Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Leveraging Large Language Models for Review Classification and Rating Estimation of Mental Health Applications

Qile Wang*, Moath Erqsous*, Prerana Khatiwada, Abhishek Karwankar, Fatimah Mohammad Alhassan, Aishwarya Chandrasekaran, Benita Abraham, Faith Lovell, Andrew Anh Ngo, Matthew Louis Mauriello

University of Delaware, USA

{kylewang, merqsous, preranak, karwabhi, alhassan, aishc, beniabra, faithlov, ango, mlm}@udel.edu

Abstract

Large Language Models (LLMs) can analyze large datasets semantically. However, research on applying LLMs for mental health text classification is relatively new and developing. Existing methods often use supervised, deep, and reinforcement learning, which rely heavily on fine-tuning and reward models. To investigate whether LLMs can assist in recommending mental health apps based on user reviews, our study collected approximately 200k user reviews from 73 mental health mobile applications. We instructed selected LLMs to classify individual reviews into 1-5 star ratings, subsequently averaging these results to derive an overall rating for each app reflecting current user feedback. While the best supervised learning method in our experiments achieved an F1-Score of 0.79 which required significantly more human effort, the GPT-4 and Gemini 1.5 Pro delivered a strong 'out-of-the-box' performance with an overall F1-Score of 0.76. We provide further statistical comparisons and discussions of the performance of these models for the text classification task. Using a crowdsourcing platform to determine agreement levels, we observed that human ratings align closely with GPT ratings. In addition, we analyze specific features and concerns highlighted in mental health app reviews. Alongside our analysis, we make our data available for further experimentation and benchmarking.

Datasets — https://github.com/Sensify-Lab/MHARD

1 Introduction

According to the WHO[1], around 970 million people globally are affected by mental health conditions. With more than 10,000 mental health mobile applications (apps) available, individuals depend on app ratings, user reviews, social media feedback, and personal recommendations for decision-making (Schueller et al. 2018).

These apps offer scalable resources for individuals lacking access to traditional care; however, reliable methods are needed to assess their quality, which should take into account app ratings and large-scale user reviews (Nguyen et al. 2024; Schueller et al. 2018; Miner et al. 2016). Large Language Models (LLMs) have demonstrated outstanding performance across a wide array of natural language processing (NLP) tasks, including text classification (Yang, Cao, and Fan 2024; Sun et al. 2019; Howard and Ruder 2018), sentiment analysis (Raffel et al. 2020; Miner et al. 2016), and language generation (Radford et al. 2019; Yang et al. 2019; Liu et al. 2019). These models, pre-trained on vast amounts of text data, exhibit remarkable capabilities in generating high-quality, human-like text with minimal task-specific fine-tuning (Brown et al. 2020; Vaswani et al. 2017; Devlin et al. 2018). Recent studies have highlighted LLMs' potential to improve text classification without extensive feature engineering (Sun et al. 2019; Yang et al. 2022; Clark et al. 2020). This has sparked interest in comparing the efficacy of LLMs with traditional supervised machine learning (ML) models in text classification tasks (Howard and Ruder 2018; Arslan et al. 2021). With the goal of building an app recommendation system from reviews, we first need to explore whether LLMs can accurately classify ratings. This context offers a valuable testbed for evaluation when compared to supervised ML models. Since the use of LLMs for mental health app review classification is under-explored, we focus on evaluating *"out-of-the-box"* LLMs for these reviews, expecting that our findings will generalize to other online review systems. Our study addresses the following research questions (RQs):

- ***RQ1****: How does the performance of LLM compare to supervised machine learning models regarding a text classification task, specifically for mental health apps?*
- ***RQ2:*** *What is the impact of LLM-generated predictions on mental health app rankings?*
- ***RQ3****: What is the estimated level of agreement between human rating evaluations and LLM-generated ratings in app review assessments?*
- ***RQ4****: What aspects lead to divergent ratings in mental health apps?*

To address these questions, we collected over *200,000* user reviews from 73 mental health apps. We utilize three different embeddings in combination with five supervised classification algorithms along with three additional deep-learning methods and compare their performance with advanced LLM-generated ratings. Particularly, we aim to benchmark the performance of GPT (Ouyang et al. 2022), Gemini (Reid et al.

*These authors contributed equally.

[1]https://www.who.int/news-room/fact-sheets/detail/mental-disorders

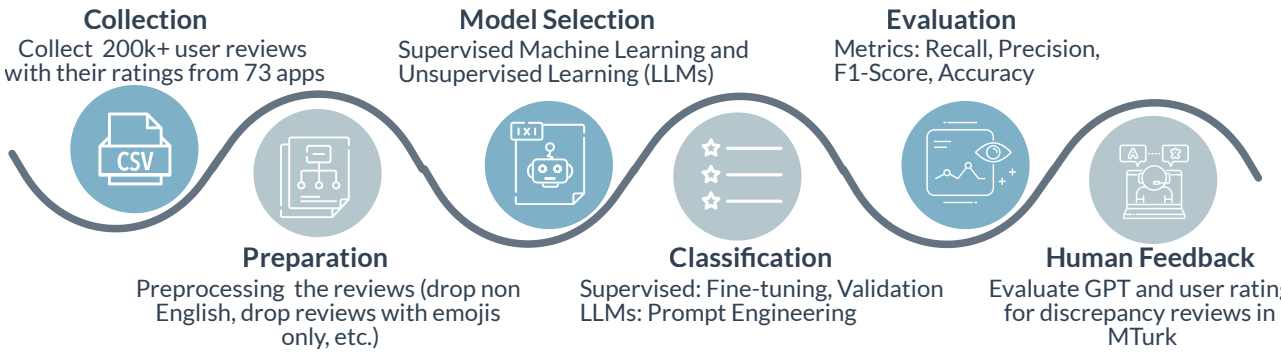

Figure 1: Overview of methods used to predict ratings for user reviews of mental health apps.

2024), and Llama (Touvron et al. 2023), in a 5-class classification task. Figure 1 illustrates each stage of the process. Our results indicate that LLMs generally achieve comparable performance. Where discrepancies exist, we engaged Amazon Mechanical Turk (MTurk) workers to provide human assessments of star ratings and their rationale. We found that over half of the GPT-4 generated ratings align closely with human judgments, particularly within the mental health app domain. This research contributes to understanding the practical implications of deploying LLMs in real-world applications and highlights their potential to enhance the reliability and efficiency of app recommendation and rating systems, offering valuable insights that can extend into broader and diverse domains.

2 Related Work

2.1 Evaluating Mental Health Applications

There are many examples of research that have been conducted to evaluate the use and reliability of applications for a variety of mental health challenges. Bakker et al. (2016) conducted a systematic review, stressing the importance of evidence-based development of mental health apps. Grainger et al. (2020) emphasized the need for rigorous reporting and systematic methods in analyzing health app reviews, especially regarding app store searches and app appraisal for patient or clinician-focused apps. Chandrasekaran et al. (2025) highlighted the need for mental health apps to be thoroughly validated within their intended context to ensure accuracy and trustworthiness. Others have shown the advantages of recommender systems for mental health apps, such as a reduction in choice overload and improvement to the digital therapeutic alliance (Kuo and Li 2023; Etemadi et al. 2023; Ricci, Rokach, and Shapira 2021). Wisniewski et al. (2019) evaluated top-rated health apps for mental health and comorbid conditions. They analyzed their attributes, effectiveness, and quality of consumer ratings and found a correlation between update frequency and app quality.

