group bonding; it falls outside the 24 key intent categories. Knowing this gives us even more leeway to address our severe category imbalance by

reducing the volume of “off-topic” category examples. Doing this will not compromise the data from the other 24 categories.

So for our analysis, we have downsampled the “off-topic” intent category. Downsampling is a method employed to mitigate category imbalance within a dataset by decreasing the sample size of the majority category(ies) to better align with sample sizes of the minority categories. We have downsampled the instances of the “off-topic” category to mirror the average number of instances across the other 24 categories. This approach has been consistently applied to the training, test, and evaluation datasets. After that, we have fine-tuned the model using the same approach as in the Large Language Model with Fine-tuning Section.

Large Language Model with Fine-tuning, Downsampling, and Error Correction

The final step in our methodology was downsampling the dominant intent category, fine-tuning the model, correcting human annotation errors flagged by examining model-human disagreements in the intent classifications, and fine-tuning the model again. Since the datasets were manually annotated, they were susceptible to human annotation errors. Though large language models are powerful tools due to their training on extensive datasets, fine-tuning them on poorly annotated or noisy data can negatively impact their learning and classification performance.

Model-human disagreement on a message’s intent category can reveal a systematic weakness in either the model or the human annotations. The model could have incorrectly classified the user’s intent, the human annotators could have done so, or possibly both. Therefore, identifying model-human disagreements over how to classify users’ message intents and resolving the human-caused errors enables us to improve the final dataset we use for model fine-tuning. If the model and humans disagree, reviewing these disagreements helps us identify human annotation mistakes, for example, due to user messages being ambiguous or vague, or general human oversights in annotations. The human annotation errors can then be corrected in the dataset, and fine-tuning can proceed with less noisy data.

Thus, we analyzed the model-human disagreements over users’ message intents in our datasets as a debugging aid. A domain expert thoroughly analyzed each model-human disagreement, and corrected the human annotation if needed. After this, we used the cleaned-up training dataset to fine-tune the large language model yet again.

Results

Evaluation Metrics

To evaluate the different large language model training approaches for recognizing users’ message intents, we used three standard accuracy measures:

Precision: How often the model was correct when it predicted a certain intent.

$$Precision = \frac{Correct\ predictions\ for\ an\ intent}{Total\ predictions\ made\ for\ that\ intent}$$

Recall : How often the model successfully recognized an intent when a user message conveyed it.

$$Recall = \frac{Correct\ predictions\ for\ an\ intent}{Total\ actual\ cases\ of\ that\ intent}$$

F1-score: This balances both precision and recall, combining them into a single number.

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Aggregated Performance

Figure 3 illustrates the performance of the state-of-the-art public-domain large language model, Llama-3 8B from Meta, after training using four different methods. The off-the-shelf model with extensively pretraining but no domain-specific fine-tuning performed poorly at recognizing users' message intents in that domain. Out of all test cases, 991 predictions could not be matched to any of the 25 intent categories (for example, the model said "stick" instead of stickissues, or just "nrt" instead of full category). Since these could not be clearly assigned to a category, we excluded them from the final scoring. Performance metrics were therefore calculated only on predictions that directly matched one of the 25 categories. It achieved an unweighted-average F1-score of 0.29 and a weighted-average F1-score of 0.37. The unweighted-average F1-score assigned equal importance to each intent category, regardless of its size. In contrast, the weighted-average F1-score gave greater weights to categories with a larger number of instances. The lower scores observed in both metrics for the model without fine-tuning highlighted the complexity of the classification task. It demonstrated that intent classification in the mobile-health context studied was not straightforward and required specialized fine-tuning to achieve adequate performance.

The large language model with domain-specific fine-tuning achieved a much improved unweighted-average F1-score of 0.72 and a weighted-average F1-score of 0.86. This improvement demonstrated the value of fine-tuning in enhancing model performance for this domain-specific task. However, the scores also indicated there was room for improvement.

The large language model that was fine-tuned after downsampling of the dominant category improved the unweighted-average F1-score from 0.72 to 0.80 while maintaining a weighted-average F1-score of 0.85. These results demonstrate the effectiveness of downsampling to address category imbalance and allow the model to more accurately recognize the minority categories. The unweighted-average score, which considered each intent category as equally important, improved substantially, consistent with the goal of ensuring all intents were accurately predicted. Lastly, the model with fine-tuning, downsampling and error correction performed the best, bolstering the unweighted-average F1-score from 0.80 to 0.86 and the weighted-average score from 0.85 to 0.90.

