A Survey of Large Language Models in Mental Health Disorder Detection on Social Media

Zhuohan Ge[1], Nicole Hu[2*], Darian Li[1], Yubo Wang[3], Shihao Qi[1], Yuming Xu[1], Han Shi[3], Jason Zhang[1]
[1]*The Hong Kong Polytechnic University*
[2]*The Chinese University of Hong Kong*
[3]*Hong Kong University of Science and Technology*

***Abstract*—The detection and intervention of mental health issues represent a critical global research focus, and social media data has been recognized as an important resource for mental health research. However, how to utilize Large Language Models (LLMs) for mental health problem detection on social media poses significant challenges. Hence, this paper aims to explore the potential of LLM applications in social media data analysis, focusing not only on the most common psychological disorders such as depression and anxiety but also incorporating psychotic disorders and externalizing disorders, summarizing the application methods of LLM from different dimensions, such as text data analysis and detection of mental disorders, and revealing the major challenges and shortcomings of current research. In addition, the paper provides an overview of popular datasets, and evaluation metrics. The survey in this paper provides a comprehensive frame of reference for researchers in the field of mental health, while demonstrating the great potential of LLMs in mental health detection to facilitate the further application of LLMs in future mental health interventions.**

***Index Terms*—LLM, mental disorders, social media, depression, suicide risk, schizophrenia, externalizing disorders.**

I. Introduction

Globally, half of all individuals will experience or have experienced a mental health disorder [1], and mental health issues have become a significant challenge affecting the well-being of both societies and individuals. According to the World Health Organization (WHO) [2], in 2019, nearly 1 billion people worldwide (including 14% of the world's adolescents) were affected by a mental disorder, representing 12.5% of the global population [3]. Mental disorders impact all aspects of life, influencing learning, productivity, and relationships with family and friends. It is estimated that about 12 billion workdays are lost each year due to depression and anxiety disorders [4], resulting in a loss of up to $1 trillion annually, a figure projected to reach $16 trillion by 2030 [5].

In recent years, the rapid development of the internet has made online social media an essential platform for detecting mental disorders on a global scale [6]–[8]. User-generated content on social media not only reflects users' daily lives but also their emotional fluctuations and psychological states. This content includes activities such as posting updates, sharing personal status, complaining about various issues, and expressing emotions. Furthermore, users can interact with others through retweets, comments, likes, and other actions, enabling

* Corresponding Author: hulan@link.cuhk.edu.hk

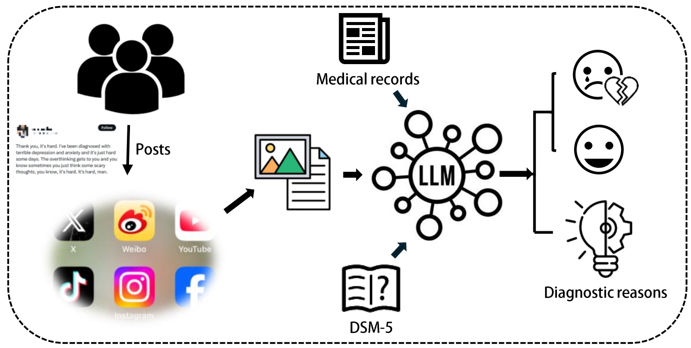


Fig. 1. The Framework for Detecting Mental Disorders on Social Media Using LLMs.

them to express their personal feelings and thoughts more comprehensively [9]. This generates a vast amount of real-time social information of significant value [10]. According to [11]–[13], the number of active users on social media platforms is immense, and these platforms have become the most important venues for social interaction and information dissemination. Hence, exploring how to effectively utilize the social media data for mental health monitoring and prediction has become a key focus for future research.

With the rapid development of natural language processing, Large Language Models (LLMs) [14]–[17] have become mainstream tools for tackling complex language understanding tasks. These models have shown significant advantages in processing large-scale textual data, offering a deep understanding of syntactic, semantic, and contextual information in language [18], [19], making them suitable for detecting mental disorders on social media [20], [21]. Figure 1 illustrates the general workflow of LLMs for mental disorders detection on social media. Users often post emotionally charged content on social media platforms. By analyzing this content, LLMs can not only identify symptoms of mental disorders but also diagnose potential diseases based on these symptoms and provide reasonable explanations.

Several surveys [22]–[36] have explored various approaches and challenges in detecting multiple mental health problems. But [22]–[28] are out-of-date considering the rapid development of the LLM research. Some other surveys [29]–[31] focus on LLM applications nearly on single disease, which makes them fail to generalize, as different diseases exhibit distinct characteristics. For example, schizophrenia detection focuses on semantic coherence rather than emotional expression, but depression is the opposite. Consequently, models designed for depression detection is naturally more oriented to word level emotional and linguistic style analyzation, but not paragraph level coherence [37], [38]. Also, many current surveys [32]–[36] focus solely on the application of LLMs in the mental health field without specializing in social media and its unique structure, including interactivity, dynamics and real-time nature, where individuals frequently express mental health disorders [7]. Additionally, social media data is easily accessible, offers real-time insights, and spans long periods. Neglecting its potential limits the understanding of how LLMs can be effectively applied to these platforms.

Specifically, this survey offers a comprehensive overview of the current state of the art in the application of LLMs for detecting mental health disorders on social media. Our contributions are summarized as follows:

- We discuss a wide range of mental health disorders, elaborating on their features, differences, and commonalities.
- We provide an overview of current mainstream social media platforms, focusing on their reliability and advantages in mental disorder detection.
- We delineate some specific applications of LLMs on social media and their key research methods, models, datasets, etc.
- We discuss the limitations of current research, challenges, and future research directions.
- We present popular social media benchmark datasets, highlighting their characteristics and applicable tasks.

Organization. We first introduce the background and preliminary concepts of mental disorders, social media platforms, and LLMs (SEC.II); then we review research on using LLMs for mental disorder detection on social media and discuss future directions in areas where research is lacking (SEC.III). Furthermore, we explore popular datasets and evaluation metrics (SEC.IV). Finally, we outline the limitations, challenges, and future directions of the research (SEC.V).

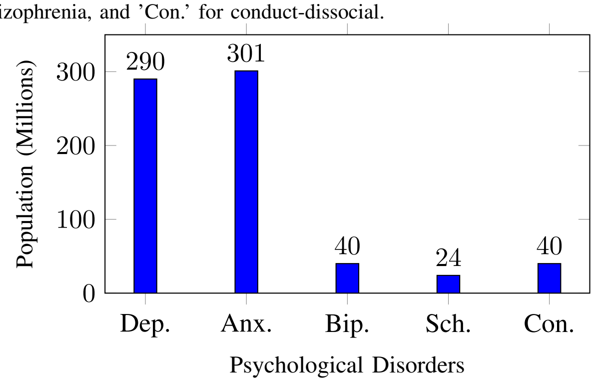
II. PRELIMINARY

Understanding and addressing mental health disorders has become a crucial area of research due to their profound impact on both individuals and society. The rapid growth of digital platforms, particularly social media, has created an unprecedented opportunity to leverage user-generated data for the early detection and intervention of mental health problems. To fully harness this potential, it is essential to examine the nature of mental health disorders, the diverse types of data available on social media, and the advanced tools that enable such analysis. We begin by providing an overview of common mental disorders and various social media data types, and then explain why LLMs are particularly advantageous in the mental health field.

A. Main Psychological Disorders

The WHO provides the affected population for some of the main disorders in 2019 [3], as shown in Figure 2, a

Fig. 2. Global population with mental disorders (2019) [3], where 'Dep.' stands for depression, 'Anx.' for anxiety, 'Bip.' for bipolar, 'Sch.' for schizophrenia, and 'Con.' for conduct-dissocial.



siginificant number of people is suffer from mental disorder. These disorders not only affect an individual's mood and behavior but also significantly impact daily life and social functioning, and can even lead to self-harm and suicidal tendencies, emphasizing the importance of early detection and intervention [39], [40].

Based on the broad diagnostic categories defined in previous studies [41], [42], mental disorders can be generally classified into internalizing disorders (including anxiety disorder, depression, disruptive mood dysregulation disorder, and post-traumatic stress disorder) and externalizing disorders (including attention-deficit disorder, oppositional defiant disorder, and conduct disorder).

Furthermore, an analysis of the classification changes of psychotic disorders [43], identified substantial heterogeneity in the symptoms and disease course of schizophrenia and other mental disorders. Therefore, while maintaining the aforementioned classification scheme, we additionally incorporate Psychotic (schizophrenia) as a separate category, ultimately classifying mental disorders into three distinct groups: Emotional Internalization, Psychotic, and Externalizing Disorder. Figure 3 presents the specific symptoms and categorizations of the main diseases of these mental disorders.

