# Toward a Mental Health Counseling System: A Bibliometric and Qualitative Analysis of Dialogue Systems for Mental Health

Jinyoung Han, Daeun Lee, Dongje Yoo, Migyeong Yang, Jihyun An

Submitted to: Journal of Medical Internet Research
on: November 26, 2024

# *Table of Contents*

# Toward a Mental Health Counseling System: A Bibliometric and Qualitative Analysis of Dialogue Systems for Mental Health

Jinyoung Han[1] PhD; Daeun Lee[1*] PhD; Dongje Yoo[1*]; Migyeong Yang[1]; Jihyun An[2] MD

[1]Sungkyunkwan University Seoul KR
[2]Samsung Medical Center Seoul KR
[*]these authors contributed equally

**Corresponding Author:**
Jinyoung Han PhD
Sungkyunkwan University
25-2, Seonggyungwan-ro, Jongno-gu
Seoul
KR

## Abstract

**Background:** The importance of mental health has been increasingly highlighted, yet many individuals still face barriers to accessing suitable interventions. Although AI-based dialogue systems for mental health enhancement have advanced notably to address this issue, comprehensive surveys in this area, particularly those considering studies that adopt large language models (LLMs), remain scarce.

**Objective:** This study aims to conduct a quantitative and qualitative review of current research trends in AI-driven dialogue systems for enhancing mental health.

**Methods:** This study performed a bibliometric analysis and a trend review analysis of AI-driven dialogue systems for mental health, covering literature from 2020 to May 2024 across three citation databases—WoS, Scopus, and ACM Digital Library. The bibliometric analysis statistically assessed the distribution of publications, while the qualitative trend review focused on three key areas: (i) highly cited publications, (ii) those using the ESConv dataset, and (iii) those employing LLMs.

**Results:** We reviewed 146 papers published between 2020 and 2024, observing a steady increase in publications over the last five years. Our bibliometric analysis examined publication distribution across sources, countries, institutions, and authors, while keyword network analysis highlighted major themes. Most of the top 10 highly cited papers focused on empathetic response generation, incorporating psychological approaches within deep learning models. For the ESConv dataset's application in counseling, prominent techniques included multi-task learning and the integration of external knowledge. Lastly, we identified notable advantages of LLMs over traditional deep learning models and explored strategies to overcome their limitations as counseling tools.

**Conclusions:** Our study identifies key areas for developing counseling dialogue systems, such as incorporating psychological knowledge, improving data access, applying LLMs, and refining evaluation methods. By examining current research trends and establishing a foundational framework, this work offers future directions to enhance the effectiveness of AI counseling systems, contributing to both the machine learning and psychology fields.

(JMIR Preprints 26/11/2024:69266)
DOI: https://doi.org/10.2196/preprints.69266

### Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

Original Paper

# Toward a Mental Health Counseling System: A Bibliometric and Qualitative Analysis of Dialogue Systems for Mental Health

Dongje Yoo[1*], Daeun Lee[2*], Migyeong Yang[2], Jihyun An[3] and Jinyoung Han[2,4†]
[1]Department of Immersive Media Engineering, Sungkyunkwan University, Seoul, Korea
[2]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, Korea
[3]Department of Psychiatry, Samsung Medical Center, Seoul, Korea
[4]Department of Human-AI Interaction, Sungkyunkwan University, Seoul, Korea

* Equal contribution.
†Corresponding author.

## Abstract

**Background:** The importance of mental health has been increasingly highlighted, yet many individuals still face barriers to accessing suitable interventions. Although AI-based dialogue systems for mental health enhancement have advanced notably to address this issue, comprehensive surveys in this area, particularly those considering studies that adopt large language models (LLMs), remain scarce.

**Objective:** This study aims to conduct a quantitative and qualitative review of current research trends in AI-driven dialogue systems for enhancing mental health.

**Methods:** This study performed a bibliometric analysis and a trend review analysis of AI-driven dialogue systems for mental health, covering literature from 2020 to May 2024 across three citation databases—WoS, Scopus, and ACM Digital Library. The bibliometric analysis statistically assessed the distribution of publications, while the qualitative trend review focused on three key areas: (i) highly cited publications, (ii) those using the ESConv dataset, and (iii) those employing LLMs.

**Results:** We reviewed 146 papers published between 2020 and 2024, observing a steady increase in publications over the last five years. Our bibliometric analysis examined publication distribution across sources, countries, institutions, and authors, while keyword network analysis highlighted major themes. Most of the top 10 highly cited papers focused on empathetic response generation, incorporating psychological approaches within deep learning models. For the ESConv dataset's application in counseling, prominent techniques included multi-task learning and the integration of external knowledge. Lastly, we identified notable advantages of LLMs over traditional deep learning models and explored strategies to overcome their limitations as counseling tools.

**Conclusions:** Our study identifies key areas for developing counseling dialogue systems, such as incorporating psychological knowledge, improving data access, applying LLMs, and refining evaluation methods. By examining current research trends and establishing a foundational framework, this work offers future directions to enhance the effectiveness of AI counseling systems, contributing to both the machine learning and psychology fields.

**Keywords:** review; mental health; counseling; deep learning; large language model;

emotional support conversation; empathetic response generation

## Introduction

Mental disorders have emerged as a critical global public health issue. OECD (Organization for Economic Cooperation and Development) reports that, in the United States, 14.1 out of every 100,000 people die by suicide each year [137]. Unfortunately, over 80% of these suicides are committed by individuals suffering from mental illness [100]. While early intervention is crucial, traditional mental health services, such as counseling and psychological therapy, often encounter significant barriers, including financial constraints, time limitations, and geographic restrictions [22, 56]. Furthermore, the stigma and shame associated with discussing personal difficulties contribute to the reluctance of many individuals to seek support [86], thereby endangering their mental health. These challenges became more evident with the heightened demand for digital mental health solutions during the COVID-19 pandemic [6].

Thankfully, advancements in artificial intelligence (AI) are offering new solutions in psychological counseling [90]. Specifically, dialogue systems — AI-based models capable of engaging in coherent and contextually appropriate conversations with humans through natural language [14] — are being increasingly incorporated into mental health services, assisting clients in self-exploration, gaining insight, taking action, and fostering their own healing processes [44]. For instance, AI-based applications, such as virtual therapists [105] and robotic counselors [39], have demonstrated substantial effectiveness in improving both the quality of mental health care and its accessibility [67, 9].

However, despite this progress, several limitations still exist for effective implementation in real-world contexts. For example, many applications continue to rely on rule-based algorithms [25], which frequently result in constrained and superficial interactions [66]. While deep learning-based dialogue systems, such as those utilizing large language models (LLMs), demonstrate advanced natural interaction capabilities, they often focus on optimizing performance rather than incorporating essential counseling functions, such as generating empathetic responses [18] and relationship building [45], which are essential for practical applicability and user-friendliness in real-world settings. Besides, current evaluation metrics are inadequate for accurately assessing the effectiveness of psychological counseling systems, which rely on essential factors like empathy [20], rapport [30], and perceived helpfulness [82]. Inconsistencies in evaluation metrics across studies also complicate comparisons, limiting the ability to gauge the true effectiveness of these systems in mental health counseling [120].

Recognizing the gaps between technological advancements and practical needs of counseling, we emphasize the importance of conducting an in-depth review of research trends to help evaluate the current capabilities of AI-based dialogue systems in mental health counseling and to clarify their limitations for effective real-world implementation.

In line with this, there are several review papers that address mental health dialogue systems. For instance, Haque and Rubya [41] discussed an overview of chatbot-based mobile mental health apps, while Coghlan *et al.* [21] focused on the ethical issues that should be considered when developing a mental health chatbot. Additionally, Ahmed *et al.* [2] conducted a scoping review on chatbots for depression and anxiety interventions, and Catania *et al.* [13] explored conversational agents targeting interventions for neurodevelopmental disorders. However, limited attention has been given to investigating the technological advancements underlying deep learning-based dialogue systems for mental health enhancement. While Rangaswamy and others [90] explored AI-driven mental health counseling systems, their focus remained primarily on categorizing the types of available applications rather than delving into the AI technologies that support these systems. Furthermore, despite recent progress in LLMs for multi-turn dialogue systems [97], there are few studies evaluating their potential as counseling systems.