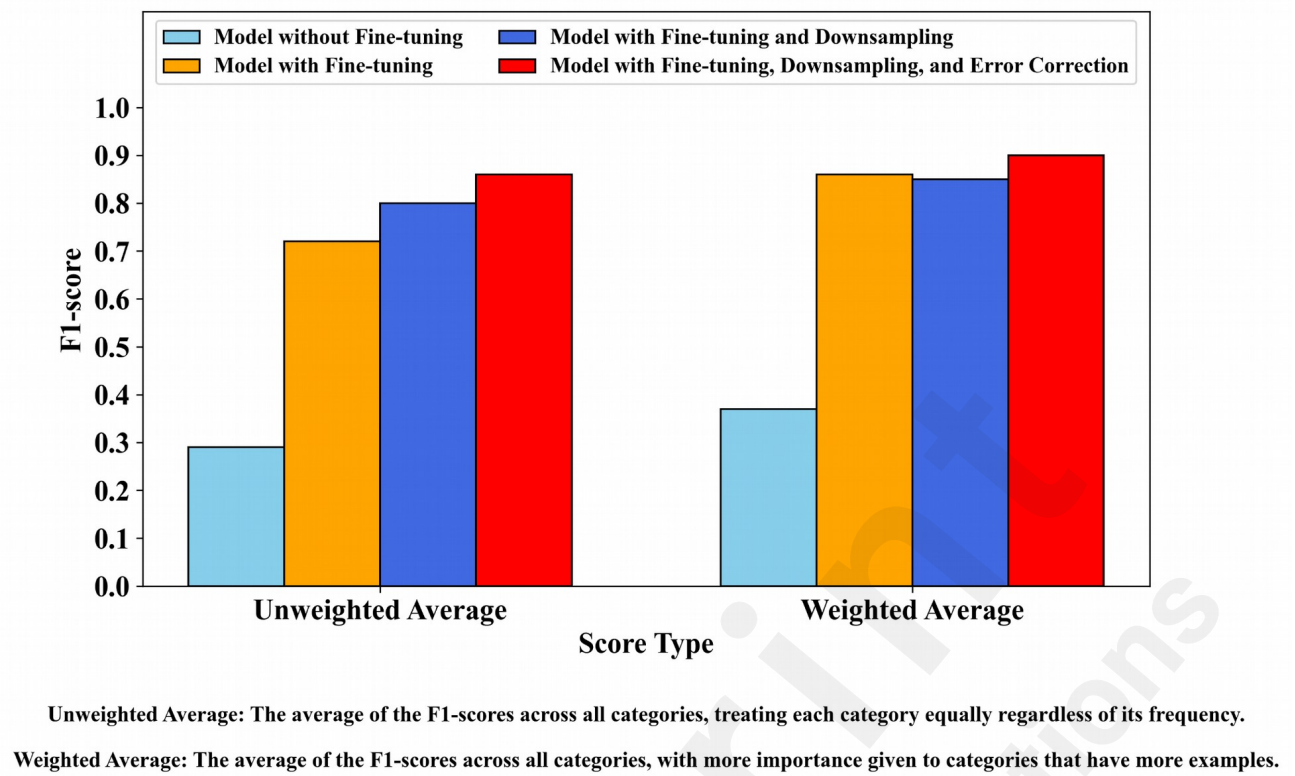


Figure 3. F1-scores Showing Recognition Accuracy for Different Large Language Model Approaches

Performance by Message Intent Category

Figure 4 depicts the performance of the different large language model training methods for each of the 25 intent categories based on unweighted F1-scores, i.e., treating each category equally regardless of its frequency. For the large language model without fine-tuning, most categories scored around 0.40, with a few categories slightly exceeding that score but the majority falling below it, indicating weak overall performance. The linguistically similar intent categories of “nrt_dontwork,” “nrt_itworks,” and “nrt_howtouse” performed especially poorly, indicating a critical need for fine-tuning.

With fine-tuning on the entire training dataset, most intent category scores improved to around 0.70. But linguistically similar intent categories such as “nrt_dontwork”, “nrt_howtouse”, and “nrt_od” continued to show relatively low performance. Unsurprisingly, the dominant “off-topic” intent category achieved a score above 0.90, reflecting a classification bias likely due to its significantly larger number of samples.
When we downsampled the dominant “off-topic” category in the training dataset for use in fine-tuning, the results demonstrated solid performance, with many categories reaching close to a 0.80 score. Most category scores improved compared to the earlier approach without downsampling. However, certain linguistically similar categories like “nrt_dontwork,” “nrt_itworks,” “nrt_howtouse,” “nrt_nauseous,” “nrt_mouthirritation,” and “nrt_od” fell below the 0.80 mark, suggesting room for further improvement.

The final approach of fine-tuning after both downsampling and error correction achieved the highest performance across virtually all intent categories. The improvement highlights the effectiveness of the error correction process and reinforces the importance of providing cleaner, less noisy data to the

large language model. One notable exception is that fine-tuning with downsampling, compared to fine-tuning alone, reduced the score for the “off-topic” category. As the “off-topic” category had an overwhelming number of samples, this boosted its score at the cost of performance on other categories, many of which plateaued at scores of around 0.70. Both downsampling methods reduced the dataset imbalance, allowing underrepresented categories to gain more importance and achieve better scores. Even after downsampling, the “off-topic” category score remains acceptable at about 0.77, especially considering that it represents irrelevant or uninformative content for our application.

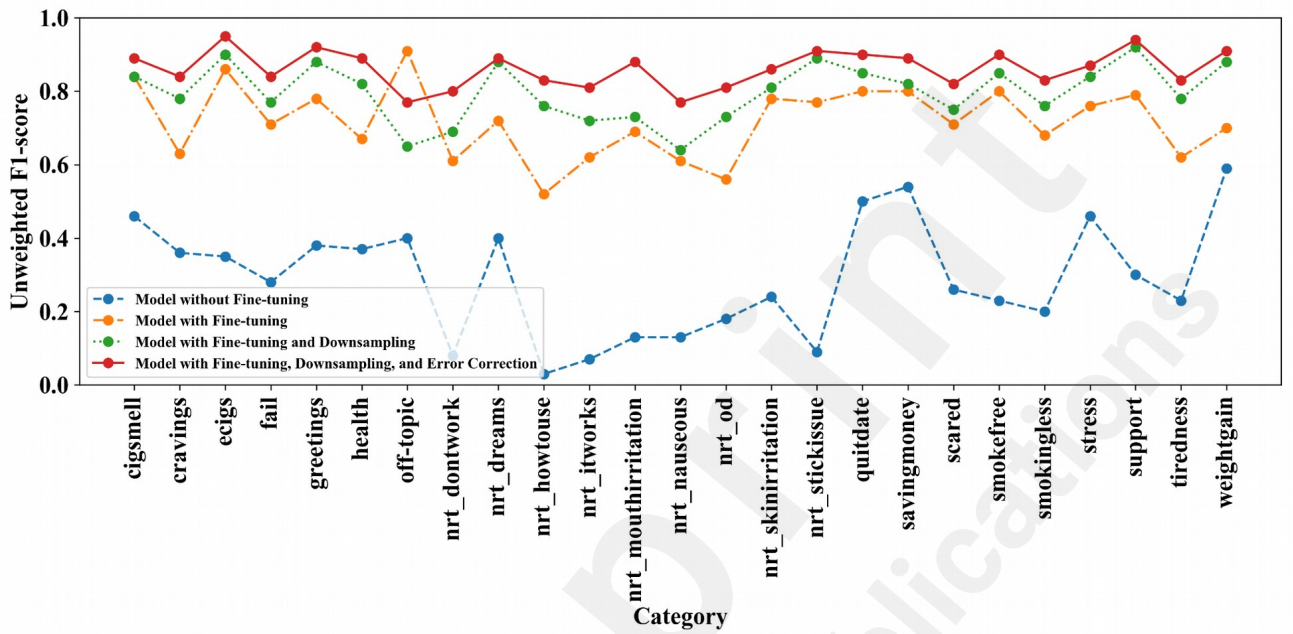


Figure 4. Unweighted F1-scores by Intent Category for Different Large Language Model Methods

Total Statistics on Model-Human Disagreements

Figure 5 illustrates the model-human disagreements across the various large language model training methods used. The left side of the figure includes the “off-topic” intent category in the test data, while the right side excludes it. In the scenario that includes “off-topic”, the large language model with no fine-tuning showed the highest rate of model-human disagreements, over 70%. Fine-tuning alone or fine-tuning with downsampling lowered the model-human disagreement rate to 14%. Fine-tuning, downsampling and error correction dropped it further to 10%. These results indicate that fine-tuning the model improves performance, and incorporating error correction further enhances it.