While researchers have analyzed mental health mobile applications for managing psychological difficulties (Badesha, Wilde, and Dawson 2022) and the perspectives of users regarding these applications (Nguyen et al. 2024; Funnell et al. 2022), the same user feedback can be leveraged to determine ratings and provide app recommendations to users accordingly. Insights from user reviews have been used in mental health apps (Alqahtani and Orji 2020), and qualitative thematic analysis is used to assess user feedback on mental health apps (Malik, Ambrose, and Sinha 2022). However, we hypothesize that predicting ratings from user reviews to generate mental health app recommendations is objective and might be more reliable.

Several studies (Haque and Rubya 2022; Jo et al. 2023; Balaskas et al. 2022) have utilized mental health user reviews to gain insights into the mobile mental health application ecosystem to understand user needs. However, these studies have typically collected only a few thousand data points (number of user reviews n = 4923, 3268, 600, respectively). In our study, we gathered a much larger dataset (n = 200,973) to enhance efficiency and significantly expand the depth of analysis. Specifically, we conduct multi-class classification to predict user ratings, which can help build a recommendation system for mental health applications. In addition, leveraging this extensive dataset can yield substantial benefits for application designers, policymakers, and researchers, similar to advancements seen in the medical field (Bond et al. 2023; Aryana, Brewster, and Nocera 2019).

2.2 LLMs for Classification Tasks

Numerous studies have focused on applying traditional NLP methods and data mining techniques to classify and summarize opinions in various contexts (Alzetta et al. 2024; Yang, Cao, and Fan 2024). The use of pre-trained transformer-based models, such as BERT and RoBERTa, has shown promising results in detection tasks using social media text data (Tian et al. 2023). Furthermore, self-supervised pre-trained language models have been proven beneficial for text classification tasks, providing a more efficient alternative to training models from scratch (Myagmar and Li 2019). The versatility and effectiveness of language models in improving text classification accuracy and performance have made them a valuable tool in domains like sentiment analysis, news article evaluation, and job description classification (Zhao, Zhang, and Hopfgartner 2021; Skondras, Zervas, and Tzimas 2023). Research has also explored the feasibility of

LLMs and GPT models in decision-making (Brin et al. 2023; Checco et al. 2021). However, challenges such as model interpretability and ethical considerations persist in integrating these advanced models into decision-making frameworks. Despite these challenges, the research highlights the transformative impact of LLMs and GPT models in revolutionizing decision-making processes and improving predictive capabilities across various domains.

Specifically, LLMs are transforming text classification tasks by using their extensive training on vast corpora to categorize and analyze textual data accurately. Understanding these models facilitates nuanced and context-aware classification across various applications, from sentiment analysis to topic detection (Devlin et al. 2018; Brown et al. 2020). Their ability to discern intricate patterns in data demonstrates their pivotal role in improving automation and insight extraction in numerous sectors. Research on applying LLMs for classification tasks extends to various applications. Studies have explored LLMs to predict product reviews (Databricks 2023), which demonstrates how LLMs can streamline the process of sifting through vast quantities of user feedback to identify prevalent themes and sentiments, enhancing decision-making and responsiveness in customer-centric industries. Liu et al. (2023) benchmark LLMs on recommendation tasks such as rating predictions focusing on the Beauty category from the Amazon recommendation dataset. Limited research has explored user classification tasks based on health app reviews. A recent review of empirical evaluations of AI-based language assessments has highlighted many advantages of using LLMs to analyze natural language for assessing mental health instead of relying on traditional rating scales, such as richer information and a broader range of expression (Kjell, Kjell, and Schwartz 2024). Our work builds on this foundation by focusing on LLMs for review classification and rating estimation, specifically within the context of mental health applications, aiming to make automated assessments more accurate and useful in this important area.

2.3 Crowdsourced Labeling and Annotation

Crowdsourcing has become a widely adopted and effective method for managing labeling and annotation tasks of large datasets by using the collective intelligence of a large group of individuals. Researchers have used crowdsourcing for more efficient and cost-effective labeling purposes, often using platforms like MTurk to assign tasks from various labeling-intensive fields such as computer vision and natural language processing to a large number of workers (MacLean and Heer 2013; Su, Deng, and Fei-Fei 2012; Borromeo and Toyama 2015; Snow et al. 2008). However, one of the primary drawbacks is the potential for low-quality labels due to the varying levels of expertise among crowd workers (Allahbakhsh et al. 2013). This inconsistency in annotations can compromise the overall quality of the labeled data and impact the performance of downstream tasks (Sheng, Provost, and Ipeirotis 2008). Despite these drawbacks, crowdsourcing remains a popular choice for labeling tasks due to its cost-effectiveness and scalability (Drutsa et al. 2020; Rodrigues and Pereira 2018). In our study, we use crowdsourcing to compare the labels generated by models with the labels originally provided by app users. This comparison helps us understand and justify which rating they prefer. By evaluating how well the models' labels match up with those from actual users, we can provide a stronger reason for selecting one app over the other based on the accuracy and relevance of the labels.

3 Method

We collected user reviews from mental health apps to predict ratings using supervised and self-supervised models. Figure 1 illustrates an overview of methods used in this study.

3.1 Dataset

In this study, we collected user reviews from mental health mobile apps (with over 100k downloads) from the Google Play Store using SerpAPI[2] tool. To maintain consistency and privacy, non-English texts (8.16%) were excluded, and personal information was anonymized during text cleaning, like replacing the usernames with numbers. The resulting Mental Health App Reviews Dataset (MHARD) contains 200,973 user ratings and reviews across 73 unique apps, spanning between March 29, 2011, and July 11, 2023. It also includes additional attributes such as the number of likes and responses from the respective app business owners. The distribution of user ratings in our dataset is imbalanced, with approximately 60% of the ratings being five stars, 12.5% four stars, 5.5% three stars, 4.4% two stars, and 17.2% being only one star. This imbalanced dataset reflects the distribution (known as the J curve) found in online rating systems (Hu, Pavlou, and Zhang 2009). To address performance discrepancies across classes 1-5, we manually created two additional balanced datasets from the original dataset. Two-star ratings have the lowest 8,898 reviews. Therefore, our first balanced dataset (*n = 44490*) applied random undersampling without replacement. To match the original data size, our second balanced dataset (n = 200k) used random oversampling with replacement, resulting in 40,000 reviews for each class.

Since these ratings are not linearly scaled, we treat them as an ordinal categorical variable. Recognizing the ability of emojis to convey emotions and sentiments, they were retained in the user reviews. (e.g., *"I love the meditation and the content, but I wish that I didn't have to pay to get everything on the app! It's kind of maddening"* - rated 4 out of 5). The collection of this dataset captured a broad range of user reviews, from positive (77.43%) to negative (14.47%), as determined and annotated by VADER[3]. To protect privacy, we de-identified users' information by converting usernames into numerical IDs.