Next, we will elaborate on the differences between these three types of mental disorders:

- **Emotional Internalization Disorders** are characterized by internal distress, primarily manifested through persistent negative emotions, excessive worry, and a sense of hopelessness toward life. Unlike psychotic disorders, where individuals may lose touch with reality through hallucinations or delusions and experience a disruption in their core sense of self, individuals with emotional internalization disorders remain cognitively aware of their struggles, with their distress primarily centered on internal suffering. Additionally, in contrast to externalizing disorders, which involve outwardly disruptive behaviors, emotional internalization is defined by self-directed suffering rather than aggression or defiance against social norms. However, comorbidity between these disorders may occur in certain cases [44].
- **Psychotic Disorders** are characterized by perceptual disturbances and a fundamental break from reality, often manifest-

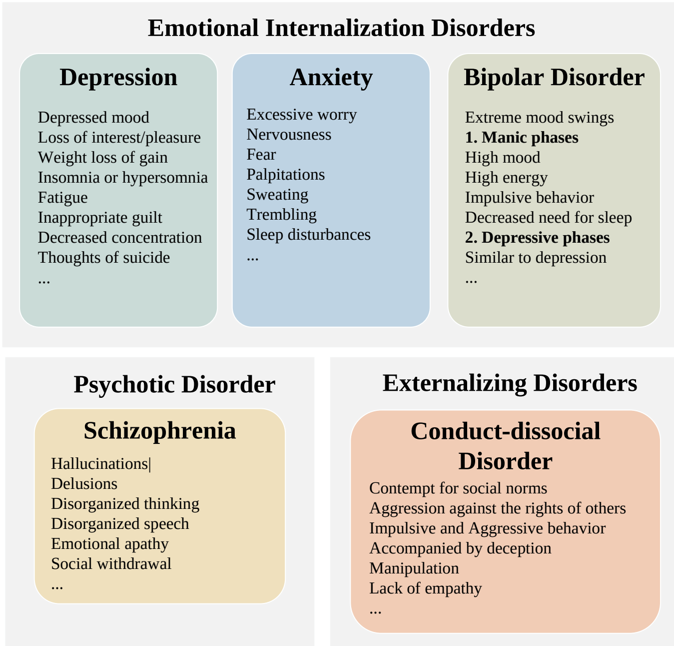


Fig. 3. Classification and symptom introduction of psychological disorders.

ing as hallucinations, delusions, and disorganized thinking, which lead to severe impairments in cognition and social functioning. Unlike emotional internalization disorders, where distress remains grounded in reality, schizophrenia and related conditions distort an individual's perception of the world, resulting in profound cognitive disruptions and difficulties in maintaining logical thought processes. Furthermore, in contrast to externalizing disorders, which primarily involve behavioral dysregulation, psychotic symptoms do not necessarily lead to impulsive or antisocial behaviors; rather, they cause significant impairments in perception, communication, and daily functioning.

- **Externalizing Disorders** are primarily characterized by impulsivity, aggression, and rule-breaking behaviors. Conditions such as conduct disorder and attention-deficit/hyperactivity disorder (ADHD) often lead individuals to engage in disruptive actions, typically manifesting as hostility, defiance, or difficulties in self-regulation. Unlike individuals with emotional internalization disorders, who may withdraw from social interactions due to distress, those with externalizing disorders tend to engage in overtly disruptive behaviors, which can lead to conflicts with peers or authority figures. Similarly, while psychotic disorders involve distorted thoughts and perceptions, externalizing disorders are primarily indicative of dysfunctional behavioral control rather than detachment from reality.

Figure 4 illustrates how these three types of disorders manifest in social media data.

B. Social Media and Data Types

With the development of the Internet, a wide variety of social media platforms have emerged, each with distinct functions and user-generated content, meaning that each type of social media data has its own unique characteristics. Table I presents the mainstream social media platforms [11].

TABLE I
NUMBER OF MONTHLY ACTIVE USERS ON SOCIAL MEDIA PLATFORMS IN 2023/2024 [11].

Platform	Main Function	Active Users (Millions)
X/Twitter	Short text sharing	421
Facebook	Social networking service	3080
Instagram	Image and video sharing	2250
Reddit	Discussion forums, subreddits	850
Weibo	Similar to Twitter	587
YouTube	Video sharing	2700
TikTok	Short video sharing	1587

In addition, social media data includes not only text but also multimodal data, such as images and videos. These diverse data types provide rich resources for mental health detection. The following is a detailed introduction to the types of data found in mainstream social media:

- **Text data**: On social media, text data typically appears in the form of tweets, status updates, comments, and more. For example, users may post tweets on Twitter to express emotions (e.g., "I feel helpless") or seek support on Reddit (e.g., "How can I deal with anxiety?"). These texts are often unstructured, with diverse grammar, and may include slang, abbreviations (e.g., "idk" for "I don't know"), or emoticons (e.g., ":-(" to express negative emotions).
- **Audio data**: Audio data on social media primarily includes voice in videos, background music in shared content, etc. For instance, on Instagram and TikTok, users may select cheerful music to express joy in shared content or use sad music to convey negative emotions.
- **Image data**: Image data on social media mainly includes selfies and user-posted pictures. For example, on Instagram, users may post a blurry or low-light picture with text describing a low mood, which could indicate depression. Specific image features, such as color (e.g., gray tones may suggest negative emotions) and background elements (e.g., isolated scenes), also contribute to emotional expression.
- **Multimodal data**: Multimodal data on social media combines text, voice, images, and other types of information. For example, on TikTok, YouTube, and Facebook, users may post a blurry or low-light picture with text describing a low mood, which may suggest depression. In videos, users may express emotions simultaneously through facial expressions, voice, and subtitles. This combination provides a richer source of information for mental health detection. For example, a short video posted by a user may contain a dubbing text (e.g., "I am tired"), a low-pitched voice, and a dark background scene.

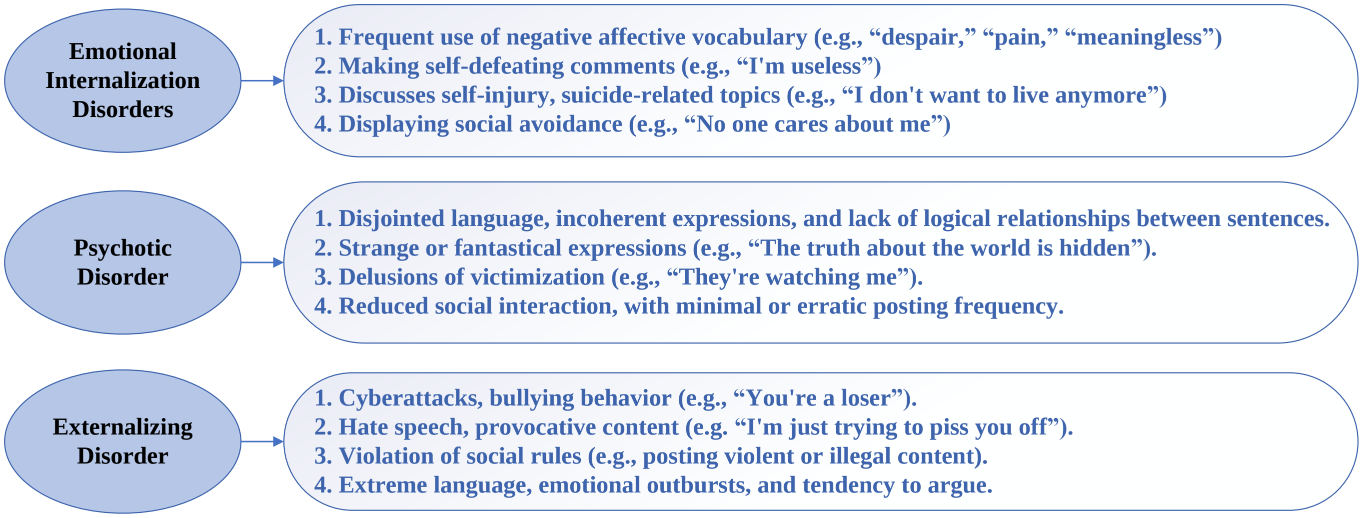


Fig. 4. Manifestations of Mental Disorders on Social Media.

C. *Large Language Models for Mental Disorders*

LLMs are trained on large-scale textual data to learn language patterns, syntax, semantics, and world knowledge, enabling them to perform natural language understanding, generation, and logical reasoning tasks [17]. Mainstream LLMs, such as the GPT series [45]–[47], LLaMA [48], [49], and DeepSeek [50]–[52], are based on the transformer architecture [53], which follows an encoder-decoder structure that efficiently captures long-range dependencies in sequence data, significantly improving the computational efficiency and expressive power of the models. These models have been successfully applied to a wide variety of tasks, including natural language processing [54]–[56] and computer vision [57]–[59].