Comparable rates of model-human disagreement for the fine-tuned large language model in the test dataset, both with and without downsampling in the training dataset, might misleadingly suggest that downsampling for training has no impact. However, the underlying reason is that, without downsampling, there was a disproportionately large amount of “off-topic” data. This caused the model to become biased toward predicting “off-topic,” and, as the dominant category, correct predictions of it lowered the overall disagreement rate in the test dataset.

The right side of Figure 5, which excludes the “off-topic” data in the test dataset (only), provides a clearer picture. Here, the large language model without fine-tuning still performed the worst, with around 67% model-human disagreement. Fine-tuning alone lowered the disagreement rate to 26% revealing that its aforementioned performance was inflated due to the bias from excessive “off-topic” examples. Fine-tuning and downsampling lowered the disagreement rate substantially more to 14%, clearly demonstrating the benefit of downsampling. Finally, fine-tuning, downsampling, and error

correction reduced the disagreement rate to just 9.5%, highlighting the effectiveness of this combined approach.

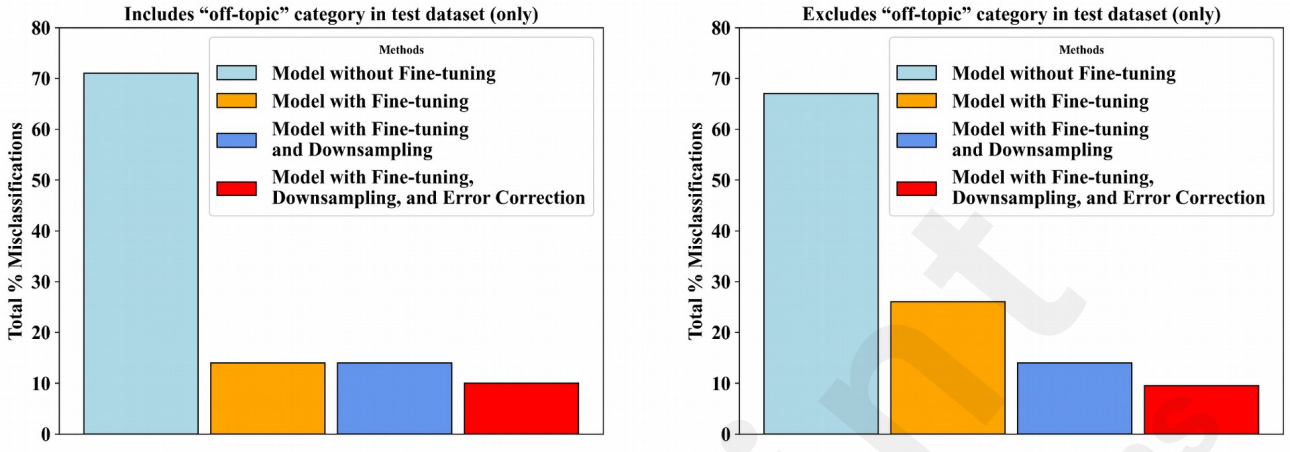


Figure 5. Model-Human Disagreements for Different Large Language Model Methods

Note– The left panel includes the dominant “off-topic” intent category in the test dataset, while the right panel excludes it from the test dataset, to show the benefit of downsampling “off-topic” in the initial training dataset.

Statistics on Model-Human Disagreements

Figure 6 shows the percentages of model-human disagreements by intent category for the different large language model training methods. It highlights which categories are commonly confused with others. This information provides insight into each training method’s specific weaknesses.

For the large language model with no fine-tuning, the bar graph shows a heavy presence of pink bars, indicating that the category “nrt_howtouse” was frequently and incorrectly predicted for many other categories. For some categories, over 50% of their instances were misclassified as “nrt_howtouse”. In fact, for 17 out of the 25 categories, this incorrect prediction dominated, suggesting that the model without fine-tuning was heavily biased toward the “nrt_howtouse” category and failed to properly distinguish among the other categories.

A similar issue surfaced with the fine-tuned large language model, but this time the bias was toward using the “off-topic” category. Here, 24 out of 25 categories were most commonly and wrongly predicted as “off-topic,” showing that the overrepresentation of “off-topic” examples in the training data led to a skewed model. This result highlights a major drawback of fine-tuning alone, as category imbalance is not addressed and leads to poor predictions for underrepresented categories.

Compared to these simpler methods, fine-tuning and downsampling showed more balance in mispredictions. No single category dominated, as shown by the absence of any overwhelmingly frequent color in the bar graph. This demonstrates that downsampling effectively reduced bias toward choosing the dominant category and led to a more equitable treatment across categories. Still, minor weaknesses remained, for example, the “smokefree” and “support” categories were incorrectly predicted for other categories.

Finally, fine-tuning with downsizing and error correction further improved performance by reducing the already low error rates. Only the “support” category was incorrectly predicted for other classes, seven in all, but with a very low misprediction rate of about 5%. This suggests that this combined approach produced the most robust and balanced classification results among all the methods.