Therefore, this study aims to provide technological insights for future research in deep learning- based mental health dialogue systems, with the goal of making these systems more effective and accessible for counseling. Specifically, we employ a quantitative bibliometric analysis to examine research trends in deep learning mental health dialogue systems. In addition, we conduct a qualitative trend analysis to identify the gap between technological advancements and the requirements of psychological counseling across three categories: (i) highly cited publications, (ii) publications utilizing the well-known counseling open dataset, ESConv [69], and (iii) publications employing LLMs. In the discussion section, we address the implications and recommendations for advancing mental health counseling systems and explore the study's limitations and outlines directions for future research.

## Method

### Data Sources and Search Strategy

We sourced papers from three major citation databases: Scopus, Web of Science (WoS), and the Association for Computing Machinery (ACM). Scopus [138] is one of the largest citation repositories, encompassing a wide range of scientific journals, conference papers, and books. WoS [139] includes reputable publications categorized under the Science Citation Index Expanded (SCIE), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI). The ACM Digital Library [140], a prominent organization in computing, provides an extensive digital library that includes journals, conference proceedings, and technical magazines, making it a key resource for research in computer science and information technology. We retrieved relevant publications where the search terms appeared in the title, abstract, or keywords. The search queries were developed based on previous research on conversational agents for mental health [69, 91, 118], as illustrated in Figure 1.

The selection criteria for studies included those published in English, appearing in peer-reviewed scientific journals or conference proceedings, which are recognized as high-quality publication venues in engineering and computer science. Studies categorized as closed access were excluded. Papers published between 2020 and 2024 were retrieved as of 7 May 2024. In order to account for the rapidly evolving advancements in machine learning, papers from 2024 were also considered, acknowledging that the year remained open for further developments.

Figure 1: Search Query Categories with Results

## Selection Criteria

We specifically included studies that aimed to enhance individuals' mental health and involved the utilization or development of deep learning models. These criteria were further refined based on the core conditions; papers focused solely on mental health detection were excluded, as they lacked direct interaction with individuals to support mental health improvement. Likewise, studies centered on dialogue systems for tasks such as negotiation generation or emotional response rewriting were excluded, as they did not primarily target mental health enhancement. Only studies proposing novel deep learning models, rather than evaluations of existing applications or chatbots, were considered. The inclusion and exclusion criteria related to outcomes are outlined in Figure 2. As a result, out of a total of 1,332 papers retrieved through the database search, 146 papers were ultimately obtained after thorough screening.

Figure 2: Selection Criteria Overview

## Analysis Methodology

In this study, we applied bibliometric analysis, a widely employed quantitative method for examining literature, which plays a significant role in AI/ML healthcare research by highlighting developmental trends and ensuring the generation of measurable, consistent, and unbiased results [55]. We first began by analyzing the distribution of publications across different categories, such as sources, countries, institutions, and authors. We also performed a network analysis of commonly used keywords to uncover dominant themes and emerging trends within the literature. Note that statistical evaluations were carried out using Python and Microsoft Excel. Moreover, we conducted a trend analysis across three categories: (i) highly cited publications, (ii) publications utilizing the well-known counseling open dataset, ESConv [69], and (iii) publications employing LLMs. We examined the characteristics of the papers in each category and discussed the challenges in developing fully functional psychotherapy dialogue systems. While the bibliometric analysis is largely software-generated, the trend analysis is subjective and guided by the authors.

## Bibliometric analysis

### Overall Publication Trend

Table 1 demonstrates a steady increase in publications from 2020 to 2024 (up to May 2024). In 2020, only 9 papers were screened, whereas publication activity surged in 2023, with 62 papers screened. Based on the screening date in May 2024, it is anticipated that the number of screened papers for the remainder of 2024 will closely mirror the totals observed in 2023.

Table 1: Number of Publications by Year

| Year | Count, n (%) |
|---|---|
| 2020 | 9 (6.1%) |
| 2021 | 25 (17.1%) |
| 2022 | 36 (24.7%) |
| 2023 | 62 (42.4%) |
| 2024 | 14 (9.6%) |
| Total | 146 (100.0%) |

### Productive Publication Source

We examined the publication sources of the selected papers, which included journal articles, conference proceedings, and book chapters. Table 2 displays the sources with the highest number of publications across Scopus, WoS, and ACM. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), a top-tier natural language processing conference in computer science, was identified as the leading source, closely followed by Lecture Notes in Computer Science, with both contributing more than 10 publications.

Table 2: Top Sources for Publications

| Rank | Source | Count, n (%) |
|---|---|---|
| 1 | Annual Meeting of the Association for Computational Linguistics (ACL) | 16 (10.9) |
| 2 | Lecture Notes in Computer Science | 13 (8.9) |
| 3 | IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 6 (4.1) |
| 3 | Findings of the Association for Computational Linguistics (EMNLP Findings) | 6 (4.1) |
| 5 | Conference on Empirical Methods in Natural Language Processing (EMNLP) | 5 (3.4) |
| 6 | Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) | 4 (2.7) |
| 7 | AAAI Conference on Artificial Intelligence (AAAI) | 4 (2.7) |
| 8 | Sun SITE Central Europe Workshop (CEUR Workshop) | 3 (2.1) |
| 8 | Knowledge-Based Systems | 3 (2.1) |

| | | |
|---|---|---|
| 8 | The Web (formerly, World Wide Web Conference) | 3 (2.1) |

## Predominant Countries

As shown in Table 3, more than 10 countries were recognized as the most productive based on their publication output. China was the leading contributor, followed by India and the United States.

Table 3: Top Countries for Publications

| Rank | Country | Count, n (%) |
|---|---|---|
| 1 | China | 54 (37.0) |
| 2 | India | 25 (17.1) |
| 3 | United States | 10 (6.8) |
| 3 | Hong Kong | 8 (5.5) |
| 5 | South Korea | 7 (4.8) |
| 6 | Japan | 4 (2.7) |
| 7 | United Kingdom | 4 (2.7) |
| 8 | Taiwan | 4 (2.7) |
| 8 | Australia | 3 (2.1) |
| 8 | Singapore | 3 (2.1) |

## Productive Institutions

A total of 90 institutions were associated with the 146 publications. The top-ranked institutions are listed in Table 4. Tsinghua University in China was the most productive institution, followed by the Indian Institute of Technology.

Table 4: Top Institutions for Publications

| Rank | Institution | Country | Count, n (%) |
|---|---|---|---|
| 1 | Tsinghua University | China | 8 (5.5) |
| 2 | Indian Institute of Technology | India | 7 (4.8) |
| 3 | Institute of Information Engineering | China | 5 (3.4) |
| 4 | Shandong University | China | 4 (2.7) |
| 4 | Northeastern University | China | 4 (2.7) |
| 4 | Harbin Institute of Technology | China | 4 (2.7) |
| 7 | Tianjin University | China | 3 (2.1) |
| 7 | National Cheng Kung University | Taiwan | 3 (2.1) |
| 9 | The University of Tokyo | Japan | 2 (1.4) |
| 9 | Beijing University | China | 2 (1.4) |
| 9 | Seoul National University | South Korea | 2 (1.4) |

## Predominant Authors

The top 10 researchers in this field are presented in Table 5, ranked by their publication count. Six of these researchers are affiliated with institutions in China, while two are associated with the Indian Institute of Technology. The most prolific researcher was Professor Su Y from Northwest Normal University, who had three publications.