LLMs are highly effective for processing large-scale text data from diverse sources such as social media, blogs, and forums. They utilize self-attention mechanisms to capture long-distance dependencies and subtle nuances in language, enabling them to perform complex tasks like text categorization, sentiment analysis, and psychological assessments. Through pre-training, fine-tuning, and reinforcement learning from human feedback (RLHF) [60], LLMs demonstrate impressive cross-task generalization, making them adaptable to a wide range of applications. Furthermore, their few-shot and zero-shot learning capabilities allow them to make inferences from minimal training examples, enhancing their flexibility for mental health prediction tasks. Some LLMs even extend their capabilities to multimodal data [61]–[63], processing text, images, audio, and video simultaneously, thereby enriching their understanding and generation of information.

III. LLMS FOR MENTAL DISORDERS DETECTION

Given the powerful performance of LLMs, an increasing number of researchers are applying them to the detection and analysis of psychological disorders. This section introduces the applications of LLMs in three categories of psychological disorders (Sec. II-A).

For areas where relevant research is lacking, we propose potential future research directions, emphasizing their strengths and values.

A. *LLMs for Emotional Internalization Disorders*

In terms of emotional internalization disorders, anxiety and depression often co-occur, and the depressive phase of bipolar disorder exhibits similar manifestations to depression. Researchers predict that depression will become the leading cause of the global burden of disease by 2030 [64]. Consequently, much of the current research primarily focuses on depression. Additionally, individuals with depression face a higher risk of suicide, with some studies indicating that about 60% of people who commit suicide suffer from depression [65], and that the risk of dying from suicide in people with depression is higher than that of the general population, making suicide detection another critical area of study. Table II summarizes the research on LLMs in depression and suicide risk detection.

Early studies [20], [66], [67] focused on basic depression detection using pre-trained models like BERT [68] and RoBERTa [69], applied to social media posts. Over time, researchers have incorporated more complex techniques, such as combining LLMs like GPT-4 with domain-specific adaptations (e.g., MentalBERT [70]), and introducing methods like fine-tuning and prompt engineering to improve prediction accuracy, as seen in [71]–[77].

In recent years, there has been a shift towards analyzing users' language changes before and after a depression diagnosis, as demonstrated in [29], [78], [79]. This user-level analysis has gained prominence due to its practicality, as relying solely on individual posts to judge depression can be unreliable. By examining changes in users' posts over an extended period, researchers can gain a more persuasive and comprehensive understanding of their mental health status. Furthermore, explainable approaches, as illustrated in [80]–[83], not only enhance predictive capabilities but also offer valuable insights into the underlying factors influencing individuals' mental health. These approaches emphasize the importance of interpretability, allowing for more transparent and clinically meaningful analysis of mental health data. Additionally, [84] focused on synthesizing new data using LLMs to enhance dataset, addressing data scarcity in mental health research.

Overall, these studies illustrate the evolving role of LLMs and deep learning in mental health research. While pre-trained models such as BERT and MentaLLaMA [85] continue to show high accuracy in depression detection, direct classification using LLMs remains unreliable due to inconsistencies in cue-based approaches. LLMs like GPT have been fine-tuned with natural conversation as one of the optimization goals but have not yet demonstrated strong performance for specific tasks such as classification prediction. Most research has used these LLMs to perform auxiliary tasks like annotation and data enhancement, leaving specific task execution to other models. However, LLMs like GPT also contribute by providing explanations for predictions, which enhances the interpretability of results, making the process more transparent and clinically meaningful. Future research should focus on refining domain-specific LLMs, integrating multimodal, long-range data, and improving interpretability for real-world clinical applications. Hybrid systems combining LLM-based annotation, data augmentation, and pre-trained models show promise in improving both accuracy and interpretability, offering a more reliable and interpretable approach to mental health analysis.

B. LLMs for Psychotic Disorders

Schizophrenia is a serious mental disorder characterized by significant impairments in thinking, perception, and emotions. The social burden caused by schizophrenia is severe [91], with society not only bearing the direct burden of patients' treatment and care but also facing indirect burdens such as productivity loss due to injury and premature death, the strain on nursing staff, and the impact on families and communities. Early detection can significantly improve the care and treatment of patients with schizophrenia [92], positively influencing their quality of life and reducing the burden on caregivers and society.

Traditional machine learning algorithms, such as SVM [93], ANN [94], and XGBoost [95], mostly rely on hand-extracted features, such as sentiment markers in tweets, posting frequency, and emoticon usage. While these features are effective, they have limitations. For instance, the performance of models like SVM and ANN often depends on the quality of feature engineering and may fail to capture deeper linguistic dependencies and sentiment changes. The detailed information is provided in Table III.

In contrast, pre-trained models (e.g., BERT) can effectively capture linguistic contextual information by self-supervised learning on large amounts of textual data, which is crucial for sentiment analysis and pattern recognition of social media texts. BERT is able to understand complex syntactic and semantic structures in text without explicit annotation, providing support for the detection of mental health problems such as schizophrenia. LLMs have strong reasoning ability and can give more accurate diagnostic explanations based on some of the user's symptoms for clinicians' reference. In addition, LLM can be further adapted to specific domains through fine-tuning, providing greater accuracy and reliability.

It is worth noting that LLMs have also been successfully applied in detecting schizophrenia through EEG data [96]. Additionally, social media data, with its real-time, spontaneous, and diverse nature, provides a rich source of behavioral and emotional indicators. The continuous flow of content on social media platforms enables the identification of subtle shifts in language and emotional state over time, which can serve as early markers for psychotic disorders. However, the use of LLMs for detecting schizophrenia on social media is still in its early stages. Despite this, LLMs show significant potential in handling the complexity of natural language, identifying subtle emotional shifts, and enabling automated analysis. With further domain adaptation and model optimization, LLMs are expected to evolve into a powerful tool for detecting schizophrenia and other mental health issues in the future.

C. LLMs for Externalizing Disorders

Externalizing disorders are primarily observed during the pre-adolescent and adolescent years. Those occurring before adolescence classified as early-onset externalizing disorders [102], [103], while those occurring during or after adolescence classified as late-onset externalizing disorders. Additionally, externalizing disorders are frequently comorbid with anxiety. Among children with early-onset externalizing disorders, 30-60% report experiencing anxiety [104], [105]. Similarly, 10-15% of individuals with anxiety in general population report having early-onset externalizing disorders [105], with the prevalence rising to 22% in clinical samples [106].

A study conducted on various adolescent populations through probability sampling [103] found that, in most cases, anxiety disorders do not exhibit a negative correlation with externalizing disorders; in some instances, a positive correlation is even observed. Furthermore, both externalizing and anxiety disorders are linked to depression [107], which experiences a sharp increase in incidence during adolescence. Therefore, the relationship between early and late-onset externalizing disorders and anxiety disorders, as well as the role of depression, warrants further investigation.

Externalizing disorders are characterized by outwardly directed behaviors such as aggression, rule-breaking, and impulsivity. Social media platforms generate a large amount of user speech information that reflects users' language models and characteristics, offering great potential for analyzing whether users suffer from externalizing disorders. However, due to the widespread occurrence of online violence [108], [109], it is difficult to accurately analyze and identify users with externalizing disorders from the public's aggressive comments. Currently, there is limited research on the application of LLMs for detecting externalizing disorders, such as conduct disorder or antisocial behavior, on social media or other platforms.

TABLE II
SUMMARY OF STUDIES ON DEPRESSION AND SUICIDE RISK DETECTION USING LLMS.