3.2 Supervised Machine Learning

For the supervised machine learning approach, our comparative analysis focused on three embeddings: FastText (Bojanowski et al. 2017), all-MiniLM-L12-v2 (Wang et al. 2020), and BERT-Large (Devlin et al. 2018). Data preprocessing, including normalization, tokenization, and removal of stop words, was autonomously managed by the language

[2]SerpAPI: https://serpapi.com

[3]https://pypi.org/project/vaderSentiment/

models. The chosen ML classification algorithms are K-Nearest Neighbor (**KNN**) (Peterson 2009), Random Forest (**RF**) (Breiman 2001), Decision Tree (**DT**) (Song and Ying 2015), Support Vector Machine[4] (**SVM**) (Hearst et al. 1998), and Neural Networks (**NN**) using PyTorch (Subramanian 2018). The selection of these algorithms was guided by their demonstrated efficacy and varied methodologies in handling classification tasks (Kowsari et al. 2019). Altogether, we categorize these 15 combinations of embeddings and classifiers as traditional machine learning. For instance, the use of BERT-Large embeddings paired with a neural network architecture. This hybrid approach enhances flexibility for specific tasks and can accommodate alternative frameworks such as RNNs. Furthermore, we include three state-of-the-art deep learning models in our experiment: direct BERT-Base [5] classifier (*number of parameters = 110M*), BERT-Large [6] classifier(*n = 340M*), and RoBERTa-Large (Liu et al. 2019) classifier (*n = 355M*).We chose BERT and RoBERTa for classification tasks due to their strong foundations in the transformer architecture introduced by (Naseer, Asvial, and Sari 2021). BERT, developed by Google, is effective in various NLP tasks because of its bidirectional context understanding. RoBERTa, developed by Facebook AI, enhances BERT by using a larger training dataset, training for longer durations, and removing the next sentence prediction objective, resulting in improved performance and robustness.

For our evaluation, we measure the Precision (**P**), Recall (**R**), F1-Score (**F**), and Accuracy (**Acc.**) of the classification models (Yacouby and Axman 2020; Chicco and Jurman 2020; Thabtah et al. 2020; Zhou et al. 2023). We use an 80-20 split for training and testing. Neural networks have been fine-tuned to attain acceptable performance. We explored how different layer configurations affect the classification task performance, identifying neural networks as optimal due to their flexibility. Our architecture starts with a fully connected layer to reduce input dimensionality, followed by a ReLU activation and a 0.5 dropout rate to prevent overfitting. Similar to the first sequence, another layer sequence refines feature reduction, leading to a final linear layer that categorizes inputs into five classes.

3.3 Large Language Model

We used several leading LLMs offered by the OpenAI[7], Meta[8] and Google DeepMind[9]. We initiated our experiment with GPT-3.5 Instruct (Model: *"gpt-3.5-turbo-instruct"*), distinguished by its succinct execution of commands. Subsequently, during our experiment, we incorporated the most updated version (as of March 2024) of GPT-3.5 Turbo (*"gpt-3.5-turbo-1025"*). For a broader analysis, we employed the advanced GPT-4 (*"gpt-4-0125-preview"*), which provides an expanded context window. As part of the benchmarking model comparison, we included Gemini 1.5 Flash (*"gemini-1.5-flash"*) and Gemini 1.5 Pro (*"gemini-1.5-pro"*), both recently updated in September 2024. Additionally, we compared lightweight Llama 3.1 with 8 billion parameters (*"Llama-3.1-8B-Instruct"* [10]) and a more recent Llama 3.3 with 70 billion parameters (*"Llama-3.3-70B-Instruct"* [11]).

We set the temperature setting to 0 to control randomness in order to increase reproducibility in generated outputs. Inspired by prompt recommendations made by Bsharat, Myrzakhan, and Shen (2023), we conducted iterative testing with different Zero-shot (i.e., when no examples are given) prompt variations on a sample dataset to fine-tune performance for our specific classification task. For our final predictions and comparisons, we utilized the following prompt for all selected LLMs:

> **Prompt:** *"You are an Expert APP RATER specializing in MENTAL HEALTH applications. Your task is to ACCURATELY CLASSIFY the review into one of the rating classes where 1 being the lowest and 5 being the highest. Only return a single number from Classes: [1, 2, 3, 4, 5] {input data}"*

This prompt formulation provided clear instructions to the model for consistent and accurate predictions across different instances. We implemented error handling for instances where GPT did not return a single number. The error rate for GPT-4 is 0.3% ($n = 502$), compared to lower GPT-3's error rate of 0.1% ($n = 191$). In our final comparison, the dataset comprises 200,471 entries.

3.4 Amazon Mechanical Turk

MTurk is a widely used platform for crowdsourcing tasks that require human intelligence (Sheehan 2018). On this platform, requesters recruit online workers globally to complete Human Intelligence Tasks (HITs). As a result, MTurk can be efficiently used to gather labels to create datasets for supervised learning models and other tasks, supporting the development of various ML applications. To address our third research question, we used MTurk to determine which rating source for a review was most appropriate and investigated the rationale behind their choice. To accomplish this task, we presented the MTurk workers with two questions for each review:

- Q1: *Two raters have assigned a star rating (1-5) to a review posted by a user of a mental health application. Based on your interpretation of the review displayed, which star rating is the most appropriate?*
- Q2: *What about the review led you to choose the rating? (Check all that apply).*

For Q2, two researchers manually inspected reviews with discrepancies between GPT and user ratings, focusing on common themes and specific issues mentioned. After identifying these themes, we generated a potential list of response options (see Figure 6). We then created a HIT for workers to review each discrepancy. The average completion time of the task was 23 seconds, and workers were compensated

[4]10% of the dataset was utilized due to computational limitations
[5]https://huggingface.co/google-bert/bert-base-uncased
[6]https://huggingface.co/google-bert/bert-large-uncased
[7]https://openai.com/
[8]https://www.llama.com/
[9]https://deepmind.google/
[10]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[11]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

$0.1 USD for each task completed (equivalent to $15 per hour). To ensure English proficiency and quality, workers were required to be based in the US and have a HIT approval rate of at least 90% to qualify for participation in the task. Thirty-four unique workers labeled a total of 1,728 discrepancies in predictions between GPT-4 predictions and reviewers' star ratings.

4 Results

4.1 Individual Review Classification: Supervised ML Models and LLMs

Table 1 presents an extensive comparison of evaluation metrics across various algorithms for supervised learning in the testing dataset (i.e., 20% of the original dataset). Evaluating the average performance across traditional ML, classifiers such as Neural Networks (NN) have the highest average F1-Score of 0.70, followed by Random Forest (RF) (F1-Score of 0.67), K-Nearest Neighbor (KNN) (F1-Score of 0.67), Support Vector Machine (SVM) (F1-Score of 0.66). The Decision Tree (DT) has the lowest performance with an F1-Score of 0.61. The embeddings MiniLM [12] with overall average F1-Score of 0.65 and BERT-Large[13] (F1-Score of 0.65) generally yielded better results than FastText [14] (F1-Score of 0.61), despite FastText's advantage of requiring significantly fewer computational resources. Evaluating the state-of-the-art ML models, all three deep learning classifiers outperformed traditional ML models in terms of average F1-Score and overall accuracy (i.e., 0.8). The RoBERTa classifier archives the highest F1-Score of 0.79. However, when examining individual rating classes, the random forest model showed better performance for classes 2, 3, and 4 in precision. Although these deep learning models achieve superior performance with extensive parameter fine-tuning, this process is computationally intensive.