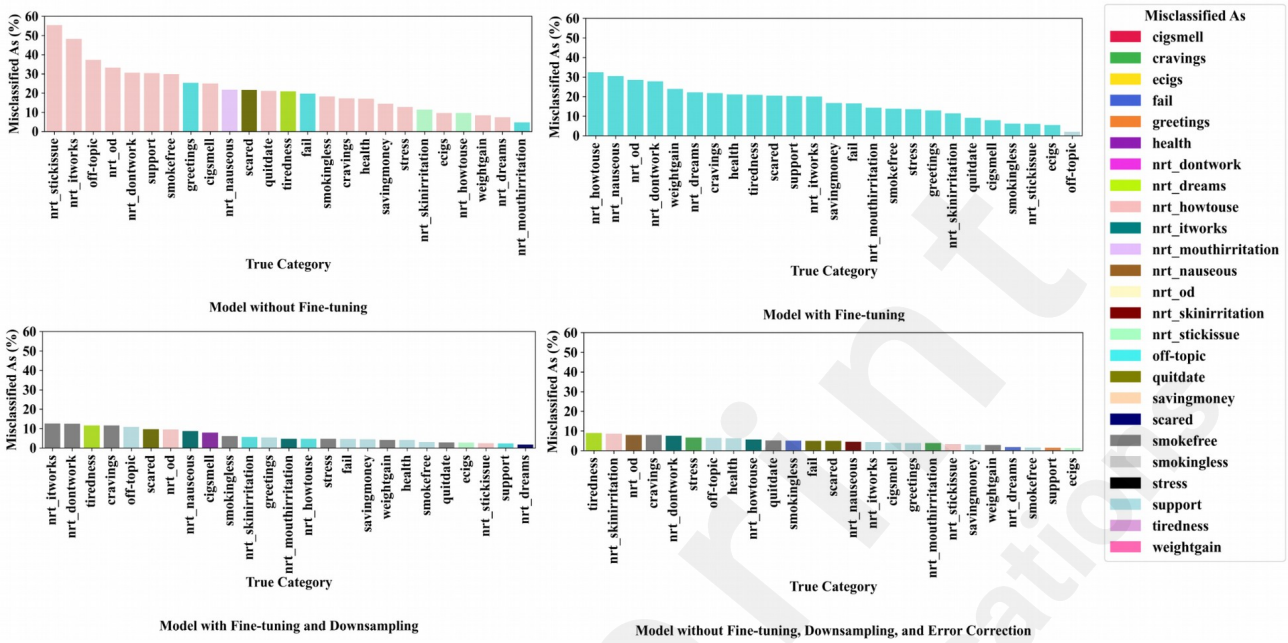


Figure 6. Most Common Model-Human Disagreements by Intent Category

Human-Model Disagreements and Human Error Correction

Table 3 presents statistics related to when the model and humans disagreed on the intent of user messages due to model mispredictions or human annotation errors or both. These statistics offer insights into the accuracy of human annotators and the potential of a large language model to sometimes outperform them.

In the training dataset (75% of all available data except the test data), the model and human annotators disagreed on the message intent 0.7% of the time. The human annotators were incorrect in 59% of these instances and yet, in all these cases, the model managed to accurately predict the users’ intents. The model mispredicted in 30% of these instances, and both the humans and model were erroneous in 11% of the instances. In sum, while the large training dataset did not include many model-human disagreements (<1% of all user messages), many such disagreements were due to human annotation errors, indicating that retraining the model on corrected data should lead to performance improvements.

In the evaluation dataset (25% of all available data except the test data), the model and human annotators disagreed on the message intent 15% of the time. Human annotators were wrong in 42% of these instances and notwithstanding the model managed to predict these intents correctly. The

model mispredicted in 47% of these instances, and both humans and model were wrong in 11% of these instances. In the test dataset (Holdout data from group 30 to 36), the model and human annotators disagreed on the message intent 15% of the time. Human annotators made errors in 39% of these cases and nevertheless the model accurately predicted the users' intents. The model mispredicted in 55% of these instances, and both model and human were incorrect in 6% of these instances.

Overall, this analysis reveals the value of using large language models not only for classifying the intent of users' messages, but also for uncovering human annotation errors in annotating these intents. The relatively high rate of model correctness in cases where human annotations were wrong reinforces the idea that correction and retraining can meaningfully improve model performance.

Table 3. Data Corrections Statistics for Model-Human Disagreements on Intents

Dataset	% of Total Dataset with Human-Model Disagreements	Humans Wrong, Model Correct	Model Wrong, Humans Correct	Both Humans and Model Wrong
Train	0.7%	59%	30%	11%
Evaluation	15%	42%	47%	11%
Test	15%	39%	55%	6%

Discussion

Principal Findings

Large language models have become the new state-of-the-art models in natural language processing due to their ability to recognize complex linguistic patterns and subtle semantic distinctions. Leveraging large language models for classification tasks like recognition of users' message intents can significantly enhance performance over earlier models. However, our analysis indicates that large language models may be insufficient for intent recognition in mobile health interventions when user messages are short and subtle message differences can profoundly affect meanings. We studied user messages in online smoking-cessation support groups, where there was a profound difference between saying "quit" or "not yet quit," "NRT works" or "NRT doesn't work," and "quit day 1" or "quit day 21" and our off-the-shelf large language model was not always able to detect these nuances. The model's performance was unacceptably poor, until we fine-tuned it on a domain-specific dataset where users' message intents were annotated using laborious human annotation. The model before fine-tuning performed poorly because the domain-specific text often contained unique terminology, implicit meanings, and contextual subtleties that a general-purpose model was unable to fully grasp without exposure to relevant training data.

We found that fine-tuning the LLM on domain-specific data led to substantial performance gains. The fine-tuning process allowed the model to familiarize itself with the characteristics of the dataset and better capture the intents behind the texts. Our fine-tuned large language model very substantially outperformed the same model without fine tuning; in fact, fine-tuning produced the greatest level of improvement in intent recognition, more than our three additional steps combined. A

key challenge we identified during fine-tuning was data imbalance, particularly due to the dominance of one category of intents annotated “off-topic.” This imbalance caused the model to overpredict the majority (dominant) category, thereby degrading performance on all other categories. To address this issue, we implemented downsampling of the majority category before fine-tuning. This adjustment produced a notable 11% increase in unweighted F1-score, reflecting a significant improvement in balanced performance across all intent categories. Individual category F1-scores also showed consistent gains.

Another critical insight from our study was the detection of dataset inconsistencies and noise, often stemming from human error in the annotation of users’ message intents. To tackle this, we analyzed model-human disagreements, identified incorrectly annotated samples, and corrected them. Training the model on the corrected dataset led to an additional 8% improvement in unweighted F1-score, highlighting the importance of clean and accurate dataset annotations. This reinforces the conclusion that large language models perform best at specialized classification tasks when trained on high-quality datasets.