Ref	Year	Sample size & Data type	Models	Task	Key Methodologies	Input / Output
Wang et al. [20]	2020	13,993 microblogs, Text	BERT, RoBERTa, XLNeT	Depression Prediction	Use of pre-trained models and autoregressive models	Microblogs / Level 0-4
Metzler et al. [66]	2022	3202 tweets, Text	BERT, XLNet	Suicide-Related Content Detection	Diversified labelling to automatically categorise suicide-related social media content, distinguishing potentially harmful or protective content.	tweets / classifications (6 posts categories, actual suicide or off-topic)
Sabaneh et al. [86]	2023	1058 tweets, Text	GPT-3.5	Depression Detection	Optimising Arabic text analysis using LLM and UMLS, combined with TF-IDF and BOW techniques.	Tweets / Depression or non depression
Owen et al. [87]	2023	770 users, 0.4 million posts, Text	BERT, ALBERT, BioBERT, Longformer, MentalBERT, MentalRoBERTa	Early Depression Detection	Analysing language changes in users prior to diagnosis to determine when models can identify depressive symptoms earliest.	Posts (Different time ranges) / Negative, neutral, positive
Qin et al. [72]	2023	2000 users (TMDD), Text + Image 2000 users (WU3D), Text + Image	ChatGPT-3.5, GPT-3, BERT	Explainable Depression Detection	Combining text and image, introducing DSM-5 criteria for depression, using Chain-of-Thought to allow LLM to reason, and designing an interactive Prompt.	Posts (text + image) / Depressed or non-depressed
Verma et al. [67]	2023	27,972 entries of mental health corpus, Text 7,650 entries of Reddit, Text	RoBERTa	Depression Detection	Linguistic and cognitive profiles were analysed using an enhanced version of the BERT (RoBERTa) model.	Posts / Depressed or non-depressed
Lamichhane [74]	2023	3553 posts, Text	GPT-3.5, BERT	Mental Health Disorders Classification	ChatGPT was used, but not fine-tuned, and only Zero-shot classification was used.	Posts / Stress (2-class), Depression (2-class), Suicidality (5-class)
Bhaumik et al. [73]	2023	2,32,000 posts, Text	ALBERT, Bio-Clinical BERT, GPT-3.5, LLaMa-2	Suicide Risk Detection	ALBERT and Bio-Clinical BERT for fine-tuning, GPT-3.5 using Prompt Engineering, Llama-2 combined with RAG.	Posts / Suicidal or nonsuicidal
Qi et al. [71]	2024	1249 posts, Text 3407 posts, Text	BERT, GPT-3.5, GPT-4, GLM-4, Llama-2, Alpaca	Cognitive Distortions and Suicidal Risks Detection	Multiple Prompting strategies are used: Zero-shot, Few-shot, Role Definition, Scene Definition, Hybrid Prompting, etc.	Posts / Cognitive Distortions multi-label classification, suicide binary classification
Song et al. [78]	2024	500 user timelines from Talklife, Text	TH-VAE (BART-BASE [88]), LLaMA-2	Timeline Summarization	Combining Hierarchical VAE with LLMS to generate clinically significant summaries from social media user timelines.	User's timeline / Timeline summary
Alhamed et al. [79]	2024	120 users, 1.9 million tweets, Text	Alpaca, BERT, RoBERTa, MentalBERT, GPT-3.5, Bard	Depression Detection	Analysing users' language changes before and after a depression diagnosis.	Posts chunks / Before or after diagnosis
Lan et al. [83]	2024	2000 users, 1.38 million posts, Text	GPT-3.5, MentalRoBERTa, BERT, MentalLLama	Depression Detection and Explanation	Combining medical knowledge, LLMs and classifiers for text annotation, Mood course modelling, depression detection.	Posts / Classification results + symptoms + mood course descriptions
Wang et al. [80]	2024	3,107 TREC files, Text 170 users (63,317 writings), Text	Llama-2, SUS-Chat-34B, Neural-chat-7b-v3	Explainable Depression Classification	Calculate the correlation between social media posts and BDI scale questions, evaluate them using LLMs, and generate corresponding explanations.	Writings (posts, comments) + files / Classification + explanation
Liu et al. [75]	2024	54,412 posts (SWHD), Text 8,554 posts (PsySym), Text	BERT, RoBERTa, BioBERT, ClinicalBERT, MentalBERT, MentalRoBERTa, GPT-4, LLaMA-2, MentaLLaMA	Multi-task Mental Health Classification	A Multi-Task Learning framework is proposed to detect multiple mental health states simultaneously.	Posts / Specific diseases classifications
Radwan et al. [76]	2024	2929 users, 3,553 labeled data points (100 tokens in length), Text	GPT-3, BERT	Stress Detection	Converting posts into vector using LLMs embeddings to capture semantic information and linguistic details.	Posts / Indicative or not of stress disorders
Shin et al. [77]	2024	91 participants, 428 diaries, Text	GPT-3.5, GPT-4	Depression Detection	LLMs combine the PHQ-9 and BSS to score users for depression and suicide risk.	Diaries / Score + depressed or non-depressed
Bauer et al. [82]	2024	2.9 million posts, Text	BERT, GPT-4	Suicidality Analysis	BERT is used for sentence embedding. GPT-4 combined with ProtoDash [89] for generating explanations and analysis.	Posts / Prototypical and extreme postings
Ghanadian et al. [84]	2024	733 users (UMD Suicidality), Text	GPT-3.5, Flan-T5, LLaMA-2, ALBERT, DistilBERT	Synthetic Data for Suicidal Ideation	Enhance the dataset by generating data using different LLMs.	Dataset + synthetic dataset / Multi-Class classification and binary classification
Singh et al. [81]	2024	934 users (UMD Suicidality), Text	Mixtral7bx8, Tulu-2-DPO-70B	Extraction of Suicidal Ideation Evidence	Using different Prompting strategies (Zero-shot, Few-shot, Chain-of-Thought, Direct).	Posts + meta-information / Evidence extraction and summary of suicidal ideation
Xu et al. [90]	2024	7 datasets for 6 tasks, Text	BERT, Mental-RoBERT, FLAN-T5, Mental-FLAN-T5, GPT-3.5, GPT-4, Alpaca, Alpaca-LoRA, Mental-Alpaca, LLaMA-2	Depression & Suicide Risk & Stress Prediction	Adopting zero-shot and few-shot prompting to prompt multiple models on different tasks.	Posts + input promopt / Predict results + explanation

Future research in this area could focus on refining LLMs to better identify the linguistic and emotional cues associated with externalizing behaviors. Incorporating multimodal data, such as integrating text with other behavioral indicators (e.g., images or audio), could further enhance the accuracy of detection. Additionally, adapting models to understand context,

TABLE III
SUMMARY OF STUDIES ON SCHIZOPHRENIA DETECTION.

Ref	Year	Sample size & Data type	Models	Task	Key Methodologies	Input / Output
McManus et al. [97]	2015	296 users, Text	ANN, SVM, NB	Schizophrenia Detection	Analyzing Twitter usage patterns, including emoticon use, posting time, and frequency, to distinguish individuals with schizophrenia from controls.	Posts / Schizophrenia diagnosis (True/False)
Mitchell et al. [98]	2015	174 users (3200 tweets per user), Text	SVM	Schizophrenia Detection	Analyzing the language characteristics of Twitter users and identify potential signs of schizophrenia	Posts / Schizophrenia diagnosis (True/False)
Birnbaum et al. [99]	2017	671 users, Text	SVM, LR, NB, RF	Schizophrenia Detection	Combining computational linguistic analysis with clinical appraisals to identify linguistic markers of schizophrenia on social media.	Posts / Schizophrenia diagnosis (True/False)
Kim et al. [100]	2020	228,060 users (488,472 posts), Text	CNN, XGBoost	Mental disorder classification	Analyzing and learning the post information written by users, and developed six independent binary classification models for each symptom.	Posts / Classification (e.g., depression, Schizophrenia)
Bae et al. [101]	2021	265,396 users, 485,350 posts, Text	SVM, LR, NB, RF	Schizophrenia Detection	Combining linguistic feature extraction with topic modeling to identify linguistic markers of schizophrenia.	Posts / Schizophrenia diagnosis (True/False)

such as differentiating between playful banter and harmful aggression, will be essential. As LLMs continue to evolve, their application in detecting externalizing disorders has the potential to become a crucial tool for early intervention, providing both predictive insights and helping reduce the stigma associated with mental health issues.

IV. DATASETS AND EVALUATION METRICS

This section describes several widely used social media datasets, including their sources, formats, composition, collection methods, and sizes. Additionally, we provide the corresponding tasks applicable to each dataset. Table IV summarizes the general information for the eleven datasets. We then introduce some standard evaluation metrics, as well as future research and application directions.

A. Popular Datasets for Mental Disorders

*a) **CLPsych Shared Task (UMD Suicidality Dataset)** [110], [111]:* The UMD dataset contains posts and comments from users on Reddit about suicidal intent or behavior. The dataset collected 1,556,194 posts from 11,129 users, and after filtering out users with fewer than 10 posts, 934 users were selected for annotation through random sampling. The scope of the dataset spans several years and includes the content, location, and time of posts and comments.

*b) **Dreaddit** [112]:* The dataset is a collection of posts from ten subreddits on Reddit between January 1, 2017 and November 19, 2018 across five domains: social, anxiety, abuse, PTSD, and financial. A team of experts independently assessed snippets of posts to determine whether they conveyed a sense of stress, and subsequently integrated their respective scores to generate final binary labels. The dataset is suitable for binary stress prediction.