Table 5: Top 10 Most Productive Authors for Publications

| Author | Institution | Country | Count, n (%) |
|---|---|---|---|
| Su Y | Northwest Normal University | China | 3 (2.1) |

| | | | |
|---|---|---|---|
| Saha T | Indian Institute of Technology | India | 2 (1.4) |
| Bi G | Institute of Information Engineering | China | 2 (1.4) |
| Majumder N | Singapore University of Technology and Design | Singapore | 2 (1.4) |
| Zhou J | Tsinghua University | China | 2 (1.4) |
| Peng W | Institute of Information Engineering | China | 2 (1.4) |
| Shen S | University of Michigan | United States | 2 (1.4) |
| Mishra K | Indian Institute of Technology | India | 2 (1.4) |
| Wang J | Hong Kong Polytechnic University | Hong Kong | 2 (1.4) |
| Li Q | Shandong University | China | 2 (1.4) |

## Author Keyword Co-occurrence

We analyzed the main keywords selected by the authors, which represent the central themes of the publications. In Figure 3, the co-occurrence of these keywords is visualized through a network graph, a widely used method in bibliometric analysis [60, 55]. Each node represents a keyword, and the edges between nodes indicate the co-occurrence of those keywords within individual publications. To improve clarity, edges representing fewer than three co-occurrences were removed after constructing the network graph.

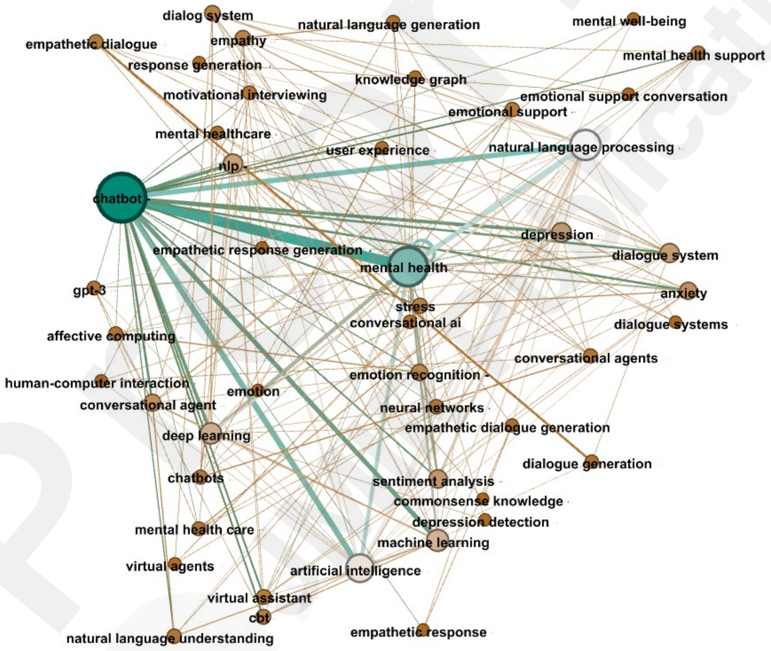



Figure 3: Keyword Co-occurrence Network Graph from 2020 to May 2024

Based on the most frequent occurrences of specific keywords, we categorized the main areas of research focus as follows: (i) ‘Mental Health,’ (ii) ‘Chatbot,’ (iii) ‘Artificial Intelligence,’ and (iv) ‘Natural Language Processing (NLP).’ Focusing on these keywords, we identified additional prominent terms. In the context of ‘Mental health,’ ‘sentiment analysis’ was the most frequently cited keyword, followed by ‘anxiety,’ ‘NLP,’ and ‘dialogue system.’ For ‘Chatbot’-related terms, ‘machine learning’ was the most commonly used, followed by ‘depression,’ ‘NLP,’ ‘deep learning,’ ‘anxiety,’ and ‘dialogue system.’

Given that major trends can fluctuate annually, we also analyzed the author keywords on a year-by-year basis. Accordingly, we generated annual co-occurrence graphs for author keywords that appeared more than twice, as illustrated in Figure 4. Firstly, in 2020, only 9 papers were published, leading to a relatively small number of overall keywords. The primary keywords were concentrated around ‘natural language processing,’ ‘machine learning,’ ‘chatbot,’ and ‘depression,’ all of which were related to dialogue systems. In 2021, prominent keywords related to mental health, such as ‘depression,’ ‘anxiety,’ and ‘COVID-19’, as well as chatbot technologies like ‘conversational agent,’ ‘dialogue system,’ and ‘conversational AI,’ frequently emerged as central themes. Next, in 2022, keywords centered around artificial intelligence began to appear, including ‘empathy,’ ‘motivational interviewing,’ and ‘virtual assistant.’ In 2023, which had the highest number of publications, a diverse range of keywords was identified. Empathy-related terms, such as ‘empathetic dialogue generation’ and ‘empathetic response,’ were particularly notable. Additionally, keywords like ‘emotional support conversation,’ ‘motivational interviewing,’ and ‘conversational robot’ were also distinguished. This suggests that research on the mental health dialogue system became more specialized and diverse in 2023. Interestingly, keywords like ‘large language model’ and ‘GPT-3’ appeared in 2024, reflecting a growing trend in the use of LLMs in mental health dialogue systems. There has also been an increase in keywords such as ‘emotional support’ and ‘emotional support conversation,’ as described in Figure 4.

Figure 4: Yearly Keyword Co-occurrence Network Graphs (2020–2024)

## Qualitative Trend Analysis

### Overview of Highly Cited Top 10 Publications

As presented in Table 6, Scopus and WoS reported over 800 and 60 annual citations of published papers, respectively (as of May 7, 2024). Aligned with the increasing publication trends shown in Table 1, the annual citation count has also demonstrated a consistent upward trend. Note that publications from ACM were excluded due to the unavailability of citation count data.

Table 6: Number of Publication Citations per Year in Scopus and WoS

| Year | SCOPUS | WOS |
|---|---|---|
| 2020 | 30 | 22 |
| 2021 | 177 | 63 |

| | | |
|---|---|---|
| 2022 | 378 | 88 |
| 2023 | 722 | 90 |
| 2024 | 806 | 68 |
| Total | 2,113 | 331 |

We then performed a qualitative analysis of the top 10 most frequently cited papers. In particular, we examined (i) the psychological perspective, focusing on aspects such as the ultimate goals for mental health improvement and the psychological approaches integrated into deep learning models. Additionally, we evaluated (ii) computational strategies, including the datasets utilized in the development of dialogue systems, the methodologies and theoretical frameworks applied in counseling systems, and the evaluation metrics used to assess their effectiveness. Table 7 provides a comprehensive summary of the top 10 most highly cited publications.

Table 7: Overview of research methodologies utilized in highly cited publications.