Comparison with Previous Work

Previous work [17] studied intent classification with the same dataset, comparing Random Forest to BERT which was a revolutionary large language model introduced by Google in 2018. The study dataset, annotated by humans to indicate the different user message intents, was identical to that used here, except the two smallest intent categories were combined. The findings showed that BERT, fine-tuned on the domain-specific annotated dataset, outperformed Random Forest. BERT yielded an unweighted F1-score of 67.4% and a weighted F1-score of 84.3% which were reasonable scores but inadequate for implementation of the large language model in a mobile health intervention. This work did not include a strategy to address the substantial imbalance in intent category exemplars in the dataset.

In the current work, we started with a more up-to-date, public-domain large language model, Llama-3 8B from Meta, to conduct intent classification with our 25 message intent categories. Llama-3 8B, fine-tuned on the same domain-specific annotated dataset, achieved an unweighted F1-score of 72% and a weighted F1-score of 86%, outperforming BERT by 7% in unweighted and 2% in weighted scores, respectively.

Our final large language model training method with fine-tuning, downsampling, and error correction yielded the best results, producing an unweighted F1-score of 86% and a weighted F1-score of 90%. These results represent a 28% improvement in the unweighted score and a 7% improvement in the weighted score over the BERT model, highlighting the effectiveness of our comprehensive approach.

Limitations and Future Directions

Our final approach to large language model training with fine-tuning, downsampling, and error correction achieved impressive performance in our dataset, but it had some limitations. One notable issue was the model’s confusion in handling temporal expressions, particularly when distinguishing between similar intent categories such as “quitdate” and “smokefree.” For instance, a user message like “Day 1, very excited” was correctly identified as “quit date” since it referred to the user’s first

day of quitting. However, user messages like “Day 2, going good” and “Day 3!” were often classified as “quit date” when the correct intent classification was “smokefree” indicating they had successfully remained smokefree past their quit date. The model mispredictions suggest it may have relied too heavily on specific keywords or patterns (e.g., “Day X”) rather than truly understanding the underlying intent conveyed in the message.

This behavior highlights an important limitation of relying solely on fine-tuning to adapt large language models to domain-specific datasets. While fine-tuning allows the model to become familiar with the data, it may not be sufficient to capture deeper contextual or semantic distinctions, especially those that are subtle, underrepresented, or inconsistent within the dataset. Certain types of domain-specific reasoning and implicit knowledge are difficult for the model to infer from training data alone.

To overcome this, our future work will explore ways to explicitly introduce domain knowledge into the learning process. This involves identifying patterns and semantic rules that are not easily learnable through data-driven methods and incorporating them into the model’s decision-making process. By combining data-driven learning with carefully introduced domain knowledge, we aim to help the model better understand nuanced intent distinctions, reduce mispredictions, and ultimately achieve more consistent and interpretable predictions across all intent categories.

Conclusion

Mobile health smoking-cessation interventions that provide real-time interactions with users have considerable potential to provide timely advice and support for quitting. However, their effectiveness is often limited by low user engagement and lack of personalization. Detecting user intent from their utterances can address this problem by identifying individual needs and enabling timely, intelligent responses through an artificial AI agent or chatbot. However, existing systems struggle with a large number of categories of users’ message intents, nuanced language, and imbalanced datasets where critical message intents are often underrepresented. To overcome these challenges, we have created a large language model that successfully conducts intent detection with complex and imbalanced data. We compared its performance across several settings: A state-of-the art but off-the-shelf large language model without fine-tuning, the model after fine-tuning on domain-specific data, fine-tuning with downsampling of the most dominant category, and lastly, fine-tuning with downsampling and retraining after correction of human annotation errors detected from model-human disagreements. Our final method, which combines all three enhancements, achieves 86% unweighted and 90% weighted F1-score and shows 28% and 7% improvements respectively over previous work on the same dataset [17].

Conflicts of Interest

None declared.

References

1. Shiu E, Hassan LM, Walsh G. Demarketing tobacco through governmental policies - the 4 P's revisited. J Bus Res;2007(62):269-278.

2. He L, Balaji D, Wiers RW, Antheunis ML, Krahmer E. Effectiveness and acceptability of conversational agents for smoking cessation: a systematic review and meta-analysis. Nicotine Tob Res;2023(25):1241-1250. doi:10.1093/ntr/ntac281

3. Bendotti H, Lawler S, Chan GCK, Gartner C, Ireland D, Marshall HM. Conversational artificial intelligence interventions to support smoking cessation: a systematic review and meta-analysis. Digit Health;2023(9):1-13. doi:10.1177/20552076231211634

4. Regmi K, Kassim N, Ahmad N, Tuah N. Effectiveness of mobile apps for smoking cessation: a review. Tob Prev Cessat;2017(3):12. doi:10.18332/tpc/70088

5. Tong H, Quiroz J, Kocaballi AB, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med;2021(148):106532. doi:10.1016/j.ypmed.2021.106532

6. Free C, Knight R, Robertson S, et al. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. Lancet;2011(378):49-55.

7. Free C, Whittaker R, Knight R, Abramsky T, Rodgers A, Roberts IG. Txt2stop: a pilot randomised controlled trial of mobile phone-based smoking cessation support. Tob Control;2009(18):88-91. doi:10.1136/tc.2008.026146

8. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? an experimental study. Digit Health;2019(5):2055207619880676. doi:10.1177/2055207619880676

9. Stead, L., Carroll, A., & Lancaster, T. (2017). Group behaviour therapy programmes for smoking cessation. Cochrane Database of Systematic Reviews, 3, 1-90.

10. Mersha, A. G., Bryant, J., Rahman, T., McGuffog, R., Maddox, R., & Kennedy, M. (2023). What are the effective components of group-based treatment programs for smoking cessation? A systematic review and meta-analysis. Nicotine and Tobacco Research, 25(9), 1525-1537.

11. Struik, L., Khan, S., Assoiants, A., & Sharma, R. H. (2022). Assessment of social support and quitting smoking in an online community forum: Study involving content analysis. JMIR Formative Research, 6(1), e34429.