In addition to the supervised ML models, the performance of various LLMs is reported in Table 2. The findings reveal that all selected full-scale LLM models (e.g. excluding Gemini 1.5 Flash and Llama 3.1-8B) outperform traditional supervised models in terms of average F1-Score. GPT-4 and Gemini 1.5 Pro both achieve an F1-Score of 0.76, which is comparable to the direct BERT (F1-Score of 0.78) and RoBERTa (F1-Score of 0.80) models. LLama 3.3 demonstrates competitive performance with a similar F1 score (0.75) and has the highest average accuracy of 0.75. During these experiments, GPT-3.5 Turbo (F1-Score of 0.71), Gemini 1.5 Flash (F1-Score of 0.65), Llama 3.1-8B (F1-Score of 0.61) were notable for their cost-effectiveness and optimized computing speed. Moreover, Table 2 shows varied performance across different classes. Specifically, classes 1 and class 5 have higher scores in Precision, Recall, and F1-Score across most models and embeddings, showing a stronger predictive performance for these classes. However, classes 2, 3, and 4 exhibit significantly lower scores for the same metrics, suggesting that all models struggle with accurately classifying instances from these categories. This pattern is consistent across the supervised ML models. Figure 2 displays a similar trend of variability in the performance of predicted ratings for the best-performing supervised learning model, BERT-Large with Neural Networks, alongside GPT-4. The majority of user ratings of 5 were misclassified as 4, and most of the user ratings of 1 were misclassified as 2 in both predictions. To further evaluate prediction on minority classes, we created two additional balanced datasets and presented results in Section 4.3.

Embedding	Classifier	P	R	F	Acc.
FasText	KNN	0.58	0.63	0.60	0.63
	SVM	0.63	0.73	0.66	0.73
	RF	0.73	0.69	0.61	0.69
	NN	0.67	0.59	0.62	0.59
	DT	0.56	0.55	0.56	0.55
MiniLM	KNN	0.65	0.72	0.67	0.72
	SVM	0.67	**0.74**	0.66	**0.74**
	RF	**0.75**	0.73	0.65	0.73
	NN	0.70	0.66	0.68	0.66
	DT	0.60	0.60	0.60	0.60
BERT-Large	KNN	0.59	0.66	0.62	0.66
	SVM	0.65	0.73	0.65	0.73
	RF	0.73	**0.74**	0.67	**0.74**
	NN	0.73	0.67	**0.70**	0.67
	DT	0.61	0.60	0.61	0.61
BERT (Base)		0.76	**0.80**	0.77	**0.80**
BERT (Large)		0.77	**0.80**	0.78	**0.80**
RoBERTa		**0.78**	**0.80**	**0.79**	**0.80**

Table 1: Performance comparison of traditional machine learning and deep learning models, measured by Precision (P), Recall (R), F1-Score (F), and Accuracy (Acc.). Bolded numbers indicate the highest values in each category.

4.2 Aggregated App Ratings: Performance Comparison

The integration of BERT-Large with a neural network classifier resulted in the top F1-Score (0.70) among the traditional supervised ML models. For self-supervised learning, the GPT-4 model achieved an F1-Score of 0.76 and the highest overall accuracy of 0.75, demonstrating superior performance. Thus, we selected these optimal performance models for comparative analysis at the aggregated app level, meaning we evaluated the models based on their performance, focusing on individual app instances. Figure 3 illustrates the distributions of precision and F1-Scores for each app for both models. GPT-4 demonstrates superior performance over the supervised learning model, with higher mean values in both distributions. These difference were statistically significant in Recall ($t = -6.66$, $p < .001$), Precision ($t = -4.28$, $p < .001$), and F1-Score ($t = -5.48$, $p < .001$). Therefore, this indicates that the GPT-4 model, on average, tends to predict with greater balanced accuracy.

[12] all-MiniLM-L12-v2:https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[13] BERT-Large-uncased: https://huggingface.co/google-bert/bert-large-uncased

[14] FastText: https://fasttext.cc/docs/en/supervised-tutorial.html

Model	Class	P	R	F	Acc.
GPT-3.5 Instruct	1	0.86	0.71	0.78	
	2	0.27	0.55	0.36	
	3	0.32	0.41	0.36	
	4	0.33	0.54	0.41	
	5	0.93	0.76	0.84	
	Avg	0.78	0.69	0.72	0.69
GPT-3.5 Turbo	1	0.90	0.56	0.69	
	2	0.23	**0.69**	0.35	
	3	0.30	0.34	0.32	
	4	0.32	0.52	0.40	
	5	0.92	0.76	0.83	
	Avg	0.78	0.67	0.71	0.67
GPT-4	1	0.87	0.72	0.79	
	2	0.28	0.60	**0.38**	
	3	0.30	0.41	0.34	
	4	**0.45**	0.39	0.42	
	5	0.92	**0.88**	**0.90**	
	Avg	**0.79**	**0.75**	**0.76**	**0.75**
Gemini 1.5 Flash	1	0.76	0.85	0.80	
	2	0.26	0.44	0.33	
	3	0.20	0.25	0.22	
	4	0.27	**0.62**	0.38	
	5	**0.95**	0.60	0.73	
	Avg	0.76	0.62	0.65	0.62
Gemini 1.5 Pro	1	0.82	0.83	**0.83**	
	2	**0.34**	0.37	0.36	
	3	0.34	0.45	**0.39**	
	4	0.38	0.56	**0.46**	
	5	0.93	0.81	0.86	
	Avg	**0.79**	0.74	**0.76**	0.74
Llama 3.1 (8B)	1	**0.93**	0.08	0.14	
	2	0.13	0.21	0.16	
	3	0.19	**0.60**	0.29	
	4	0.36	0.45	0.40	
	5	0.85	0.82	0.84	
	Avg	0.74	0.61	0.61	0.61
Llama 3.3 (70B)	1	0.79	**0.87**	**0.83**	
	2	0.29	0.40	0.34	
	3	**0.35**	0.24	0.28	
	4	0.39	0.54	0.45	
	5	0.92	0.82	0.87	
	Avg	0.77	**0.75**	0.75	**0.75**

Table 2: Comparison of different LLMs (self-supervised learning) with Precision (P), Recall (R), F1-Score (F) and Accuracy (Acc.) Bold numbers represent the highest values across each model for their respective evaluations within each class.

To address RQ2 and consider the influence on average numerical app ratings, we treat ratings as numerical values for each mental health app and determine the mean for both user ratings and model-generated predictions. Figure 4 demonstrates a strong correlation between the predicted ratings and the actual average app ratings, as indicated by the high R-squared values for both supervised and self-supervised models. GPT model's performance reflects marginally higher

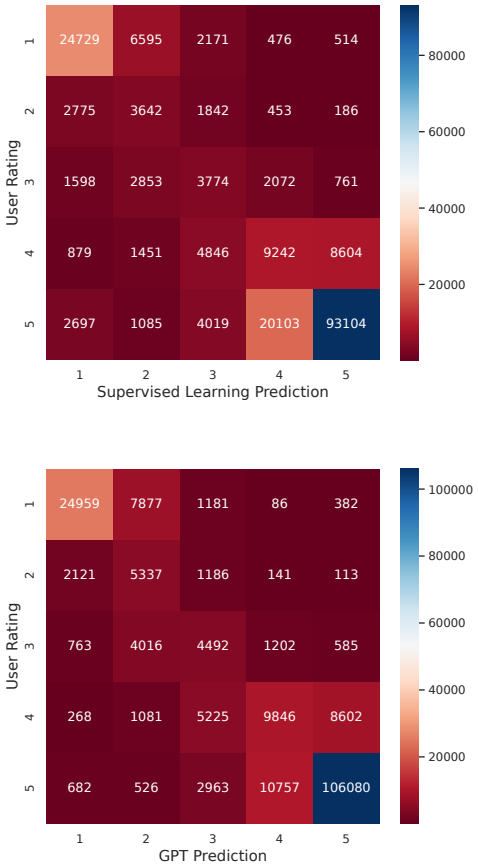

Figure 2: Confusion matrix of user ratings versus supervised learning: BERT-Large with Neural Networks (top) and GPT-4 (bottom).