| Reference | Year | Task | Psychological Background | Computational Approach | Dataset | Evaluation Metrics |
|---|---|---|---|---|---|---|
| Majumder *et al.* [74] | 2020 | Empathetic Response Generation | Empathy is expressed by mirroring the emotions of the other person [12]. | ELBO, Multi-task learning | Empathetic Dialogue [91] | (Automatic) BLEU (Human eval) Empathy, Relevance, Fluency |
| Li *et al.* [64] | 2020 | Empathetic Response Generation | Responses to the current turn inherently include feedback on the previous turn [124] | Discriminator, Multi-task learning | Empathetic Dialogue [91] | (Automatic) ACC, PPL, DIST-1, DIST-2 (Human eval) Empathy, Relevance, Fluency |
| Lin *et al.* [67] | 2020 | Empathetic Response Generation | - | Pretrain, Multi-task learning | Empathetic Dialogue [91], PersonaChat [123] | PPL, AVG-BLEU, EMO-ACC |
| Sharma *et al.* [96] | 2021 | Empathetic Rewriting | Empathetic interactions are a key factor in improving an individual's mental health [31]. | Reinforcement Learning | TALKTOME community | Change in empathy, PPL, Sentence coherence, Extrema, DIST-1, DIST-2, Edit rate |
| Liu *et al.* [69] | 2021 | Emotional Support Conversation | Selecting appropriate strategies by considering the stages in the context of emotional supportive conversations [44] | Generating strategy token | ESConv [69] | BLEU-2, ROUGE-L, Extrema |
| Sabour *et al.* [95] | 2022 | Empathetic Response Generation | Empathy is a broad construct encompassing both affective and cognitive components, with cognition playing a role in understanding situations and emotions [23]. | External knowledge (COMET) | Empathetic Dialogue [91] | ACC, PPL, DIST-1, DIST-2 |
| Li *et al.* [65] | 2022 | Empathetic Response Generation | - | External knowledge (NRC-VAD, ConceptNet), | Empathetic Dialogue [91] | (Automatic) ACC, PPL, DIST-1, DIST-2 (Human eval) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | GNN with attention | | Empathy, Relevance, Fluency |
| Wang *et al.* [112] | 2021 | Emotional Support Response Generation | - | CNN, seq2seq with attention | YouBao health community posts | (Automatic) ACC, BLEU (Human eval) Grammar, Relevance, Correctness, Willing-to-reply, Emotional Support |
| Kim *et al.* [54] | 2021 | Empathetic Response Generation, Emotion Recognition | Perspective-taking is a key component in empathetic reasoning [24]. | Baysian Inference | Empathetic Dialogue [91] | (Automatic) TOP-1/3/5 Recall, EPITOME-Exploration/ Interpretation, (Human eval) Empathy, Relevance, Fluency |
| Shin *et al.* [98] | 2020 | Empathetic Response Generation | - | Reinforcement Learning | Empathetic Dialogue [91] | DIST-1, DIST-2, DIST-3, AVG-BLEU, Extrema |

### Psychological Approach for Generating Empathetic Responses

In this section, we reviewed psychological approaches to developing mental health dialogue systems. Kim et al. [54], for example, aimed to improve empathetic engagement in dialogue systems by incorporating perspective-taking, a psychological process that facilitates understanding situations from others' viewpoints. By employing the Rational Speech Acts (RSA) framework [35], a probabilistic model that views communication as a recursive reasoning process, they utilized Bayesian network inferences with target words tied to emotional causes. This iterative approach allowed for inferring intentions and beliefs, enhancing the system's capacity to produce empathetic responses. Majumder et al. [74] also proposed a model that generates empathetic responses based on whether the emotions in prior responses are positive or negative, building on the idea that empathy involves mirroring another person's emotions [12]. Highlighting the importance of accurately understanding another person's circumstances and emotional states for effective empathy, Sabour et al. [95] employed Commonsense Transformers (COMET) [8], which leverage a commonsense knowledge graph to produce contextually relevant inferences about events, actions, and emotional states. This approach facilitates a more precise and nuanced comprehension of human behavior and interactions, enabling dialogue systems to respond with greater empathy and contextual awareness. Moreover, Sharma et al. [96] introduced a reinforcement learning model that rewards performance based on fluency and consistency, which are critical factors in enhancing empathetic communication. Both Li et al. [64] and Shin et al. [98] develop models that integrate feedback or sentiment from user responses to improve empathetic interactions, highlighting the importance of tracking client responses in fostering successful empathy. In addition, drawing on empirical evidence that social media platform users can be categorized as information-seeking or emotion-seeking types, Wang et al. [112] first classified user types using a CNN module and provided empathetic responses only to emotion-seeking users. Liu et al. [69] emphasized that counselors deliver emotional support by using psychological counseling strategies based on the context and information shared by clients. Consequently, the researchers designed the decoder to produce a specialized strategy token aligned with the prior conversational context, followed by generating a response conditioned on this token. Human evaluation outcomes indicated that this method significantly improved the model's emotional support quality, particularly in the areas of Fluency, Identification, Suggestion, and Comforting.

### *Computational Approach for Generating Empathetic Responses*

**Training Datasets:** We observed that the main objective of dialogue systems for mental health improvement is influenced by the choice of dataset. To be specific, our analysis revealed that numerous studies employed the publicly available Empathetic Dialogue dataset [91], comprising 25,000 conversations designed to address the emotional cues of dialogue partners. On the other hand, some studies constructed their own datasets for training deep learning models. For instance, Wang et al. [112] collected post-response pairs from a pregnancy healthcare community developed to provide informational and emotional support to pregnant women. Likewise, Sharma et al. [96] collected posts and responses from TalkLife, an online peer-to-peer support platform, which were then refined by human experts to improve empathetic quality, converting interactions with lower empathy into those with higher empathy. Although only one of the ten studies leveraged the ESConv (Emotional Support CONVersation) dataset [69], we will delve into the ESConv dataset and related studies separately in a subsequent section. This dataset has recently garnered attention for its applicability, as its dialogue sets are annotated according to psychological counseling processes [44].

**Deep Learning Models:** We noticed that most studies aimed to develop models for generating empathetic responses and built their models on the decoder architecture of transformers [108]. Furthermore, two studies employed reinforcement learning models that utilized sentiment intensity [98] and fluency and coherence [96] as reward signals to promote the generation of empathetic responses. Remarkably, external knowledge sources were utilized to fill gaps in domain knowledge. For example, Li et al. [65] incorporated NRC-VAD [8] to support the understanding of emotional tone; this resource provides human ratings for over 20,000 English words in terms of valence, arousal, and dominance. Similarly, Sabour et al. [95] used COMET [8], a commonsense knowledge graph designed to generate inferences aligned with the context of events, actions, and emotional states, to enhance the interpretation of emotions and contexts. Kim et al. [54] adopted Bayesian Inference to adjust prior beliefs based on observed data, such as emotion-inducing words, enabling the model to dynamically enhance its understanding of emotional triggers. This approach improved the contextual relevance and specificity of the empathetic responses generated.

Owing to the recent advancements in techniques, publications utilizing LLMs in dialogue systems were not included among the top 10 papers. However, given their strong performance in language generation, we will discuss LLMs and related studies individually.

**Evaluation Metrics:** We found that there is a wide range of metrics used to evaluate the performance of empathetic response generation, with no standardized approach. The description of each metric is provided in Table 8. Specifically, many studies utilized statistical automatic metrics such as ACC [64, 95, 65], DIST [64, 96, 95, 65, 98], BLEU [74, 69, 112, 98, 67], and PPL [64, 96, 69, 95, 65, 67]. Moreover, by utilizing pretrained language models, researchers have been able to evaluate a broader range of factors. For instance, Shin et al. [98] employed Bag of Word Embedding Similarity [68] to assess inter-sentence similarity, while Kim et al. [54] used a trained RoBERTa model [70] to measure counseling strategies, such as exploration and interpretation, to capture empathetic attributes.

Nevertheless, automatic metrics like BLEU have been found to correlate weakly with human assessments of response quality [68]. Thus, human evaluation methods were adopted to measure complex empathetic responses more accurately [74]. Annotators evaluated general responses across dimensions such as ‘Empathy,’ ‘Relevance,’ and ‘Fluency’ [54, 64, 74, 65]. In a similar effort, Wang et al. [112] assessed human and AI responses based on human judgments concerning ‘Grammar Correctness,’ ‘Relevance,’ ‘Willingness to Reply,’ and ‘Emotional Support.’ A/B testing was also frequently employed to determine which model appeared more empathetic and human-like [54, 64, 74, 69].