12. Pechmann C, Pan L, Delucchi K, Lakon CM, Prochaska JJ. Development of a twitter-based intervention for smoking cessation that encourages high-quality social media interactions via automessages. J Med Internet Res;2015(17):e50. doi:10.2196/jmir.3859

13. Pechmann C, Delucchi K, Lakon CM, Prochaska JJ. Randomised controlled trial evaluation of tweet2quit: a social network quit-smoking intervention. Tob Control;2017(26):188-194.

14. Joyce E, Kraut RE. Predicting continued participation in newsgroups. J Comput Mediat Commun;2006(11):723-747.

15. K. Kosyluk, T. Baeder, K. Y. Greene, J. T. Tran, C. Bolton, N. Loecher, et al. (2024), Mental distress, label avoidance, and use of a mental health chatbot: Results from a US survey JMIR

Formative Research Vol. 8 Pages e45959

16. J. J. Prochaska, E. A. Vogel, A. Chieng, M. Kendra, M. Baiocchi, S. Pajarito, et al. (2021), A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study Journal of Medical Internet Research Vol. 23 Issue 3 Pages e24850

17. Giyahchi T, Singh S, Harris I, Pechmann C. Customized training of pretrained language models to detect post intents in online health support groups. In: Shaban-Nejad A, Michalowski M, Bianco S, editors. *Multimodal AI in healthcare: a paradigm shift in health intelligence*. Cham: Springer International Publishing; 2023. p. 59–75. doi:10.1007/978-3-031-14771-5_5

18. Susan McRoy, Majid Rastegar-Mojarad, Yanshan Wang, Kathryn J Ruddy, Tufia C Haddad, and Hongfang Liu. Assessing unmet information needs of breast cancer survivors: Exploratory study of online health forums using text classification and retrieval. JMIR Cancer, 4(1):e10, May 2018.

19. Jihyun Park, D. Kotzias, Patty B Kuo, IV RobertLLogan, Kritzia Merced, Sameer Singh, Michael J Tanana, Efi Karra Taniskidou, J. Lafata, David C. Atkins, M. Tai-Seale, Zac E. Imel, and Padhraic Smyth. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. Journal of the American Medical Informatics Association : JAMIA, 26:1493 – 1504, 2019.

20. Zeng K, Pan Z, Xu Y, Qu Y. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. In: JMIR Publications, editor. JMIR Med Inform. 2020 Jul 1;8(7):e17832. doi:10.2196/17832.

21. Li X, Shu Q, Kong C, Wang J, Li G, Fang X, Lou X, Yu G. An intelligent system for classifying patient complaints using machine learning and natural language processing: development and validation study. In: JMIR Publications, editor. J Med Internet Res. 2025;27:e55721. doi:10.2196/55721. URL: https://www.jmir.org/2025/1/e55721

22. Babu A, Boddu SB. BERT-based medical chatbot: enhancing healthcare communication through natural language understanding. Explor Res Clin Soc Pharm;2024(13):100419. doi:10.1016/j.rcsop.2024.100419

23. Yu HQ, McGuinness S. An experimental study of integrating fine-tuned large language models and prompts for enhancing a mental health support chatbot system. J Med Artif Intell;2024(7):e8991. https://jmai.amegroups.org/article/view/8991

24. Zamani A, Reeson M, Marshall T, Gharaat MA, Foster AL, Noble J, Zaiane OR. Intent and entity detection with data augmentation for a mental health virtual assistant chatbot. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents; 2023 Sep 19-22; Würzburg, Germany. New York, NY: Association for Computing Machinery; 2023. p. 41. doi:10.1145/3570945.3607324

25. Aftab H, Gautam V, Hawkins R, Alexander R, Habli I. Robust intent classification using Bayesian LSTM for clinical conversational agents (CAs). In: Gao X, Jamalipour A, Guo L, editors. Wireless Mobile Communication and Healthcare. Cham: Springer International Publishing; 2022. p. 106-118. doi:10.1007/978-3-031-06368-8_8

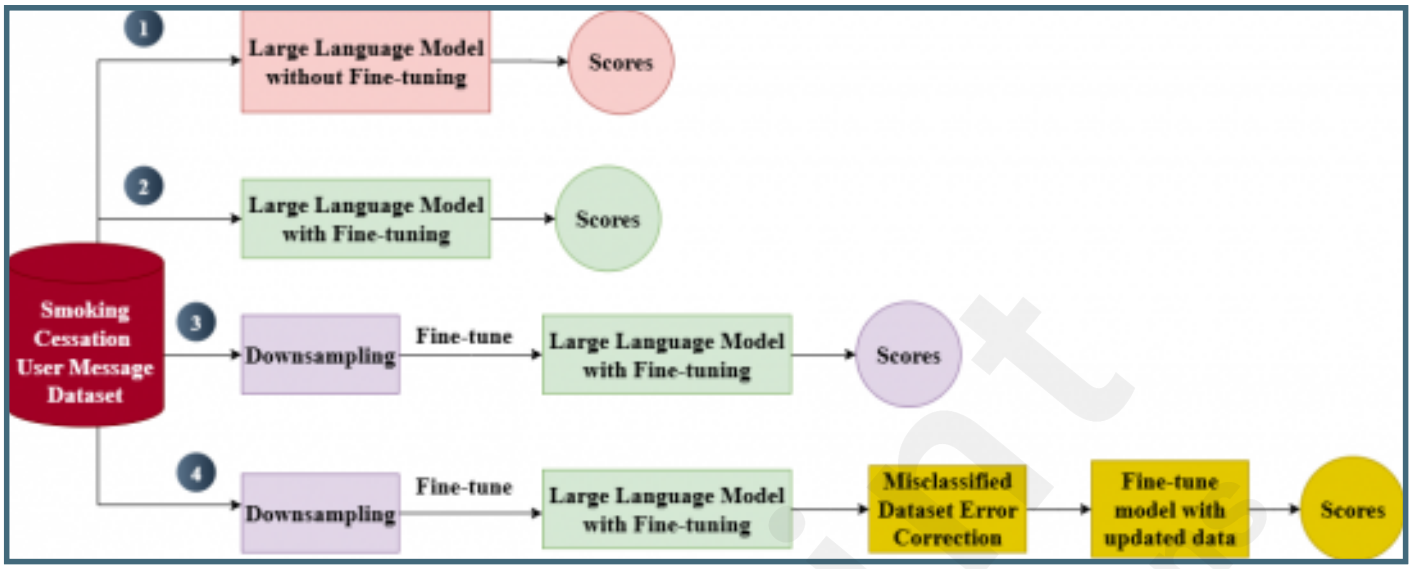
26. Zhang B, Tu Z, Hang S, Chu D, Xu X. Conco-ERNIE: complex user intent detect model for smart healthcare cognitive bot. ACM Trans Internet Technol;2023(23):21. doi:10.1145/3574135