R-squared and lower RMSE, with a slightly better accuracy in prediction. In examining the top 10 apps ranked by the users, GPT-4 average predictions successfully identified 8 of these apps (Gratitude, Moodpress, Intellect, Breath Ball, Healthy Minds, Dare, Mood Tracker, Daylio Journal). From the original top 10 rankings, the Voidpet Garden appeared at No.20 and Finch at No.14 by GPT-4. Conversely, GPT-4 included the Voice and Stop Panic apps within its top 10.

4.3 Performance on Balanced Datasets

Similar to Ishikawa, Yakoh, and Urushihara (2022); Ahmed and Ghabayen (2022); Shaikh et al. (2021), further evaluating the imbalance in our dataset would provide additional observations about model behavior. Thus, we conducted experiments on two additional crafted balanced datasets using the hybrid BERT-NN model, GPT-4, and the direct BERT-Large classifier. The results are presented in Table 3. When classes were uniformly distributed, we observed consistent performance patterns compared to the original unbalanced dataset for GPT-4 and hybrid BERT-NN. Conversely, the BERT classifier (Avg F1-0.77) outperforms GPT-4 (F1-0.59)

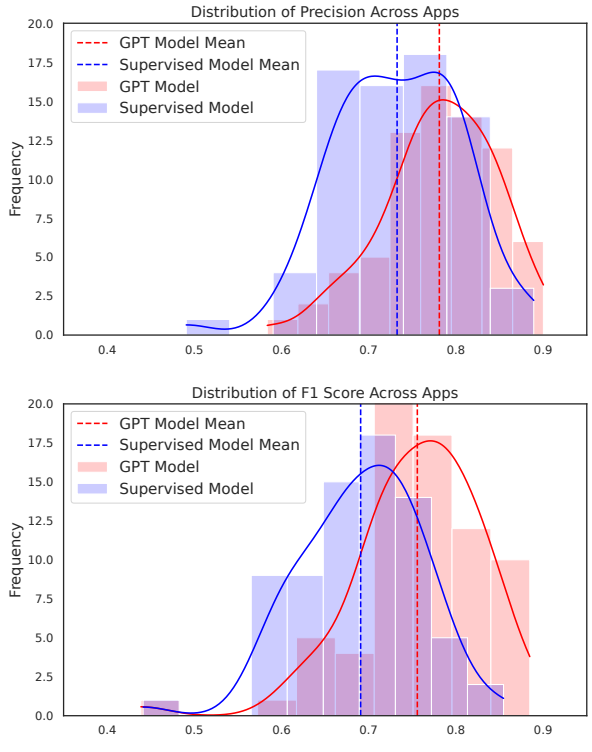

Figure 3: Distributions of Precision (top) and F1 Scores (bottom) across 73 apps. The Precision from GPT had higher values ($M = 0.78$, $SD = 0.06$) than the supervised model ($M = 0.73$, $SD = 0.07$). The F1-Score from GPT had higher values ($M = 0.76$, $SD = 0.07$) than the supervised model ($M = 0.69$, $SD = 0.07$).

	D1 (n=44k)			D2 (n=200k)		
Class	Hybrid	GPT-4	BERT	Hybrid	GPT-4	BERT
1	0.64	**0.71**	0.67	0.64	0.70	**0.84**
2	0.42	**0.54**	0.52	0.42	0.54	**0.81**
3	0.37	0.45	**0.46**	0.38	0.45	**0.77**
4	0.40	0.49	**0.51**	0.41	0.49	**0.67**
5	0.69	**0.76**	0.68	0.70	0.76	**0.77**
Avg	0.51	**0.59**	0.57	0.51	0.59	**0.77**

Table 3: Comparison of F1-scores in two balanced datasets. The hybrid model uses BERT-Large with neural networks, while BERT refers to the direct BERT-Large classifier. Bold numbers represent the highest values for each model within each dataset for each class.

in the second balanced dataset due to the increased number of samples for classes 2-4, allowing BERT to train more effectively (i.e. Class-2 F1-0.81)

4.4 MTurk Study Findings

Our study aimed to assess the alignment between LLM predictions and human evaluations. We identified 49,757 instances where GPT-4's predicted ratings diverged from user

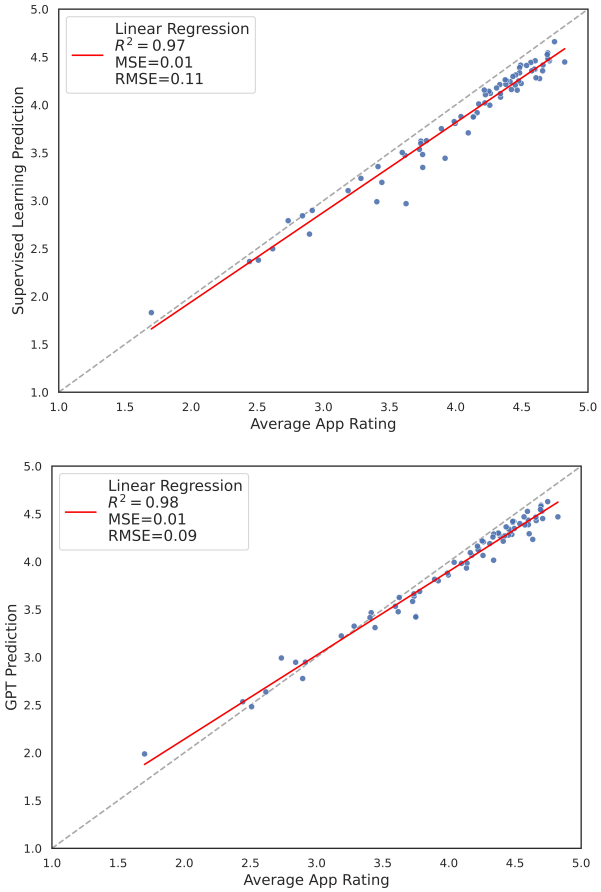

Figure 4: Correlation between actual user and predicted app ratings using supervised (top) and GPT-4 models (bottom).

ratings. To further investigate, we randomly selected 1,728 of these reviews and asked human annotators on MTurk to determine the most appropriate rating. The results indicate a substantial alignment with GPT ratings over the original user ratings. Specifically, MTurk workers agreed to approve 1050 (60.76%) reviews to match the ratings generated by GPT. 88 (5.09%) reviews were undecided, tagging them as "neither." Meanwhile, only 588 (34.03%) reviews were distinctly tagged with the original user rating. Moreover, responses in which individuals disagreed with the GPT rating were more likely to be evaluated as "neither agreeing nor disagreeing" in subsequent assessments (Figure 5). The majority of MTurk workers believe there is an overwhelming positive benefit. Figure 6 illustrates all rationales behind the choices made by MTurk workers participants.