Table 8: Explanation of Evaluation Metrics for Automatic and Human Assessment

| Approach | Metric | Description |
|---|---|---|
| Automatic Evaluation - Statistical | **ACC** | Accuracy; Measure the percentage of correct predictions |
| | **PPL** | Perplexity; Evaluate how well a model predicts a sequence; lower is better |
| | **BLEU-N** | Score text similarity based on n-gram matches (N refers to the length of the n-grams used to calculate the overlap) |
| | **DIST** | Measure diversity of generated text based on unique n-grams |
| | **METEOR** | Measure quality by assessing accuracy, fidelity, word order, and lexical diversity |
| | **ROUGE-L** | Evaluate longest common subsequence between texts (L denotes the longest common subsequence) |
| | **CIDEr** | Evaluate text similarity using n-gram TF-IDF scores to emphasize important words |
| | **Extrema** | Measure semantic similarity by comparing the highest word embedding values between generated and reference |
| | **NIDF** | Normalized Inverse Document Frequency; Measure response informativeness by calculating the rarity of specific words or phrases within a large dataset |
| | **EMOACC** | Measure emotion accuracy about empathetic response generation |
| | **Edit Rate** | Calculate the number of modifications made in the rewritten response compared to the original, indicating the precision and conciseness |
| Automatic Evaluation - Pretrained Model | **cES** | Conversation Emotional Support; Measure the percentage of correct predictions. |
| | **tES** | Turn-level Emotional Support; Evaluate how well a model predicts a sequence |
| | **cDC** | Context Dialogue Coherence; Evaluate coherence with the context |
| | **fDC** | Future Dialogue Coherence; Evaluate coherence with the user’s future utterance |
| | **BERTScore** | Evaluate text quality by measuring semantic similarity between predicted and reference sentences using token embeddings. P, R, and F mean Precision, Recall, and F1-score, respectively. |
| | **BARTScore** | Assess text generation quality by using a pretrained BART model to score the likelihood of generated text given the reference |
| | **INTENTACC** | Measure the response intent accuracy using a fine-tuned BERT model on the EMPIN-TENT dataset [116] |
| | **Emotion Reaction** | Expressing emotions such as warmth, compassion, and concern, experienced by response |
| | **Exploration** | Improving understanding of the user by exploring the feelings |
| | **Interpretation** | Communicating an understanding of feelings and experiences inferred from the user’s utterance |
| | **Empathy** | Measure the increase or decrease in empathy levels in rewritten responses compared to the original |

| | | |
|---|---|---|
| | **Sentence Coherence** | Calculate the number of modifications made in the rewritten response compared to the original, indicating the precision and conciseness |
| **Human Evaluation** | **Identification** | Which bot better analyzed your situation and identified your problems? |
| | **Comforting** | Which bot was more effective in providing comfort? |
| | **Suggestion** | Which model offered more useful advice? |
| | **Fluency** | Which bot's responses were clearer and more natural? |
| | **Knowledge** | The extent to which useful knowledge is provided |
| | **Empathy** | Which model showed more suitable emotional responses, like warmth and concern? |
| | **Coherence** | Which bot's response better fits the context across turns? |
| | **Supportiveness** | Which bot was more effective at shifting the user's emotions positively? |
| | **Informativeness** | Which bot's response was more varied, detailed, and informative? |
| | **Naturalness** | Measure how smooth and natural the model's response feels in conversation |
| | **Realism** | Measure how closely the agent interaction aligns with authentic human conversation characteristics |
| | **Valence** | Evaluate the positivity or negativity of the emotional response elicited by the agent, reflecting the overall emotional tone |
| | **Arousal** | Measure the level of physiological and emotional activation elicited by the agent, ranging from calm to excited states |
| | **Veracity** | Measure whether the response generated by the system is correct and relevant to the question |
| | **Evidence** | Measure whether response by the system includes references to relevant studies, clinical trials, or guidelines that can validate it |
| | **Helpfulness** | Assess the relevance and usefulness of the model's response to the user's needs |
| | **Rapport** | Measure the connection established between clients and counselors in empathetic dialogs |
| | **Relevance** | Measure the degree to which a response is contextually aligned with the thematic content of the preceding dialogue |
| | **Safety** | Measure whether the model's response avoids harmful, offensive, or legally sensitive content |

## Overview of Emotional Support Conversation (ESConv)

Providing emotional support is an essential skill in mental health interventions, aiming to alleviate emotional distress and assist individuals in navigating the challenges they encounter [58]. In line with this, Liu et al. [69] developed the Emotional Support Conversation dataset (ESConv), which consists of dialogues between trained crowd workers acting as supporters and help-seekers, who were required to complete a pre-chat survey about their problems and emotions, as well as provide feedback during and after the conversations. As described in Figure 5, each conversational turn in ESConv is annotated by selecting the appropriate strategy and corresponding stage in the counseling process based on principles from Helping Skills Theory [44]. This approach mirrors that of trained psychological counselors, who select strategies and stages based on the context of the conversation [44]. Moreover, the dataset includes the help-seeker's final emotional intensity, as well as assessments of the supporters' empathy and the relevance of their responses following each conversation.

Figure 5: We present figures from the original paper [69] to illustrate three stages of the ESConv framework and the corresponding eight counseling strategies. According to Liu *et al.* [69], this framework comprises three stages, each with specific support strategies. The *exploration* stage aims to help individuals identify underlying issues; the *comforting* stage focuses on providing empathy and understanding; and the *action* stage involves offering practical information or suggestions. Typically, the emotional support process follows a sequential order from 1. Exploration → 2. Comforting → 3. Action, as indicated by black arrows, can also be adjusted to suit the conversation's needs, as represented by dashed gray arrows.

The ESConv dataset is highly aligned with psychological counseling processes and has been widely cited, with 215 citations on Google Scholar[1]. Recognizing its importance, we conducted a comprehensive review of eight studies that leveraged this dataset, highlighting at least one citation from the selected studies. The summary of selected papers is described in Table 9.

Table 9: Overview of research methodologies employed in publications utilizing the ESConv [69].

| Reference | Year | Psychological Background | Computational Approach | Automatic Evaluation Metrics | Human Evaluation Metrics |
|---|---|---|---|---|---|
| Liu *et al.* [69] | 2021 | Emotional support through strategy selection [44] | Multi-task learning | BLEU-2, ROUGE-L, Extrema | Fluency, Identification, Comforting, Suggestion |
| Peng *et al.* [83] | 2022 | Understanding the cause of seeker's problem and the intention of the seeker [89] | External knowledge (COMET), Hierarchical Graph, Attention Network, Multi-task learning | PPL, BLEU-4, DIST-2, ROUGE-L | Fluency, Identification, Comforting, Suggestion |
| Tu *et al.* [106] | 2022 | Enhancing the comprehension of the seeker's emotion by utilizing external knowledge | External knowledge (COMET), Cross-attention, Multi-task learning | ACC, PPL, DIST-2, BLEU-4, ROUGE-L, METEOR | Fluency, Knowledge, Empathy |

[1] Accessed on October 24, 2024

| | | | | | |
|---|---|---|---|---|---|
| Peng *et al.* [85] | 2023 | Considering seeker's feedback when selecting strategy | Gate mechanism, Strategy dictionary, Multi-task learning | ACC, PPL, DIST-2, BLEU-4, ROUGE-L, METEOR | Fluency, Identification, Comforting, Suggestion |
| Cheng *et al.* [17] | 2022 | Long-term strategy planning through forecasting the future's seeker's feedback | Emotion cause detection, External knowledge (ERC-VAD), A* algorithm | PPL, BLEU-4, ROUGE-L, METEOR, CIDEr | Fluency, Identification, Comforting, Suggestion |
| Deng *et al.* [26] | 2023 | Capturing the transition of strategy, seeker's emotion, conversation's semantic | External knowledge (COMET, HEAL), Graph Retrieval | PPL, BLEU-4, ROUGE-L | Fluency, Identification, Comforting, Suggestion |
| Zhou *et al.* [134] | 2023 | Evoking the seeker's emotion intensity | MoE, External knowledge (COMET, NRC-VAD), Reinforcement Learning | PPL, BLEU-2 DIST-2, cES, tES, cDC, fDC, | Fluency, Informativeness, Coherence, Supportiveness |
| Zhao *et al.* [129] | 2023 | Mixed-Initiative of AI for Emotional Support | External knowledge (COMET, ATOMIC), State Transition Graph, Cross-attention, Multi-task learning | ACC, PPL, DIST-2, BLEU-4, ROUGE-L | Fluency, Identification, Suggestion, Empathy |

### Psychological Approach for Emotional Support

We discovered that, despite using the same dataset, selected studies adopt differing psychological perspectives on improving emotional support, which has resulted in the development of diverse approaches. Numerous studies highlighted the importance of understanding an individual's psychological state for delivering effective emotional support. For instance, Tu et al. [106] enhanced comprehension of the help-seeker's emotions by utilizing the commonsense knowledge graph dataset, COMET [8]. Peng et al. [84] developed a hierarchical graph network to capture both the seeker's overarching concerns throughout the conversation and the peripheral intentions within each utterance, suggesting that emotional support can be enhanced by understanding the root of the problem. Similarly, Zhou et al. [134] highlighted that the primary aim of emotional support is to evoke the user's emotions, leading them to use the emotion intensity factor in ESConv as the reward function within the Reinforcement Learning framework.