27. Sakurai H, Miyao Y. Evaluating intention detection capability of large language models in persuasive dialogues. In: Ku L, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024 Aug; Bangkok, Thailand. Association for Computational Linguistics; 2024. p. 1635-1657. doi:10.18653/v1/2024.acl-long.90

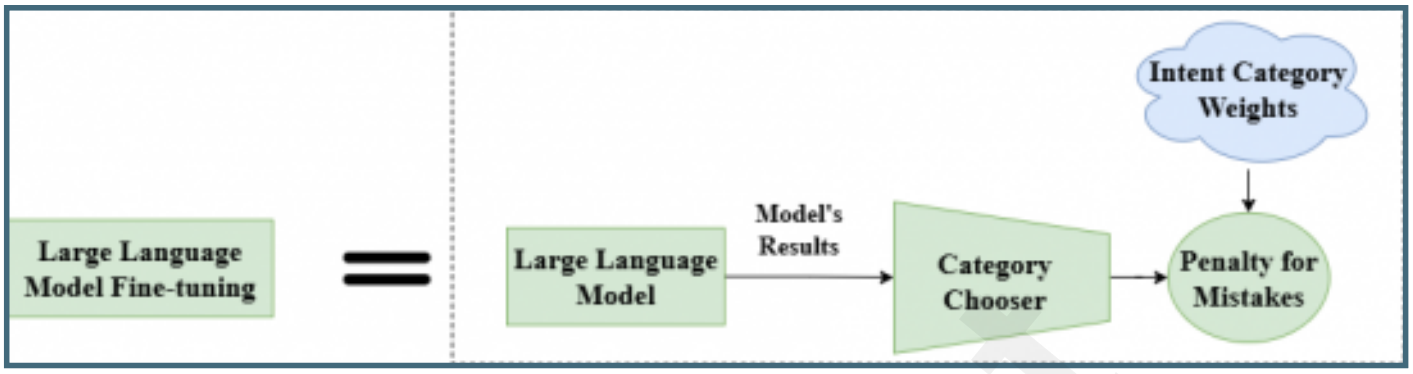
Supplementary Files

Figures

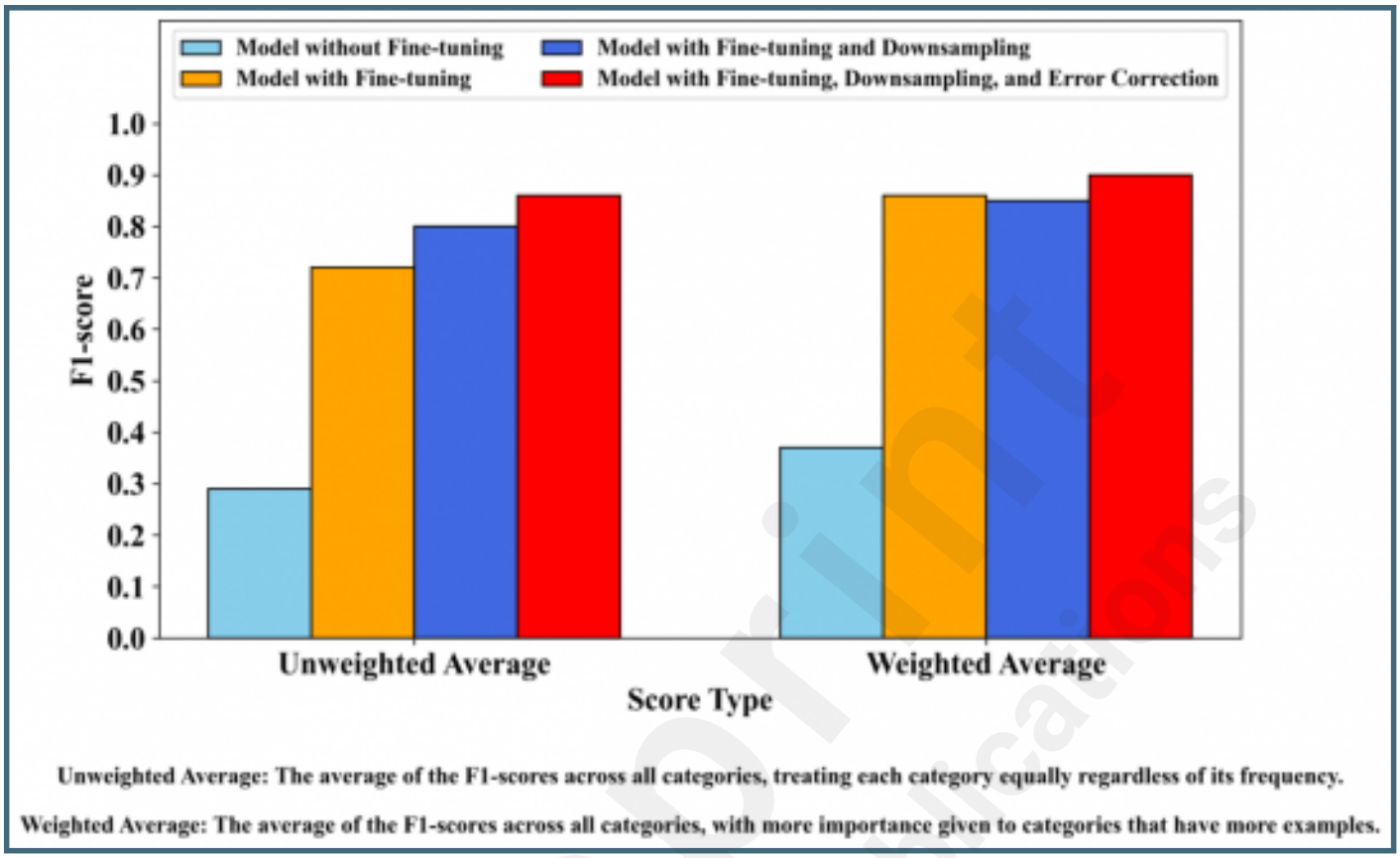
Overview of Methodology.