*c) **DepSeverity** [113]:* The dataset utilized the same posts as Dreaddit, but with a shift in focus to depression content. Two experts categorized each post into four depression severity levels (minimal, mild, moderate, and severe) based on DSM-5 criteria. The dataset was applied to the four levels of depression prediction.

*d) **SDCNL** [114]:* The dataset collects posts from communities such as r/SuicideWatch and r/Depression via Python Reddit API, covering 1,723 users. Each post was manually reviewed by experts to flag the presence of suicidal ideation. The dataset is suitable for binary suicide risk prediction.

*e) **CSSRS-Suicide** [115]:* The CSSRS-Suicide dataset contains posts collected from 15 mental health-related subreddits between 2005 and 2016. Four specialized psychiatrists manually assessed 500 users according to the guidelines of the Columbia Suicide Severity Rating Scale (C-SSRS), classifying their suicide risk into five levels: supportive, indicator, ideation, behavior, and attempt. attempt.The dataset was applied to the five levels of suicide risk prediction.

*f) **RSDD** [116]:* The Reddit Self-Reported Depression Diagnosis (RSDD) dataset contains posts from more than 9,000 users who consider themselves to have been diagnosed with depression (known as "diagnosed users"), as well as posts from more than 107,000 undiagnosed users. Importantly, any content from diagnosed users that appeared in mental health-focused subreddits or contained explicit depression-related phrases (e.g., "I was diagnosed with depression") was excluded from the dataset. The dataset was applied to binary depression predictions.

*g) **Twt-60Users** [117]:* The dataset used the Twitter API to collect tweets from 60 users during 2015. The tweets were meticulously annotated by two professionals to determine the presence of depressive signals. Notably, the dataset showed a strong imbalance, with about 90.7% of the tweets labeled as non-depressed. This is because most of the tweets did not show signs of mental disorders. The dataset is suitable for binary depression prediction.

*h) **TMDD** [118]:* The dataset was constructed in two steps, firstly, tweets from users within a certain time frame were obtained based on self-diagnosis (I am/ I was/ I've been diagnosed depression), which constitutes the text depression dataset. Subsequently, all images were collected using Twitter API based on the IDs of the said tweets. A new multimodal dataset was constructed based on these images and tweets.

TABLE IV
SUMMARY OF DATASETS FOR MENTAL HEALTH ANALYSIS.

Dataset	Source	Task	Dataset Size
CLPsych Shared Task (UMD Suicidality Dataset) [110], [111]	Reddit	Four-level Depression Detection	934 Users (Selected from 11,129 users)
Dreaddit [112]	Reddit	Binary Stress Prediction	3553 Tweets (52.3% True, 47.7% False)
DepSeverity [113]	Reddit	Four-level Depression Prediction	3553 Tweets (72.9% Minimum, 8.2% Mild, 11.3% Moderate, 7.9% Severe)
SDCNL [114]	Reddit	Binary Suicide Risk Prediction	1895 Tweets (48.3% low, 51.7% high)
CSSRS-Suicide [115]	Reddit	Five-level Suicide Risk Prediction	500 Users (21.6% Supportive, 19.8% Indicator, 34.2% Ideation, 15.4% Behavior, 9.0% Attempt)
RSDD [116]	Reddit	Binary Depression Prediction	116,484 Tweets (7.9% True, 92.1% False)
Twt-60Users [117]	Twitter (X)	Binary Depression Prediction	8135 Tweets (9.3% True, 90.7% False)
TMDD [118]	Twitter (X)	Binary Depression Prediction	Users: 2804 (50% True, 50% False) Tweets: 1,111,920 (20.9% True, 79.1% False)
SOS-HL-1K [71]	Weibo	Binary Suicide Risk Prediction	1249 Tweets (51.9% low, 48.1% high)
SWDD [119]	Weibo	Binary Depression Prediction	Users: 23,237 (16% True, 84% False) Tweets: 4,854,421 (16.2% True, 83.8% False)
WU3D [120]	Weibo	Binary Depression Prediction	Users: 32,570 (31.7% True, 68.3% False) Tweets: 2,191,910 (18.7% True, 81.3% False)

*i) **SOS-HL-1K** [71]:* The data is obtained by crawling user comments on the blog "Zoufan" in the microblogging platform. The dataset consists of textual data, mainly user comments. The dataset was annotated by a qualified psychologist. The dataset is suitable for binary depression prediction.

*j) **SWDD** [119]:* The Sina Weibo Depression Dataset (SWDD) contains samples of depressed and non-depressed users and is collected through the official API provided by Sina Weibo. The data in this dataset consists of three parts: the user's personal information, the history of tweets (including timestamps, image data), and a symptom description table (whether there are e.g., depression, panic components in the tweets, etc.). The data labeled depressed users according to a set of labeling criteria proposed in the source article. The dataset is suitable for binary depression prediction.

*k) **WU3D** [120]:* The Weibo User Depression Detection Dataset (WU3D) includes samples of normal and depressed users obtained through the official Weibo API. It contains user profile information and historical tweets (including timestamp and image data). It is worth noting that all samples identified as depressed were manually labeled by expert data annotators and subsequently validated by psychologists and psychiatrists. The dataset is suitable for binary depression prediction.

B. Evaluation Metrics

In the context of classification tasks, evaluating model performance requires a range of metrics that provide a comprehensive understanding of the model's strengths and weaknesses [121], [122]. While accuracy is often used as a baseline, it may not always reflect true performance, especially in imbalanced datasets. Therefore, it is crucial to consider other metrics that account for various aspects of the model's ability to classify both positive and negative samples effectively. The following are the most commonly used evaluation metrics in current research: accuracy, precision, sensitivity, specificity, recall, F1-score, ROC, PRC, AUC, and Kappa statistics, each offering unique insights into model performance.

C. Future Direction

1) **Integrating Mental Health Disorder Detection with Social Event Detection:**
   Integrating mental health disorders detection on social media with social security applications [123]–[126] can proactively mitigate societal risks by identifying potential threats early. By focusing on linguistic cues and posting behaviors, tailored prompts can effectively detect signs of mental disorders [85], [90], [127]. Furthermore, these signs can be assessed to determine whether they indicate violent tendencies, using prompt-driven event inference. Additionally, analyzing historical social media activity allows for the prediction of potential locations, approaches, or contexts for such events. Finally, the noise in social networks and social media should also be investigated, as it can significantly impact model predictions due to the propagation of erroneous information [128], [129].
2) **Supporting Personalized Treatment and Interventions with RAG:**
   Future research could advance personalized mental health detection, treatment, and interventions by integrating social media data with external knowledge and relational

context, tailored to specific scenarios like campuses, workplaces, and families. A promising approach uses Retrieval-Augmented Generation (RAG) [130]–[134] to enhance LLMs by retrieving psychological knowledge and medical records from external sources. It integrates with Knowledge Graphs (KG) [135]–[139] to enrich analysis with insights into social networks and interpersonal relationships.

3) **Advancing Mental Health Research with Explainable LLM-Based Data Synthesis:**
Utilizing LLM-based data synthesis [140]–[143] offers a solution to the limitations in collecting large-scale social media data for mental health studies, which are often constrained by privacy concerns and legal regulations. Although social media generates vast amounts of data, privacy issues and regulatory barriers limit its use for training models, a gap that synthetic data can effectively bridge [144]–[147]. By employing data synthesis techniques, such as prompt engineering approaches [148]–[150] and multi-step generation [151]–[153], researchers can create realistic, anonymized datasets that reflect mental health patterns. This approach has the potential to accelerate mental health research, providing scalable and ethical data solutions while simultaneously improving LLM accuracy for real-world applications. Lastly, the explainability [154] of the predictions should also be explored to ensure reliability.

V. CONCLUSION

In this survey, we provide a comprehensive review of the use of LLMs for detecting mental disorders on social media. We review recent advances in detecting psychological disorders, highlighting the strengths and limitations of state-of-the-art models, datasets, and assessment methods. Although LLMs demonstrate excellent capabilities in understanding complex and unstructured text data, their performance remains suboptimal when applied directly to classification and prediction tasks, as LLM fail to efficiently map inputs to discrete categories due to their generative design and lack of domain-specific classification head. Future research should focus on effectively utilizing these models to directly predict mental disorders in social media users, providing reasonable explanations for such predictions, integrating diverse LLMs and pre-trained models, and exploring cross-platform applications with multimodal data to enhance the accuracy and applicability of LLMs. Additionally, the safety and reliability of LLMs in the field of mental health should be explored, as LLMs often struggle with understanding numbers [17], which are common in this domain. Therefore, whether LLMs can provide reliable suggestions requires thorough investigation.