Conversely, a few studies emphasized the significance of selecting appropriate strategies to improve the emotional satisfaction of help-seekers. Liu et al. [69] proposed that strategic decision-making plays a crucial role in providing emotional support, demonstrating the effectiveness of identifying anticipated strategies in generating empathetic responses. Peng et al. [85] incorporated user feedback into strategy selection by reinforcing or discouraging specific counseling strategies based on the help-seeker's responses throughout the conversation, resulting in more accurate and user-aligned strategy predictions. Furthermore, Cheng et al. [17] underscored the importance of long-term strategy planning, prompting them to forecast future user feedback and optimize emotional support strategies across multiple conversational turns.

Although the ESConv dataset has shown promise in supporting the development of counseling dialogue systems for emotional support, it remains insufficient for capturing real-world scenarios and the full range of counseling strategies necessary for developing comprehensive psychological counseling systems. We will address the limitations of datasets for counseling system development in the discussion section.

### Computational Approach for Emotional Support

**Deep Learning Models:** We identified two main approaches commonly utilized in the selected papers: (i) multi-task learning and (ii) the integration of external knowledge resources. First, several studies adopted multi-task learning, wherein a single model is trained on multiple tasks concurrently to enhance performance through shared representations. For example, Peng et al. [84] enabled the model to simultaneously generate responses and predict the type of issue presented by the help-seeker, while other studies allowed the model to generate responses and identify the counseling strategies applied by the supporter [69, 106, 85]. Additionally, the joint prediction of keywords and emotions was also investigated [134, 129]. As related tasks can significantly improve model performance by facilitating knowledge transfer between tasks [33], we noticed that many studies leveraged the training of similar tasks together to maximize these benefits.

Second, external knowledge sources have also been incorporated to address the limitations of domain knowledge in training data and models. For instance, Deng et al. [26] employed HEAL [117], a knowledge graph specifically designed for mental health conversations. COMET [8] has also been utilized to infer users' emotions [106, 134, 129], intentions, and underlying psychological causes [84]. In addition, NRC-VAD [76] has been used for emotion detection [17, 134].

Furthermore, we examined the underlying backbone models adapted to implement those approaches. Zhao et al. [129] introduced a State Transition Graph (STG) [80] to represent the dynamic behavior of directed graphs. Using this approach, they tracked semantic, emotional, and strategic transitions throughout conversations by constructing a separate graph for each stage of counseling. Likewise, Peng et al. [84] built a hierarchical graph attention network to model the relationships among the global cause, local intention, and dialogue history. Tu et al. [106] suggested a strategy probability distribution method to select subsequent counseling strategies, mapping the probability of each strategy's selection to a discrete latent space. This approach allows the model to consider multiple strategies in a dynamic rather than fixed manner, facilitating the generation of responses that are both supportive and contextually relevant. Furthermore, Cheng et al. [17] applied the A* algorithm for strategy planning, an optimal path-finding algorithm that calculates the shortest path using both cost and heuristic functions [42]. Zhou et al. [134], in addition, implemented a Mixture of Experts (MoE) architecture for emotion and keyword predictions, combined with reinforcement learning that uses emotion intensity and coherence as reward signals. MoE is a neural network architecture designed to improve efficiency and performance by routing inputs to specialized sub-models [49]. Finally, Peng et al. [85] managed the selection of subsequent strategies and context representation through a gating mechanism, a trainable component that regulates information flow by learning to selectively allow or block signals through multiplicative operations [38].

**Automatic Evaluation Metrics:** A variety of statistical evaluation metrics have been applied to assess the performance of emotional support systems, including PPL [69, 83, 106, 85, 17, 26, 134, 129], BLEU [69, 83, 106, 85, 17, 26, 134, 129], ROUGE-L [69, 83, 106, 85, 17, 26, 129], DIST [83, 106, 85, 134, 129], METEOR [17], CIDEr [17], Extrema [69], and ACC [106, 85, 129]. Zhou et al. [134] introduced evaluation metrics—cES and tES for assessing emotional elicitation intensity, and cDC and fDC for measuring response coherence. These metrics were evaluated using the pre-trained emotion classification model DistilRoBERTa [43] along with BERT [27].

Detailed explanations of these metrics can be found in Table 8.

However, our findings indicate that the evaluation metrics predominantly assess fluency and accuracy, often overlooking the semantic richness of responses. These metrics are notably limited in their ability to capture semantic equivalence between sentences that differ in wording yet express similar meanings [72]. To more accurately evaluate the semantic content of responses, it is essential to incorporate diverse and semantically aware evaluation methods, such as BERTScore [125], which leverages cosine similarity to align sentence embeddings.

**Human Evaluation Metrics:** To effectively assess emotional support conversations, it is essential to incorporate human evaluation [69], which may provide valuable insights into subtle elements such as empathy, rapport, and perceived helpfulness that automated systems may not fully capture. Our analysis found that most publications considered human A/B evaluations, which compare responses from the target model against a baseline to determine which is superior. The factors evaluated to determine the quality of emotional support included Suggestion [69, 85, 17, 134, 129, 26, 83], Identification [69, 83, 17, 85, 129, 26], Empathy [17, 129], Informativeness [134], Coherence [134], and Supportiveness [134]. In contrast, Tu et al. [106] assembled three experts with backgrounds in linguistics or psychology to independently assess Fluency, Knowledge, and Empathy, rating each on a scale from 0 to 2. Detailed descriptions of each evaluation aspect can be found in Table 8.

## Overview of Publications using Large Language Models (LLMs)

LLMs have demonstrated excellence in engaging in human-like interactions and following instructions to provide contextually relevant feedback [29]. These abilities make LLMs suitable not only for general applications but also for specialized domains such as mental health [59], particularly in counseling systems [79]. Their capability to manage complex, multi-turn dialogues allows for nuanced, empathetic, and adaptive interactions [7, 29].

Here, we investigated the applications and limitations of LLMs within dialogue systems that emulate counseling environments. To this end, we filtered papers from the entire collection that included specific author keywords, such as 'large language models,' 'GPT-3,' 'ChatGPT,' 'GPT-4,' and 'GPT-3.5.' This filtering process yielded a total of 10 papers for examination, as illustrated in Table 10.