4.5 Thematic Analysis of App Reviews

Overall, while GPT provides a balanced view. To address our fourth research question, we manually analyzed selected reviews (n=529) with discrepancies between GPT-generated ratings and actual app store ratings. Users expressing dissatisfaction, like in reviews stating, *"Says free, Asks for payment ASAP, Avoid, A scam"* or *"I've sent three emails, no response"*

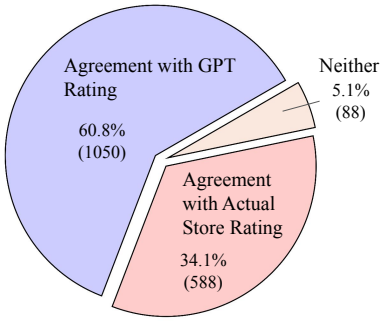

Figure 5: Ratings agreement distribution with human evaluation.

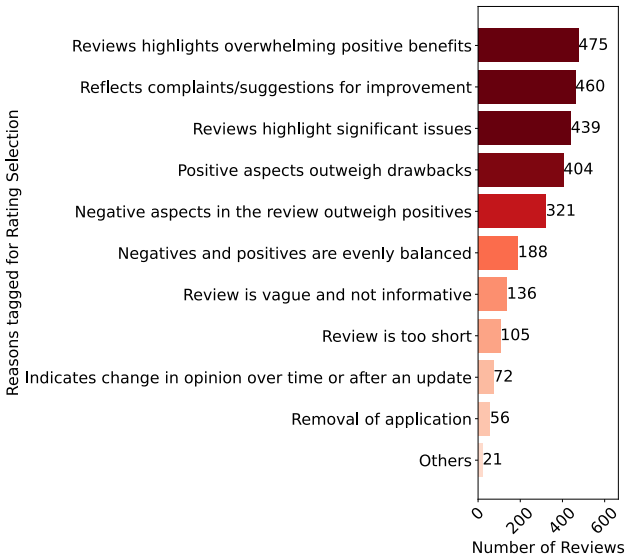

Figure 6: Distribution of reasons provided for rating selection by MTurk workers.

drove the lower GPT rating. Additionally, reviews containing unclear, contradictory, or irrelevant content also contributed to the rating misalignment.

GPT assigned a lower rating with decreases ranging from 1 to 4 stars in 296 cases (56%). This tendency may be due to GPT focusing on minor issues or critiques mentioned in the reviews, such as syncing problems, premium features, and functionality tweaks. While this feedback highlights the app's usefulness, it also points out limitations that prevent a perfect score, aligning generally with MTurk ratings. For example, one review stated, *"I use this every night and do look back at it whenever I'm asked how my mental health has been. Great for tracking the basics."* This captures basic sentiment and minor drawbacks that may not justify a 5-star rating. Similarly, another review noted, *"I like that one of the accessory options is hearing aids. I've never seen that in any game before...I LOVE that representation"* although the user desired additional features without cost. Most users still find the app highly useful, especially for managing anxiety or depression, and they value the free content even if some features require payment. They might also have lower expectations from an app compared to a therapist, which affects their valuation of the app despite its imperfections. Many users who gave 5-star ratings highlighted its positive features but experienced issues affecting their overall satisfaction. For instance, one user mentioned, *"This is a great app for using CBT techniques for self. Although it's not like talking to a real person, give it a chance."*

On the other hand, GPT gave higher ratings than the original in 232 cases (44%), with increments of only 1 or 2 stars. This suggests GPT tends to be more optimistic. For example, one review stated, *"Well, this app gives a new type of experience. I found this app to have a positive impact on my life."* While the user rated the app 4 stars, GPT assigned it 5 stars, which aligns with MTurkers' evaluations. This indicates that users who leave highly positive reviews may rate conservatively (4 stars), while GPT more optimistic approach leads to higher ratings. In cases where MTurkers aligned with GPT with more moderate 2-star ratings compared to the original 1-star reviews, users expressed mixed feedback. While some appreciated features like meditation, many were frustrated by technical issues, data loss, and inadequate support. Common sources of disappointment included changes in the payment structure, with previously free content now locked behind paywalls, repetitive meditations, limited accessibility, and a lack of emotional logging options. Additional frustrations included streak resets and missing playlist controls. For reviews with original 3-star ratings, GPT often interpreted them as more positive, assigning 4- or 5-star ratings. In these cases (n = 12), users highlighted the app's insightful features for self-understanding, mood tracking, and goal-setting. However, they also noted drawbacks such as spelling errors, unclear session guidelines, and the need for additional features like voice recording and customized affirmations. Users acknowledged the clean UX design but expressed a desire for greater affordability and expanded health insurance coverage.

5 Discussion

Community participation has shown increasing effectiveness in supporting individuals with mental health disorders (Saha and Sharma 2020). Our work explores new approaches and attempts to assist users in selecting mental health apps based on ratings. To achieve this, we framed four key RQs and discussed how we answered each of our RQs based on our analysis and the results obtained.

To address RQ1, our findings indicate that both LLM and supervised machine learning techniques can effectively classify user feedback. As discussed in Section 4.1, the overall F1-Score of 0.76 achieved by the GPT-4 model and Gemini 1.5 Pro is particularly noteworthy due to the inherent advantages of self-supervised learning methods, which require significantly less human effort than supervised approaches. Similarly, Llama 3.3 also achieves comparable performance. On the other hand, with smaller-sized LLMs, Gemini 1.5 Flash delivers low latency and cost optimization, though it demonstrated lower performance compared to Gemini 1.5 Pro in our experiments and other benchmark tasks (Team et al.

2024). The *"out-of-the-box"* performance of these LLMs demonstrates the potential for effective text classification tasks without the need for extensive labeled datasets. This capability reduces the labor-intensive process of manual labeling (Yang, Cao, and Fan 2024). Meanwhile, the combination of the BERT-Large and Neural Network achieved an optimal balance of precision and recall. We found that GPT-4 and GPT-3.5 Instruct were superior to GPT-3.5 Turbo. However, GPT-3.5 Instruct is a cost-effective choice for resource-limited projects. Despite the high accuracy, less than 1% of GPT requests returned unexpected results. These included bracketed numbers, lengthy explanations, or error messages. Furthermore, we encountered challenges with rate limitations and the extensive time required to process over 200k requests.

To address RQ2, we aggregated GPT-4 prediction at the app level and found a significant correlation between GPT-generated average ratings and original user evaluations, closely aligning with App Store overall rankings, as discussed in Section 4.2. This suggests that LLMs are potentially suitable for assessing app quality via large text corpora.

Investigating RQ3 with a human study, we found GPT predictions align closely with MTurk evaluations, which could provide a reliable measurement that captures user sentiments effectively and with less subjectivity. However, GPT tends to assign lower scores to reviews affected by bots, spam, or mixed feedback, where users share both positive and negative comments or leave random text and emojis. Such reviews, despite originally receiving higher star ratings, often raise issues like technical glitches or frequent changes in app permissions. The alignment between LLM classifications and MTurk feedback across diverse review sentiments demonstrates potential applicability to broader platforms, including user-generated reviews on Google Maps and Amazon product reviews. This highlights the potential of LLMs in analyzing user reviews and underscores the necessity for enhancements in scalability and performance.

Answering our RQ4, rating discrepancies in mental health apps arise primarily due to differences in how users and GPT prioritize praise and criticism within reviews. As mentioned in Section 4.5, users often balance their high ratings with critical feedback, such as expressing satisfaction with an app's features while simultaneously wishing for more free options. In such cases, GPT tends to focus on the critiques, leading to lower ratings than users initially provided. This discrepancy occurs especially when reviews contain mixed sentiments, where the overall tone may be positive but includes notable drawbacks. GPT tends to emphasize these negative aspects more heavily, resulting in divergent ratings. Furthermore, when reviews highlight minor flaws alongside otherwise positive feedback, GPT often overweight the critical points, diverging from user ratings that may have downplayed these issues in their final assessment.