REFERENCES

[1] J. J. McGrath, A. Al-Hamzawi, J. Alonso, Y. Altwaijri, L. H. Andrade, E. J. Bromet, R. Bruffaerts, J. M. C. de Almeida, S. Chardoul, W. T. Chiu et al., “Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries,” The Lancet Psychiatry, vol. 10, no. 9, pp. 668–681, 2023.

[2] W. H. Organization. (2025) Who website. Accessed: 2025-02-03. [Online]. Available: https://www.who.int/

[3] World Health Organization. (2022) Mental disorders fact sheet. Accessed: Feb. 3, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[4] ——. (2024) Mental health at work. Accessed: Feb. 3, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work

[5] T. L. G. Health, “Mental health matters,” The Lancet. Global Health, vol. 8, no. 11, p. e1352, 2020.

[6] I. Pantic, “Online social networking and mental health,” Cyberpsychology, Behavior, and Social Networking, vol. 17, no. 10, pp. 652–657, 2014.

[7] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner, “Social media and mental health: benefits, risks, and opportunities for research and practice,” Journal of technology in behavioral science, vol. 5, no. 3, pp. 245–257, 2020.

[8] J. Torous, S. Bucci, I. H. Bell, L. V. Kessing, M. Faurholt-Jepsen, P. Whelan, A. F. Carvalho, M. Keshavan, J. Linardon, and J. Firth, “The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality,” World Psychiatry, vol. 20, no. 3, pp. 318–335, 2021.

[9] A. M. Kaplan and M. Haenlein, “Users of the world, unite! the challenges and opportunities of social media,” Business horizons, vol. 53, no. 1, pp. 59–68, 2010.

[10] G. Barbier and H. Liu, “Data mining in social media,” Social network data analytics, pp. 327–352, 2011.

[11] B. of Apps. (2025) Social app report. Accessed: Feb. 3, 2025. [Online]. Available: https://www.businessofapps.com/data/social-app-report/

[12] ——. (2025) Twitter statistics. Accessed: Feb. 3, 2025. [Online]. Available: https://www.businessofapps.com/data/twitter-statistics/

[13] P. Newswire. (2025) Weibo announces third quarter 2024 unaudited financial results. Accessed: Feb. 3, 2025. [Online]. Available: https://www.prnewswire.com/news-releases/weibo-announces-third-quarter-2024-unaudited-financial-results-302309557.html

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

[15] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, “A survey on model compression for large language models,” Transactions of the Association for Computational Linguistics, vol. 12, pp. 1556–1577, 2024.

[16] H. Li, Y. Li, A. Tian, T. Tang, Z. Xu, X. Chen, N. Hu, W. Dong, Q. Li, and L. Chen, “A survey on large language model acceleration based on kv cache management,” arXiv preprint arXiv:2412.19442, 2024.

[17] H. Li, X. Chen, Z. XU, D. Li, N. Hu, F. Teng, Y. Li, L. Qiu, C. J. Zhang, Q. Li et al., “Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models,” arXiv preprint arXiv:2502.11075, 2025.

[18] A. Mudrik, G. N. Nadkarni, O. Efros, B. S. Glicksberg, E. Klang, and S. Soffer, “Exploring the role of large language models in haematology: A focused review of applications, benefits and limitations,” British Journal of Haematology, vol. 205, no. 5, pp. 1685–1698, 2024.

[19] M. Omar Sr, S. Soffer Sr, A. Charney Sr, I. Landi, G. Nadkarni, and E. Klang Jr, “Applications of large language models in psychiatry: A systematic review.” medRxiv, pp. 2024–03, 2024.

[20] X. Wang, S. Chen, T. Li, W. Li, Y. Zhou, J. Zheng, Q. Chen, J. Yan, B. Tang et al., “Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis,” JMIR medical informatics, vol. 8, no. 7, p. e17958, 2020.

[21] X. Wang, S. Chen, T. Li, W. Li, Y. Zhou, J. Zheng, Y. Zhang, and B. Tang, “Assessing depression risk in chinese microblogs: a corpus and machine learning methods,” in 2019 IEEE International conference on healthcare informatics (ICHI). IEEE, 2019, pp. 1–5.

[22] A. Wongkoblap, M. A. Vadillo, and V. Curcin, “Researching mental health disorders in the era of social media: systematic review,” Journal of medical Internet research, vol. 19, no. 6, p. e228, 2017.

[23] R. Skaik and D. Inkpen, “Using social media for mental health surveillance: a review,” ACM Computing Surveys (CSUR), vol. 53, no. 6, pp. 1–31, 2020.

[24] K. Harrigian, C. Aguirre, and M. Dredze, “On the state of social media data for mental health research,” arXiv preprint arXiv:2011.05233, 2020.

[25] A. Lejeune, B.-M. Robaglia, M. Walter, S. Berrouiguet, and C. Lemey, “Use of social media data to diagnose and monitor psychotic disorders:

systematic review," Journal of medical Internet research, vol. 24, no. 9, p. e36986, 2022.

[26] A. Malhotra and R. Jindal, "Deep learning techniques for suicide and depression detection from online social media: A scoping review," Applied Soft Computing, vol. 130, p. 109713, 2022.

[27] M. Garg, "Mental health analysis in social media posts: a survey," Archives of Computational Methods in Engineering, vol. 30, no. 3, pp. 1819–1842, 2023.

[28] N. H. Di Cara, V. Maggio, O. S. Davis, and C. M. Haworth, "Methodologies for monitoring mental health on twitter: Systematic review," Journal of Medical Internet Research, vol. 25, p. e42734, 2023.

[29] D. Owen, A. J. Lynham, S. E. Smart, A. F. Pardiñas, and J. Camacho Collados, "Ai for analyzing mental health disorders among social media users: Quarter-century narrative review of progress and challenges," Journal of Medical Internet Research, vol. 26, p. e59225, 2024.

[30] S. M. Shah, M. M. Aljawarneh, M. A. Saleem, and M. S. Jawarneh, "Mental illness detection through harvesting social media: a comprehensive literature review," PeerJ Computer Science, vol. 10, p. e2296, 2024.

[31] M. Omar and I. Levkovich, "Exploring the efficacy and potential of large language models for depression: A systematic review," Journal of Affective Disorders, 2024.

[32] Y. Hua, F. Liu, K. Yang, Z. Li, H. Na, Y.-h. Sheu, P. Zhou, L. V. Moran, S. Ananiadou, A. Beam et al., "Large language models in mental health care: a scoping review," arXiv preprint arXiv:2401.02984, 2024.

[33] Y. Hua, H. Na, Z. Li, F. Liu, X. Fang, D. Clifton, and J. Torous, "Applying and evaluating large language models in mental health care: A scoping review of human-assessed generative tasks," arXiv preprint arXiv:2408.11288, 2024.

[34] L. Ke, S. Tong, P. Cheng, and K. Peng, "Exploring the frontiers of llms in psychological applications: A comprehensive review," arXiv preprint arXiv:2401.01519, 2024.

[35] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, K. Li et al., "Large language models for mental health applications: Systematic review," JMIR mental health, vol. 11, no. 1, p. e57400, 2024.

[36] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, and M. J. Bell, "The opportunities and risks of large language models in mental health," JMIR Mental Health, vol. 11, no. 1, p. e59479, 2024.

[37] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proceedings of the international AAAI conference on web and social media, vol. 7, no. 1, 2013, pp. 128–137.

[38] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in Proceedings of the 5th annual ACM web science conference, 2013, pp. 47–56.

[39] S. Hetrick, A. Parker, I. Hickie, R. Purcell, A. Yung, and P. McGorry, "Early identification and intervention in depressive disorders: towards a clinical staging model," Psychotherapy and psychosomatics, vol. 77, no. 5, pp. 263–270, 2008.

[40] E. J. Costello, "Early detection and prevention of mental health problems: developmental epidemiology and systems of support," Journal of Clinical Child & Adolescent Psychology, vol. 45, no. 6, pp. 710–717, 2016.

[41] M. R. Garabiles, C. K. Lao, Y. Xiong, and B. J. Hall, "Exploring comorbidity between anxiety and depression among migrant filipino domestic workers: a network approach," Journal of affective disorders, vol. 250, pp. 85–93, 2019.

[42] A. Caspi, R. M. Houts, A. Ambler, A. Danese, M. L. Elliott, A. Hariri, H. Harrington, S. Hogan, R. Poulton, S. Ramrakha et al., "Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the dunedin birth cohort study," JAMA network open, vol. 3, no. 4, pp. e203 221–e203 221, 2020.