Table 10: Overview of research methodologies applied in studies utilizing LLMs.

| Reference | Year | Task | Psychologic Background | Computation Approach | LLM | Dataset | Automatic Evaluation Metrics | Human Evaluation Metrics |
|---|---|---|---|---|---|---|---|---|
| Lai *et al.* [57] | 2024 | Mental Health QA | - | Pretraining on corpus, Finetuning | WenZhong [113], PanGU [122] | PsyQA [101] | PPL, DIST-1, DIST-2, ROUGE-L | - |
| Llanes-Jurado *et al.* [71] | 2024 | Empathetic Response Generation | - | Providing context with GPT-3 followed by response generation | GPT-3 [10] | - | - | Naturalness, Realism, Valence, Arousal |
| Kharitonova *et al.* [53] | 2024 | Mental Health QA | - | Retrieval Augmented Generation (RAG) | GPT-3 [10], Llama [103], | Synthetic QA from LLM | - | Coherence, Varacity, Evidence |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Llama-2 [104], | | | |
| Kaysar and Shiramatsu [52] | 2023 | Providing suggestion for mental health problem | - | Natural Language Understanding (intent, emotion), Finetuning | GPT-3 [10] | Customized Conversational Datasets | BLEU, ROUGE | - |
| Firdaus *et al.* [34] | 2023 | Empathetic Response Generation | - | Few-shot learning | DialogGPT [126] | DailyDialog [63], EmotionLines [47], EmoWOZ [32] | BLEU, ROUGE-L | - |
| Lee *et al.* [61] | 2022 | Empathetic Response Generation | Expressing empathy requires emotional and cognitive insights [23] | Few-shot learning | GPT-3 [10] | Empathetic Dialogues [91] | DIST-2, NIDF, PPL, INTENTACC, EMOACC, Interpretation, Exploration, Emotion Reaction | - |
| Zhang *et al.* [127] | 2023 | Emotional Support Conversation | - | Providing knowledge from GPT-3.5 as context | GPT-3.5 [94] | ESConv [69], BlendedSkillTalks [93] | BLEU-4, ROUGE-L, BERTScore, BARTScore | - |
| Chen *et al.* [15] | 2024 | Emotional Support Conversation | - | Finetuning, Voting | CHATGLM2-6B [37] | ESConv [69](en) | BLEU-2, BLEU-4 DIST-1, DIST-2 | Empathy, Coherence, Helpfulness, Rapport |
| Qian *et al.* [87] | 2023 | Empathetic Response Generation | - | Few-shot learning, Using knowledge-base for context | GPT-3 [10], GPT-3.5 [94], CHATGPT [94] | Empathetic Dialogues [91] | Dist1, Dist2, P-BERTScore, R-BERTScore, F-BERTScore, BLEU-2, BLEU-4 | Fluency, Identification, Empathy, Coherence |
| Chen *et al.* [16] | 2023 | Counseling Data Augmentation | - | Rewriting using ChatGPT, Finetuning | CHATGLM-6B [37] | SoulChatCorpus SMILECHAT [88] | BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L | Naturalness, Empathy, Helpfulness, Safety |

## *Types and Attributes of Applied LLMs*

In this section, we examined the specific LLMs utilized in each study and their distinct characteristics. Our findings revealed that the choice of LLM often varies based on its attributes and the intended application purpose. Particularly, GPT-3 [10] was the most prevalent LLM, applied in five of the ten papers [71, 53, 52, 87, 61], while two papers [127, 87] utilized GPT-3.5 [94]. GPT-3 [10], an autoregressive language model with 175 billion parameters, is trained on a comprehensive 570GB dataset to generate human-like text. This capability allows it to excel in tasks such as text generation, translation, summarization, and question-answering. GPT-3.5 [94], an upgraded version of GPT-3, is developed to advance language comprehension and generation abilities, delivering increased accuracy, efficiency, and adaptability across a range of natural language tasks. We posit that the selection of these models is likely due to their superior generative abilities and ease of use. In contrast, Kharitonova et al. [53] chose the open-source models Llama [103] and Llama2 [104], which support local customization, to mitigate potential cost and privacy issues associated with the closed-source GPT series. Meanwhile, LLMs fine-tuned for specific languages, such as Chinese, were exploited, including CHATGLM-6B [37], WenZhong [113], and PanGu [122].

### Advanced Techniques for Optimizing LLM Effectiveness

We studied the specialized techniques that distinguish LLMs from previous deep learning models.

Our analysis identified zero-shot and few-shot learning as the frameworks that are predominantly utilized for LLMs. These approaches involved prompting methods that provide the model with either no examples or a few examples of questions and answers, enabling it to effectively address similar tasks. Qian et al. [87] supplied GPT-3 with random examples of empathetic responses and relevant background knowledge to support empathetic response generation. Likewise, both Lee et al. [61] and Firdaus et al. [34] leveraged few-shot learning to produce empathetic responses, with the former focusing on aligning responses closely with the input query and the latter enhancing response empathy through emotion recognition. Due to the generalizability of LLMs, they can attain high performance without further training, offering advantages over traditional deep learning models that demand vast amounts of training data [10] — a notable limitation in the mental health domain due to data scarcity.

An alternative approach is Retrieval-Augmented Generation (RAG), a technique that integrates information retrieval with language generation. RAG initially retrieves pertinent information from a knowledge base and subsequently employs the LLMs to generate a response informed by this retrieved content [62]. Kharitonova et al. [53] applied this approach by retrieving knowledge items to respond to input queries in a psychological context. Comparably, Chen et al. [15] implemented ChatGPT to produce multiple response candidates and used its reasoning capabilities to select the most suitable strategy for providing emotional support.

### Limitations of LLMs in Psychological Counseling Applications

We identified major limitations of LLMs as psychological counseling models and examined strategies to address these limitations.

***Hallucination*** emerged as a primary issue, defined as the phenomenon in which LLMs produce incorrect or irrelevant outputs in response to given inputs [48, 109]. To reduce hallucination, additional information was supplied to support the generation of appropriate responses, such as incorporating emotionally similar contexts to align with input data [61, 34]. Furthermore, Qian et al. [87] trained the LLM using random contexts, enabling it to encounter and respond effectively to a range of scenarios. In comparison, Kharitonova et al. [53] essentially assessed the risk of hallucination in LLMs by integrating a separate knowledge base. This approach enables the LLM to infer from carefully selected scenarios, thereby supporting the generation of safe responses.

***Limited mental health-related knowledge*** represents another critical limitation, as recent LLMs often demonstrate unreliability or inconsistency [1], potentially due to inadequate understanding of mental health domains [119]. Therefore, exploring approaches to strengthen the domain knowledge of LLMs is essential. We confirmed that only Lee et al. [61] developed their model by considering the psychological foundations, understanding that empathetic responses require awareness of both the user's emotional and situational context. They applied few-shot learning techniques, using examples that closely align with similar emotional and situational characteristics. Alternatively, Domain adaptation through prompting or fine-tuning was also applied to facilitate the effective use of LLMs as counseling models, primarily due to their limited exposure to relevant data during pretraining [114, 40]. For example, prompting techniques are employed to initially identify the interlocutor's emotions [52, 34, 87] or intentions [52] to establish contextual understanding. This approach enables the model to condition responses based on the detected emotions or intentions, resulting in more appropriate replies. Additionally, Lai et al. [57] trained their model on a comprehensive psychology corpus and fine-tuned it using PsyQA [101] to enhance adaptation to the psychological domain. PsyQA is a Chinese dataset designed for generating long-form counseling responses in mental health support, consisting of 22,000 questions and 56,000 structured answers. Impressively, Zhang et al. [127] underscored the knowledge limitations of smaller LLMs and positioned the LLM as a knowledge expert, given its extensive knowledge base acquired through diverse sources during training [135]. To facilitate emotional support, they prompted the LLM to inquire about the seeker's emotions, underlying causes, and potential solutions, subsequently using this information as contextual input to generate supportive responses.

Lastly, to address ***the scarcity of counseling datasets***, Chen et al. [16] introduced SoulChat, utilizing ChatGPT's text rewriting capabilities. By employing prompt-based transformations, they converted an initial dataset comprising 215,813 single-turn psychological counseling questions across 12 topics, along with 619,725 paired responses, into multi-turn conversations. This approach ultimately yielded a Chinese-language multi-turn empathetic conversation dataset containing 2,300,248 samples.

## Discussion

This section addresses the implications and recommendations for advancing mental health counseling systems. Additionally, it explores the study's limitations and outlines directions for future research.

### Implications & Recommendations

#### *The Lack of Training Dataset for Counseling System Development*

The difficulty of developing open-access datasets that represent the psychological counseling process remains a significant obstacle to advancing research in counseling system development. Our review indicates that existing studies have a constrained focus, often due to limited available datasets, when it comes to addressing the comprehensive counseling process. For example, among the ten most-cited papers, seven center exclusively on creating models intended to generate empathetic responses using the Empathetic Dialogue dataset [91], while many recent studies focus on providing emotional support through the ESConv dataset [69].