Finally, our dataset contribution will be highly valuable to the community, similar to (Singh et al. 2022), in response to the need for large-scale, well-labeled mental health datasets with fast reproducible methods to facilitate their heuristic growth. Our work highlights the broader implications of using AI-driven systems, like GPT, to evaluate mental health apps, where accurate interpretation of user feedback is crucial for app adoption and trust. One could argue that AI might be more objective than individuals in rating systems, especially if rating scores are crucial for recommendations. Our work has implications for understanding how AI systems interpret user feedback. Specifically,as it suggests how AI-mediated rating systems might shape online communities' perceptions and decisions regarding widely used digital apps, platforms, and services.

5.1 Limitation & Future Work

Our exploration into various Zero-shot prompting designs revealed that shorter and more straightforward prompts yield greater effectiveness compared to those that are longer and more intricate. While we maintained a consistent optimal prompt across all LLM models, future work could explore the use of different GPT prompting (e.g. Few-Shot) for further improvement.

Since the majority of user reviews (91.84%) in our dataset are in English, we included only English reviews to maintain controlled conditions and ensure a fair comparison. However, future work could explore incorporating multilingual reviews using multilingual supported BERT model and LLMs to gain deeper insights into each mental health app by language. Additionally, user ratings over time could be examined using a longitudinal approach, analyzing the running average of each app rating and comparing it with model effectiveness.

In addition, we acknowledge that LLMs have been pretrained on vast amounts of data, which may include content similar to the review data collected in this study. Consequently, we recognize that fine-tuned supervised models could potentially outperform LLMs in certain scenarios, particularly when substantial resources are available for fine-tuning. Nevertheless, the primary objective of this study is to demonstrate the feasibility and practicality of utilizing *"out-of-the-box"* LLMs for review classification and app rating analysis. Our findings underscore the competitive performance of LLMs and their potential to streamline classification tasks in resource-constrained environments. Moreover, integrating LLM-generated explanations for each prediction could improve interpretability and transparency. While this study focuses on LLMs in the context of mental health app reviews, its implications extend to other domains, such as product, service, and platform reviews that lack explicit rating scales, relying solely on textual content for evaluation. Additionally, as mentioned in the survey paper (Hadi et al. 2023), this approach could also be applied to news articles, sports, and entertainment applications.

6 Conclusion

Our research demonstrates the effectiveness of LLMs in multi-class classification across large datasets. The GPT-4, Gemini 1.5 Pro, and Llama 3.3 had noteworthy performance in categorizing user reviews into ratings, with a comparable performance of state-of-the-art supervised learning techniques. Our results highlight the efficiency of LLMs in processing and understanding user feedback, eliminating the complexities of fine-tuning and reward dependencies. We provide recommendations for future mental health app development, emphasizing improved accessibility, enhanced

customer support, and other key features. Furthermore, we discovered that most crowdsourced workers concurred more with the GPT-4-derived ratings than the original user ratings. We foresee such methodologies contributing to the development of app recommendation systems that improve user decision-making. The potential for applying these approaches extends to various review contexts, such as those on entertainment and online retail. To encourage further research in this area, we have made our MHARD dataset available at https://github.com/Sensify-Lab/MHARD.git

Acknowledgments

This work originated from a hackathon event run by the Data Science Institute (DSI) and AI Center of Excellence (AICoE) at the University of Delaware. Additionally, we are grateful to Dr. Salvatore Giorgi for his valuable feedback for improving this manuscript, which was made possible through the ICWSM 2025 Mentoring Program. Furthermore, Moath Erqsous acknowledges the financial support provided by the University of Bisha, and Fatimah Mohammed Alhassan acknowledges the financial support from the Cultural Mission of the Royal Embassy of Saudi Arabia (SACM).

References

Ahmed, B. H.; and Ghabayen, A. S. 2022. Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, 13(7): 3423–3432.

Allahbakhsh, M.; Benatallah, B.; Ignjatovic, A.; Motahari-Nezhad, H. R.; Bertino, E.; and Dustdar, S. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing*, 17(2): 76–81.

Alqahtani, F.; and Orji, R. 2020. Insights from user reviews to improve mental health apps. *Health informatics journal*, 26(3): 2042–2066.

Alzetta, C.; Dell'Orletta, F.; Miaschi, A.; Prat, E.; and Venturi, G. 2024. Tell me how you write and I'll tell you what you read: a study on the writing style of book reviews. *Journal of Documentation*, 80(1): 180–202.

Arslan, Y.; Allix, K.; Veiber, L.; Lothritz, C.; Bissyandé, T. F.; Klein, J.; and Goujon, A. 2021. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*, 260–268.

Aryana, B.; Brewster, L.; and Nocera, J. A. 2019. Design for mobile mental health: an exploratory review. *Health and Technology*, 9: 401–424.

Badesha, K.; Wilde, S.; and Dawson, D. L. 2022. Mental health mobile app use to manage psychological difficulties: An umbrella review. *Mental Health Review Journal*, 27(3): 241–280.

Bakker, D.; Kazantzis, N.; Rickwood, D.; Rickard, N.; et al. 2016. Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR mental health*, 3(1): e4984.

Balaskas, A.; Schueller, S. M.; Cox, A. L.; and Doherty, G. 2022. Understanding users' perspectives on mobile apps for anxiety management. *Frontiers in digital health*, 4: 854263.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

Bond, R. R.; Mulvenna, M. D.; Potts, C.; O'Neill, S.; Ennis, E.; and Torous, J. 2023. Digital transformation of mental health services. *Npj Mental Health Research*, 2(1): 13.

Borromeo, R. M.; and Toyama, M. 2015. Automatic vs. crowdsourced sentiment analysis. In *Proceedings of the 19th International Database Engineering & Applications Symposium*, 90–95.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Brin, D.; Sorin, V.; Konen, E.; Nadkarni, G.; Glicksberg, B. S.; and Klang, E. 2023. How Large Language Models Perform on the United States Medical Licensing Examination: A Systematic Review. *medRxiv*, 2023–09.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bsharat, S. M.; Myrzakhan, A.; and Shen, Z. 2023. Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. *arXiv preprint arXiv:2312.16171*.

Chandrasekaran, A.; Bielicke, L.; Shah, D.; Janakiraman, H.; and Mauriello, M. L. 2025. "I spent 14 hours debugging just one assignment": Toward Computer-Mediated Personal Informatics for Computer Science Student Mental Health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19.

Checco, A.; Bracciale, L.; Loreti, P.; Pinfield, S.; and Bianchi, G. 2021. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1): 1–11.

Chicco, D.; and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21: 1–13.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Databricks. 2023. Automated Analysis of Product Reviews Using Large Language Models (LLMs). https://www.databricks.com/blog/automated-analysis-product-reviews-using-large-language-models-llms. Accessed: 2024-03-23.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Drutsa, A.; Fedorova, V.; Ustalov, D.; Megorskaya, O.; Zerminova, E.; and Baidakova, D. 2020. Practice of efficient data collection via crowdsourcing. *Proceedings of the 13th International Conference on Web Search and Data Mining*.

Etemadi, M.; Abkenar, S. B.; Ahmadzadeh, A.; Kashani, M. H.; Asghari, P.; Akbari, M.; and Mahdipour, E. 2023. A systematic review of healthcare recommender systems:

Open issues, challenges, and techniques. *Expert Systems with Applications*, 213: 118823.