[43] F. Biedermann and W. W. Fleischhacker, "Psychotic disorders in dsm-5 and icd-11," CNS spectrums, vol. 21, no. 4, pp. 349–354, 2016.

[44] S. H. McConaughy and R. J. Skiba, "Comorbidity of externalizing and internalizing problems," School Psychology Review, vol. 22, no. 3, pp. 421–436, 1993.

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

[46] A. Radford, "Improving language understanding by generative pre-training," 2018.

[47] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

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

[49] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[50] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu et al., "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," arXiv preprint arXiv:2401.06066, 2024.

[51] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo et al., "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," arXiv preprint arXiv:2405.04434, 2024.

[52] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang et al., "Deepseek-vl: towards real-world vision-language understanding," arXiv preprint arXiv:2403.05525, 2024.

[53] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[54] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," arXiv preprint arXiv:2307.06435, 2023.

[55] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," ACM Computing Surveys, vol. 56, no. 2, pp. 1–40, 2023.

[56] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen, "Large language models for generative information extraction: A survey," Frontiers of Computer Science, vol. 18, no. 6, p. 186357, 2024.

[57] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," Advances in neural information processing systems, vol. 36, 2024.

[58] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[59] W. Berrios, G. Mittal, T. Thrush, D. Kiela, and A. Singh, "Towards language models that can see: Computer vision through the lens of natural language," arXiv preprint arXiv:2306.16410, 2023.

[60] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, and A. Rastogi, "Rlaif: Scaling reinforcement learning from human feedback with ai feedback," 2023.

[61] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," arXiv preprint arXiv:2401.13601, 2024.

[62] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao et al., "A survey on multimodal large language models for autonomous driving," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.

[63] J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip, "Multimodal large language models: A survey," in 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023, pp. 2247–2256.

[64] Y. Zhang, X. Jia, Y. Yang, N. Sun, S. Shi, and W. Wang, "Change in the global burden of depression from 1990-2019 and its prediction for 2030," Journal of psychiatric research, vol. 178, pp. 16–22, 2024.

[65] S. Law and P. Liu, "Suicide in china: unique demographic patterns and relationship to depressive disorder," Current Psychiatry Reports, vol. 10, pp. 80–86, 2008.

[66] H. Metzler, H. Baginski, T. Niederkrotenthaler, and D. Garcia, "Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach," Journal of medical internet research, vol. 24, no. 8, p. e34705, 2022.

[67] S. Verma, R. C. Joshi, M. K. Dutta, S. Jezek, R. Burget et al., "Ai-enhanced mental health diagnosis: Leveraging transformers for early detection of depression tendency in textual data," in 2023 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). IEEE, 2023, pp. 56–61.

[68] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 conference of the North American chapter

of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[69] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[70] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," arXiv preprint arXiv:2110.15621, 2021.
[71] H. Qi, Q. Zhao, J. Li, C. Song, W. Zhai, L. Dan, S. Liu, Y. J. Yu, F. Wang, H. Zou et al., "Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media," 2023.
[72] W. Qin, Z. Chen, L. Wang, Y. Lan, W. Ren, and R. Hong, "Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media," arXiv preprint arXiv:2305.05138, 2023.
[73] R. Bhaumik, V. Srivastava, A. Jalali, S. Ghosh, and R. Chandrasekharan, "Mindwatch: A smart cloud-based ai solution for suicide ideation detection leveraging large language models," MedRxiv, pp. 2023–09, 2023.
[74] B. Lamichhane, "Evaluation of chatgpt for nlp-based mental health applications," arXiv preprint arXiv:2303.15727, 2023.
[75] J. Liu and M. Su, "Enhancing mental health condition detection on social media through multi-task learning," medRxiv, pp. 2024–02, 2024.
[76] A. Radwan, M. Amarneh, H. Alawneh, H. I. Ashqar, A. AlSobeh, and A. A. A. R. Magableh, "Predictive analytics in mental health leveraging llm embeddings and machine learning models for social media analysis," International Journal of Web Services Research (IJWSR), vol. 21, no. 1, pp. 1–22, 2024.
[77] D. Shin, H. Kim, S. Lee, Y. Cho, and W. Jung, "Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: Instrument validation study," Journal of Medical Internet Research, vol. 26, p. e54617, 2024.
[78] J. Song, J. Chim, A. Tsakalidis, J. Ive, D. Atzil-Slonim, and M. Liakata, "Combining hierachical vaes with llms for clinically meaningful timeline summarisation in social media," in Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 14 651–14 672.
[79] F. Alhamed, J. Ive, and L. Specia, "Classifying social media users before and after depression diagnosis via their language usage: A dataset and study," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 3250–3260.
[80] Y. Wang, D. Inkpen, and P. K. Gamaarachchige, "Explainable depression detection using large language models on social media data," in Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), 2024, pp. 108–126.
[81] L. G. Singh, J. Mao, R. Mutalik, and S. E. Middleton, "Extraction and summarization of suicidal ideation evidence in social media content using large language models," in Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, 2024.
[82] B. Bauer, R. Norel, A. Leow, Z. A. Rached, B. Wen, and G. Cecchi, "Using large language models to understand suicidality in a social media–based taxonomy of mental health disorders: Linguistic analysis of reddit posts," JMIR mental health, vol. 11, p. e57234, 2024.
[83] X. Lan, Y. Cheng, L. Sheng, C. Gao, and Y. Li, "Depression detection on social media with large language models," arXiv preprint arXiv:2403.10750, 2024.
[84] H. Ghanadian, I. Nejadgholi, and H. Al Osman, "Socially aware synthetic data generation for suicidal ideation detection using large language models," IEEE Access, 2024.
[85] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "Mentallama: interpretable mental health analysis on social media with large language models," in Proceedings of the ACM Web Conference 2024, 2024, pp. 4489–4500.
[86] K. Sabaneh, M. A. Salameh, F. Khaleel, M. M. Herzallah, J. Y. Natsheh, and M. Maree, "Early risk prediction of depression based on social media posts in arabic," in 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2023, pp. 595–602.
[87] D. Owen, D. Antypas, A. Hassoulas, A. F. Pardiñas, L. Espinosa-Anke, J. C. Collados et al., "Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation," JMIR AI, vol. 2, no. 1, p. e41205, 2023.
[88] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv:1910.13461, 2019.
[89] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019, pp. 260–269.
[90] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-llm: Leveraging large language models for mental health prediction via online text data," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 8, no. 1, pp. 1–32, 2024.
[91] W. Rössler, H. J. Salize, J. Van Os, and A. Riecher-Rössler, "Size of burden of schizophrenia and psychotic disorders," European neuropsychopharmacology, vol. 15, no. 4, pp. 399–409, 2005.
[92] N. Albert and M. A. Weibell, "The outcome of early intervention in first episode psychosis," International Review of Psychiatry, vol. 31, no. 5-6, pp. 413–424, 2019.
[93] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18–28, 1998.
[94] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," Heliyon, vol. 4, no. 11, 2018.
[95] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[96] M. Guerra, R. Milanese, M. Deodato, M. G. Ciobanu, and F. Fasano, "Exploring the diagnostic potential of llms in schizophrenia detection through eeg analysis," in 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024, pp. 6812–6819.
[97] K. McManus, E. K. Mallory, R. L. Goldfeder, W. A. Haynes, and J. D. Tatum, "Mining twitter data to improve detection of schizophrenia," AMIA Summits on Translational Science Proceedings, vol. 2015, p. 122, 2015.
[98] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality, 2015, pp. 11–20.
[99] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, M. De Choudhury, and J. M. Kane, "A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals," Journal of medical Internet research, vol. 19, no. 8, p. e7956, 2017.
[100] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," Scientific reports, vol. 10, no. 1, p. 11846, 2020.
[101] Y. J. Bae, M. Shim, and W. H. Lee, "Schizophrenia detection using machine learning approach from social media content," Sensors, vol. 21, no. 17, p. 5924, 2021.
[102] J. Burke and R. Loeber, "Oppositional defiant disorder and the explanation of the comorbidity between behavioral disorders and depression," Clinical Psychology: Science and Practice, vol. 17, no. 4, pp. 319–326, 2010.
[103] N. R. Marmorstein, "Relationships between anxiety and externalizing disorders in youth: The influences of age and gender," Journal of anxiety disorders, vol. 21, no. 3, pp. 420–432, 2007.
[104] M. K. Nock, A. E. Kazdin, E. Hiripi, and R. C. Kessler, "Lifetime prevalence, correlates, and persistence of oppositional defiant disorder: results from the national comorbidity survey replication," Journal of Child Psychology and Psychiatry, vol. 48, no. 7, pp. 703–713, 2007.
[105] A. Angold, E. J. Costello, and A. Erkanli, "Comorbidity," The Journal of Child Psychology and Psychiatry and Allied Disciplines, vol. 40, no. 1, pp. 57–87, 1999.
[106] T. L. Verduin and P. C. Kendall, "Differential occurrence of comorbidity within childhood anxiety disorders," Journal of Clinical Child and Adolescent Psychology, vol. 32, no. 2, pp. 290–295, 2003.
[107] A. K. Loth, D. A. Drabick, E. Leibenluft, and L. A. Hulvershorn, "Do childhood externalizing disorders predict adult depression? a meta-analysis," Journal of abnormal child psychology, vol. 42, pp. 1103–1113, 2014.