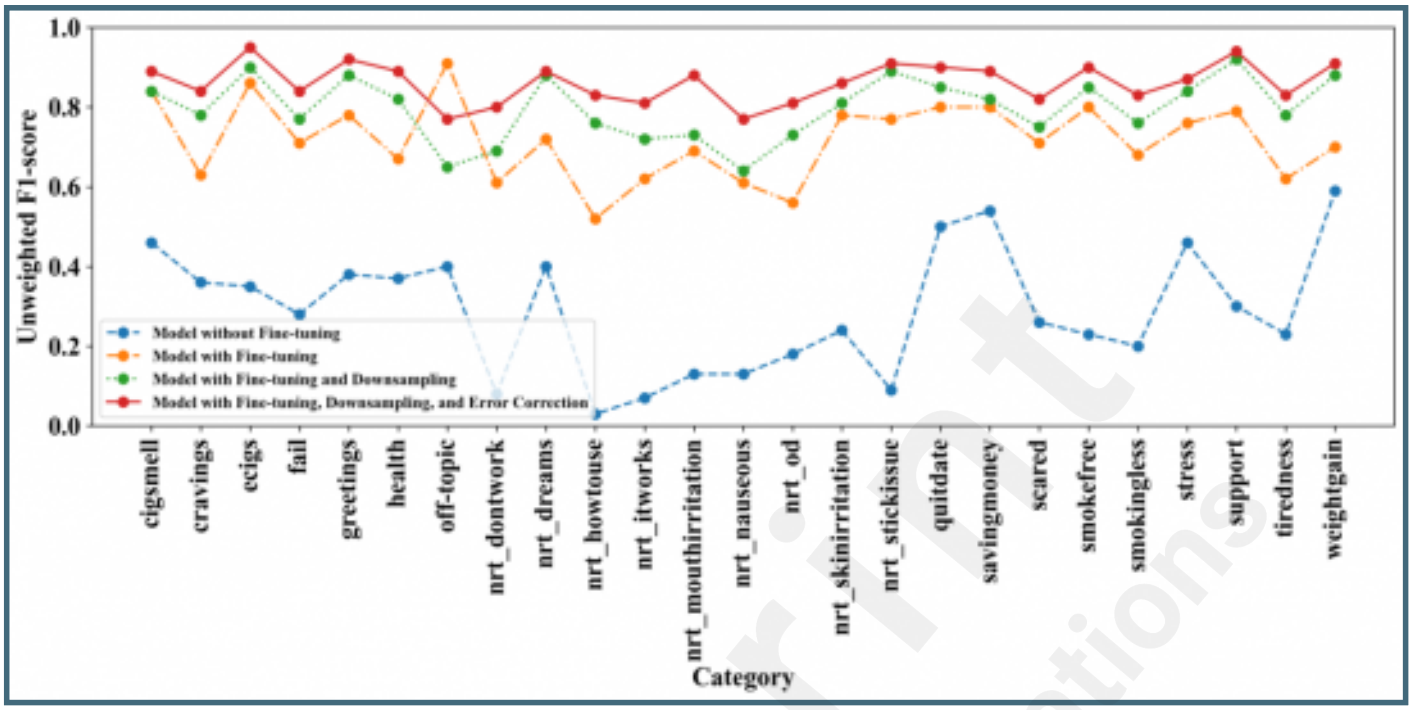
Overview of Large Language Model Fine-tuning.



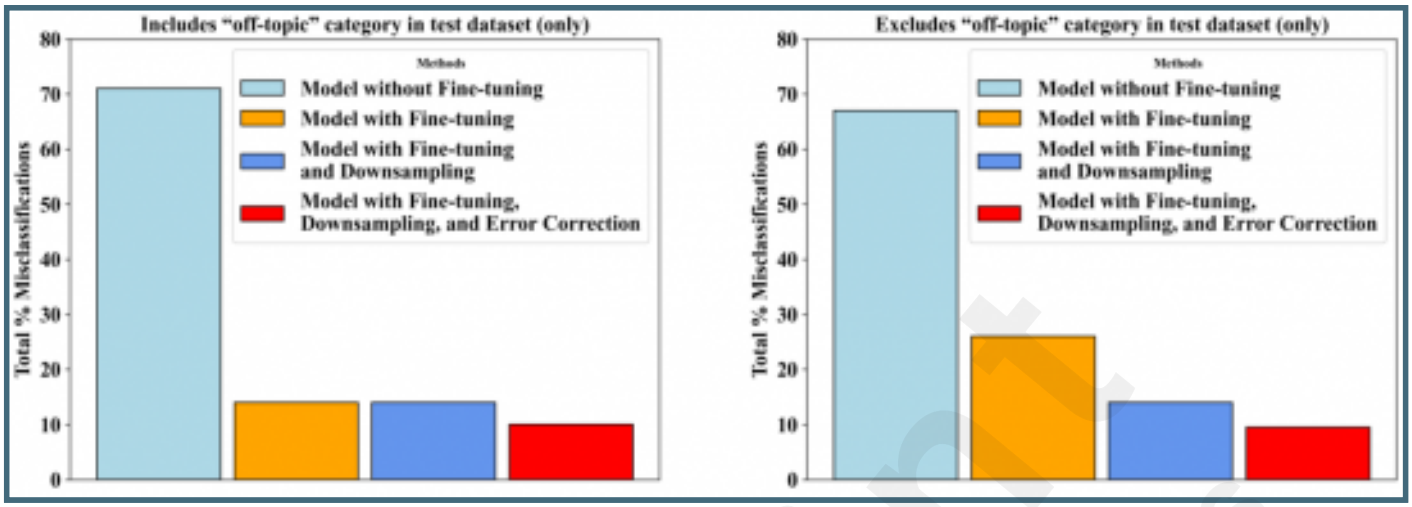
F1-scores Showing Recognition Accuracy for Different Large Language Model Approaches.



Unweighted F1-scores by Intent Category for Different Large Language Model Methods.



Model-Human Disagreements for Different Large Language Model Methods.



Most Common Model-Human Disagreements by Intent Category.