Funnell, E. L.; Spadaro, B.; Martin-Key, N.; Metcalfe, T.; and Bahn, S. 2022. mHealth solutions for mental health screening and diagnosis: a review of app user perspectives using sentiment and thematic analysis. *Frontiers in psychiatry*, 13: 857304.

Grainger, R.; Devan, H.; Sangelaji, B.; and Hay-Smith, J. 2020. Issues in reporting of systematic review methods in health app-focused reviews: a scoping review. *Health Informatics Journal*, 26(4): 2930–2945.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Haque, M. R.; and Rubya, S. 2022. " For an app supposed to make its users feel better, it sure is a joke"-an analysis of user reviews of mobile mental health applications. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–29.

Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; and Scholkopf, B. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4): 18–28.

Howard, J.; and Ruder, S. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hu, N.; Pavlou, P. A.; and Zhang, J. J. 2009. Why do online product reviews have a J-shaped distribution? Overcoming biases in online word-of-mouth communication. *Communications of the ACM*, 52(10): 144–147.

Ishikawa, T.; Yakoh, T.; and Urushihara, H. 2022. An NLP-inspired data augmentation method for adverse event prediction using an imbalanced healthcare dataset. *IEEE Access*, 10: 81166–81176.

Jo, E.; Kouaho, W.-J.; Schueller, S. M.; and Epstein, D. A. 2023. Exploring User Perspectives of and Ethical Experiences With Teletherapy Apps: Qualitative Analysis of User Reviews. *JMIR mental health*, 10: e49684.

Kjell, O. N.; Kjell, K.; and Schwartz, H. A. 2024. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333: 115667.

Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; and Brown, D. 2019. Text classification algorithms: A survey. *Information*, 10(4): 150.

Kuo, R.; and Li, S.-S. 2023. Applying particle swarm optimization algorithm-based collaborative filtering recommender system considering rating and review. *Applied Soft Computing*, 135: 110038.

Liu, J.; Liu, C.; Zhou, P.; Ye, Q.; Chong, D.; Zhou, K.; Xie, Y.; Cao, Y.; Wang, S.; You, C.; et al. 2023. Llmrec: Benchmarking large language models on recommendation task. *arXiv preprint arXiv:2308.12241*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacLean, D. L.; and Heer, J. 2013. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the american medical informatics association*, 20(6): 1120–1127.

Malik, T.; Ambrose, A. J.; and Sinha, C. 2022. Evaluating user feedback for an artificial intelligence–enabled, cognitive behavioral therapy–based mental health app (Wysa): qualitative thematic analysis. *JMIR Human Factors*, 9(2): e35668.

Miner, A. S.; Milstein, A.; Schueller, S.; Hegde, R.; Mangurian, C.; and Linos, E. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5): 619–625.

Myagmar, B.; and Li, J. 2019. Cross-domain sentiment classification with bidirectional contextualized transformer language models. *IEEE Access*, 7: 163219–163230.

Naseer, M.; Asvial, M.; and Sari, R. F. 2021. An empirical comparison of bert, roberta, and electra for fact verification. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 241–246. IEEE.

Nguyen, V. C.; Jain, M.; Chauhan, A.; Soled, H. J.; Lesmes, S. A.; Li, Z.; Birnbaum, M. L.; Tang, S. X.; Kumar, S.; and De Choudhury, M. 2024. Supporters and Skeptics: LLM-based Analysis of Engagement with Mental Health (Mis) Information Content on Video-sharing Platforms. *arXiv preprint arXiv:2407.02662*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peterson, L. E. 2009. K-nearest neighbor. *Scholarpedia*, 4(2): 1883.

Pistilli, G. 2022. What lies behind AGI: ethical concerns related to LLMs.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ricci, F.; Rokach, L.; and Shapira, B. 2021. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, 1–35.

Rodrigues, F.; and Pereira, F. C. 2018. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Saha, K.; and Sharma, A. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 590–601.

Schueller, S. M.; Neary, M.; O'Loughlin, K.; and Adkins, E. C. 2018. Discovery of and interest in health apps among those with mental health needs: survey and focus group study. *Journal of medical Internet research*, 20(6): e10141.

Shaikh, S.; Daudpota, S. M.; Imran, A. S.; and Kastrati, Z. 2021. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. *Applied Sciences*, 11(2): 869.

Sheehan, K. B. 2018. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs*, 85(1): 140–156.

Sheng, V. S.; Provost, F. J.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. *Organizations & Markets eJournal*.

Singh, A. K.; Arora, U.; Shrivastava, S.; Singh, A.; Shah, R. R.; Kumaraguru, P.; et al. 2022. Twitter-stmhd: An extensive user-level database of multiple mental health disorders. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1182–1191.

Skondras, P.; Zervas, P.; and Tzimas, G. 2023. Generating Synthetic Resume Data with Large Language Models for Enhanced Job Description Classification. *Future Internet*, 15(11): 363.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.

Song, Y.-Y.; and Ying, L. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2): 130.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the twenty-sixth AAAI conference on artificial intelligence*.

Subramanian, V. 2018. *Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch*. Packt Publishing Ltd.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, 194–206. Springer.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Thabtah, F.; Hammoud, S.; Kamalov, F.; and Gonsalves, A. 2020. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513: 429–441.

Tian, Y.; Zhang, W.; Duan, L.; McDonald, W.; and Osgood, N. D. 2023. Comparison of pretrained transformer-based models for influenza and covid-19 detection using social media text data in saskatchewan, canada. *Frontiers in Digital Health*, 5.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33: 5776–5788.

Wisniewski, H.; Liu, G.; Henson, P.; Vaidyam, A.; Hajratalli, N. K.; Onnela, J.-P.; and Torous, J. 2019. Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. *BMJ Ment Health*, 22(1): 4–9.

Xi, Z.; Rui, Z.; and Tao, G. 2023. Safety and Ethical Concerns of Large Language Models. In *China National Conference on Chinese Computational Linguistics*.

Yacouby, R.; and Axman, D. 2020. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 79–91.

Yang, C.; Cao, B.; and Fan, J. 2024. TeC: A Novel Method for Text Clustering with Large Language Models Guidance and Weakly-Supervised Contrastive Learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1702–1712.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1): 194.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhao, Z.; Zhang, Z.; and Hopfgartner, F. 2021. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, 500–507.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; Peng, H.; Li, J.; Wu, J.; Liu, Z.; Xie, P.; Xiong, C.; Pei, J.; Yu, P. S.; and Sun, L. 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv:2302.09419.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA

(e) Did you describe the limitations of your work? Yes

(f) Did you discuss any potential negative societal impacts of your work? NA

(g) Did you discuss any potential misuse of your work? NA

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We have made the dataset available at https://github.com/Sensify-Lab/MHARD.git

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Ethical Statement

Utilizing LLM models in sensitive areas like mental health app evaluations raises ethical challenges, such as bias mitigation and transparent decision-making. As discussed in Section 3.1, our dataset is fully anonymized and does not contain sensitive information. Additionally, memory storage for the LLM models was disabled to prevent any data from being stored on external servers. Although research has raised concerns about the potential for harmful content to be generated (Xi, Rui, and Tao 2023; Pistilli 2022), we did not encounter any harmful content, as we instructed the LLM to return only rating scores. Ensuring accurate and fair app ratings is crucial to avoid misleading users who rely on these tools for recommendations. Furthermore, no personally identifiable information was collected from MTurk workers during our human study.