[108] J. B. Walther, "Social media and online hate," Current Opinion in Psychology, vol. 45, p. 101298, 2022.
[109] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López, "Internet, social media and online hate speech. systematic review," Aggression and violent behavior, vol. 58, p. 101608, 2021.
[110] J. Chim, A. Tsakalidis, D. Gkoumas, D. Atzil-Slonim, Y. Ophir, A. Zirikly, P. Resnik, and M. Liakata, "Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts," in Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), 2024, pp. 177–190.
[111] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, 2018, pp. 25–36.
[112] E. Turcan and K. McKeown, "Dreaddit: A reddit dataset for stress analysis in social media," arXiv preprint arXiv:1911.00133, 2019.
[113] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early identification of depression severity levels on reddit using ordinal classification," in Proceedings of the ACM Web Conference 2022, 2022, pp. 2563–2572.
[114] A. Haque, V. Reddi, and T. Giallanza, "Deep learning for suicide and depression identification with unsupervised label correction," in Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30. Springer, 2021, pp. 436–447.
[115] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in The world wide web conference, 2019, pp. 514–525.
[116] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," arXiv preprint arXiv:1709.01848, 2017.
[117] Z. Jamil, "Monitoring tweets for depression to detect at-risk users," Ph.D. dissertation, Université d'Ottawa/University of Ottawa, 2017.
[118] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 110–117.
[119] Y. Cai, H. Wang, H. Ye, Y. Jin, and W. Gao, "Depression detection on online social network with multivariate time series feature of user depressive symptoms," Expert Systems with Applications, vol. 217, p. 119538, 2023.
[120] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "A multitask deep learning approach for user depression detection on sina weibo," arXiv preprint arXiv:2008.11708, 2020.
[121] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in Computer Science On-line Conference. Springer, 2023, pp. 15–25.
[122] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Scientific Reports, vol. 14, no. 1, p. 6086, 2024.
[123] J. Simpson, V. Albani, Z. Bell, C. Bambra, and H. Brown, "Effects of social security policy reforms on mental health and inequalities: a systematic review of observational studies in high-income countries," Social Science & Medicine, vol. 272, p. 113717, 2021.
[124] Z. Xu, M. Gahr, Y. Xiang, D. Kingdon, N. Rüsch, and G. Wang, "The state of mental health care in china," Asian Journal of Psychiatry, vol. 69, p. 102975, 2022.
[125] J. B. Kirkbride, D. M. Anglin, I. Colman, J. Dykxhoorn, P. B. Jones, P. Patalay, A. Pitman, E. Soneson, T. Steare, T. Wright et al., "The social determinants of mental health and disorder: evidence, prevention and recommendations," World psychiatry, vol. 23, no. 1, pp. 58–90, 2024.
[126] Y. Cao, G. S. Y. Kwan, Z. C. K. Tse, K. K. S. Chow, D. K. S. Kwan, W. W. Y. Lam, and D. H. K. Shum, "Managing uncertainty and loneliness: Protective and risk factors impacting on older people's mental health in hong kong," Journal of Gerontological Social Work, vol. 68, no. 2, pp. 234–255, 2025.
[127] Y. Guo, J. Liu, L. Wang, W. Qin, S. Hao, and R. Hong, "A prompt-based topic-modeling method for depression detection on low-resource data," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 1430–1439, 2023.
[128] H. Li, S. Di, C. H. Y. Li, L. Chen, and X. Zhou, "Fight fire with fire: Towards robust graph neural networks on dynamic graphs via actively defense," Proceedings of the VLDB Endowment, vol. 17, no. 8, pp. 2050–2063, 2024.
[129] H. Li, S. Di, Z. Li, L. Chen, and J. Cao, "Black-box adversarial attack and defense on graph neural networks," in 2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, 2022, pp. 1017–1030.
[130] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," Advances in neural information processing systems, vol. 33, pp. 9459–9474, 2020.
[131] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 7969–7992.
[132] Y. Wang, H. Xin, and L. Chen, "Kglink: A column type annotation method that combines knowledge graph and pre-trained language model," in 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE, 2024, pp. 1023–1035.
[133] Y. Wang, H. Li, F. Teng, and L. Chen, "Graph-based retrieval augmented generation for dynamic few-shot text classification," arXiv preprint arXiv:2501.02844, 2025.
[134] S. Es, J. James, L. E. Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, pp. 150–158.
[135] Z. Jia, H. Li, and L. Chen, "Air: Adaptive incremental embedding updating for dynamic knowledge graphs," in International Conference on Database Systems for Advanced Applications. Springer, 2023, pp. 606–621.
[136] F. Teng, H. Li, S. Di, and L. Chen, "Cardinality estimation on hyper-relational knowledge graphs," arXiv preprint arXiv:2405.15231, 2024.
[137] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, and L. Sun, "Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, 2024, pp. 18 126–18 134.
[138] D. Le, K. Zhao, M. Wang, and Y. Wu, "Graphlingo: Domain knowledge exploration by synchronizing knowledge graphs and large language models," in 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE, 2024, pp. 5477–5480.
[139] W. Chen, H. Wan, Y. Wu, S. Zhao, J. Cheng, Y. Li, and Y. Lin, "Local-global history-aware contrastive learning for temporal knowledge graph reasoning," in 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE, 2024, pp. 733–746.
[140] K. Wang, J. Zhu, M. Ren, Z. Liu, S. Li, Z. Zhang, C. Zhang, X. Wu, Q. Zhan, Q. Liu et al., "A survey on data synthesis and augmentation for large language models," arXiv preprint arXiv:2410.12896, 2024.
[141] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Pre-trained language models for text generation: A survey," ACM Computing Surveys, vol. 56, no. 9, pp. 1–39, 2024.
[142] R. Liu, J. Wei, F. Liu, C. Si, Y. Zhang, J. Rao, S. Zheng, D. Peng, D. Yang, D. Zhou et al., "Best practices and lessons learned on synthetic data," arXiv preprint arXiv:2404.07503, 2024.
[143] A. Havrilla, A. Dai, L. O'Mahony, K. Oostermeijer, V. Zisler, A. Albalak, F. Milo, S. C. Raparthy, K. Gandhi, B. Abbasi et al., "Surveying the effects of quality, diversity, and complexity in synthetic data from large language models," arXiv preprint arXiv:2412.02980, 2024.
[144] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. Springer, 2019, pp. 510–526.
[145] S. M. Bellovin, P. K. Dutta, and N. Reitinger, "Privacy and synthetic datasets," Stan. Tech. L. Rev., vol. 22, p. 1, 2019.
[146] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data," Neurocomputing, vol. 416, pp. 244–255, 2020.
[147] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," NPJ digital medicine, vol. 6, no. 1, p. 186, 2023.

[148] N. Freise, M. Heitlinger, R. Nuredini, and G. Meixner, "Automatic prompt optimization techniques: Exploring the potential for synthetic data generation," arXiv preprint arXiv:2502.03078, 2025.

[149] H. Gupta, K. Scaria, U. Anantheswaran, S. Verma, M. Parmar, S. A. Sawant, C. Baral, and S. Mishra, "Targen: Targeted data generation with large language models," arXiv preprint arXiv:2310.17876, 2023.

[150] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong, "Zerogen: Efficient zero-shot learning via dataset generation," arXiv preprint arXiv:2202.07922, 2022.

[151] R. Xu, H. Cui, Y. Yu, X. Kan, W. Shi, Y. Zhuang, W. Jin, J. Ho, and C. Yang, "Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models," arXiv preprint arXiv:2311.00287, 2023.

[152] M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. F. Chen, Z. Liu, and D. Yang, "Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation," arXiv preprint arXiv:2310.15638, 2023.

[153] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," arXiv preprint arXiv:2212.10560, 2022.

[154] W. Fan, L. Fan, D. Lin, and M. Xie, "Explaining gnn-based recommendations in logic," PVLDB, vol. 18, no. 3, pp. 715–728, 2025.