Above all, the collection and annotation of counseling datasets is not only time-intensive and costly but also requires professional expertise [36]. Furthermore, due to confidentiality principles, counseling data cannot be shared externally without explicit client consent [121, 73]. Since such data often contains personally identifiable information, anonymization through data preprocessing presents additional challenges [136]. In line with this, while ESConv is designed to simulate mental health counseling interaction, it encompasses only a narrow range of strategies, excluding techniques like [51] and Confrontation [75]—methods used by trained counselors to enhance therapeutic effectiveness—due to the lack of professional supervision and sufficient training resources. This constraint can lead to challenges in incorporating real-world scenarios into the model.

To mitigate data scarcity, recent studies have increasingly aimed to apply LLMs to generate counseling datasets that encompass a wider variety of counseling strategies. Here, we provide a brief overview of several studies published after May 7, 2024, which were not included in our analysis. To expand the dataset, Zheng et al. [130] initially fine-tuned GPT-J 6B [111] on 100 samples from the ESConv dataset and then leveraged the model to generate emotional support conversations by responding to dialogues from the Empathetic Dialogue dataset. This process yielded the AUGESC dataset, comprising 65,000 sessions across diverse topics. Experimental results and human evaluations confirmed that the dataset is non-toxic and closely emulates genuine emotional support conversations. Zheng et al. [132] utilized ChatGPT [81] to augment a conversation dataset by iteratively generating new data, using the ESConv dataset as a seed. This approach yielded the EXTES dataset, which includes 11,000 dialogues spanning 36 varied scenarios and 16 unique helping strategies. Human evaluations demonstrated that the quality of this dataset is comparable to that of the original ESConv dataset. Interestingly, Zhang et al. [128] converted 3,134 high-quality psychological counseling reports into multi-turn consultation dialogues using a two-phase approach. In the first phase, counseling reports were transformed into clinical notes framed from the perspective of a psychological supervisor, offering guidance for the counselor. Based on these notes, simulated dialogues between a psychological counselor and client were then generated. The final dialogue dataset achieved high ratings in terms of Comprehensiveness, Professionalism, Authenticity, and Safety.

### *The Need for Psychological Insights in Developing Counseling Systems*

In psychological counseling, it is essential to explain the underlying causes of a client's issues and develop tailored strategies for addressing them. This process is grounded in extensive psychological knowledge and the specialized training of counselors [107, 44]. However, our review indicated that most studies prioritized accuracy improvements through computational techniques over the integration of clinical insights and psychological knowledge, both crucial for practical application and usability. In other words, while current studies aimed at improving mental health, such as emotional support or generating empathetic responses, they have generally neglected to align their frameworks and underlying models with foundational psychological knowledge. Moreover, several studies have drawn on empirical findings rather than psychological insights. For instance, Tu et al. [106] underscored the role of fine-grained emotion detection in emotional support based on empirical evidence alone. Likewise, Cai et al. [11] prioritized commonsense knowledge over psychological expertise to enhance the system's comprehension of implicit social and emotional contexts.

In contrast, Liu et al. [69] developed a dialogue dataset and trained a model that simulated the selection of context-sensitive conversational strategies, closely aligning actual counseling practices used by therapists. This approach, grounded in psychological knowledge, produced models that achieved high preference ratings in human evaluations.

Therefore, we propose that AI models for mental health improvement, particularly for interdisciplinary research and practical applications, be designed with a foundation in psychological knowledge. This integration is essential for accurately understanding individuals' psychological states and can greatly enhance the empathy quality exhibited by counseling systems [28].

### Comprehensive Evaluation Metrics in Psychological Counseling Systems

Evaluating the effectiveness of psychological counseling systems requires a careful examination of nuanced factors like empathy [20], rapport [30], and perceived helpfulness [82], all of which are critical for successful counseling [5]. However, we found that many studies have concentrated on outcome accuracy using automatic evaluation metrics, which frequently fall short in capturing the subjective quality of responses as effectively as human assessments [68]. While a growing number of studies have adopted human evaluations to overcome these limitations, such methods exhibit significant drawbacks. Human evaluations are time-consuming and lack consistency, as their results may vary depending on the evaluators involved, complicating comparisons across different counseling systems [46, 99]. Additionally, they are prone to biases that may favor types of responses [77].

Interestingly, recent research has explored the potential of using LLMs as evaluators in psychological counseling contexts [131]. For instance, Zhang et al. [128] used GPT-4 to assess semantically complex factors, including comprehensiveness, professionalism, authenticity, and safety. Moreover, Kang et al. [50] examined biases in LLM-based dialogue systems toward specific emotional support strategies to achieve a balanced approach.

Nevertheless, there remains an unmet need, suggesting that future directions for evaluation metrics should focus on standardization and incorporate a wider array of psychological counseling dimensions.

### Challenges of LLMs in Developing Personalized Counseling Systems

Effective psychological counseling generally involves multiple sessions, making it essential to monitor and retain details of the client's emotional state, experiences, and relevant events throughout the therapeutic journey to enable personalized counseling [19]. Although personalization in chatbots has also demonstrated the potential to enhance therapy outcomes [110], most current psychological dialogue systems are often trained on single-session datasets, limiting their capacity to provide personalized therapy.

A primary challenge with LLMs lies in their limited context length [115], which restricts their ability to retain personalized events and information across extended, ongoing counseling sessions. In line with this, Zhong et al. [133] introduced MemoryBank, a method designed to enhance the long-term memory capabilities of LLMs by enabling recall of prior interactions and adaptation to user personality traits.

### Ethical Considerations in the Development of Safe Systems

For the safe use of AI counseling systems, it is pivotal to address ethical considerations. Risks include potential privacy infringements and leakage of personal information during both training and inference stages [78]. Furthermore, algorithmic biases and limitations in data may lead to culturally insensitive care or the dissemination of misinformation [90], or the generation of psychologically harmful content [102].

To this end, model development should adhere to recognized guidelines, such as the American Psychological Association's Code of Ethics [4] and AI risk management framework from NIST [3]. In constructing datasets, researchers must account for regulations such as the General Data Protection Regulation (GDPR) that cover commercial use, scientific data handling, informed consent, data deidentification, and adherence to a code of conduct [92]. Thorough ethical consideration and researcher responsibility are vital to creating a safe and reliable counseling system.

## Study Limitations & Future Directions

**Study Limitation:** Numerous journals focused on AI applications in psychological counseling have yet to be indexed in major academic databases. Also, since the paper collection concluded on May 7, 2024, some recent studies addressing the identified challenges may not have been included by the time of publication. Furthermore, while primary databases for conference proceedings, such as ACM and SCOPUS, were utilized, our search may not comprehensively capture all technically oriented publications. Despite extensive use of leading engineering databases, certain innovations remain absent from our review, likely due to a limited focus on evaluations specific to the mental health domain. Additionally, ethical considerations were not deeply addressed when discussing the current state and challenges in developing a counseling model. Although an initial pool of 146 papers was identified, only approximately 30 were subjected to qualitative analysis. The complete list of these 146 papers is available in Multimedia Appendix 1.

**Future Directions:** For future systematic review research, scrutinizing more recent studies will help capture the latest trends in AI-driven psychological counseling systems. Expanding the investigation of interdisciplinary collaboration between the fields of computer science and mental health will better align technological advances with mental health needs. In addition, conducting a broader qualitative analysis covering all 146 identified papers, or a larger sample, could provide deeper insights into emerging trends and ethical considerations, improving our understanding of the future direction of AI in advancing mental health.

## Conclusions

This study conducted a quantitative bibliometric analysis along with a qualitative trend review of publications on AI-driven dialogue systems for mental health applications. Using three citation databases—WoS, Scopus, and the ACM Digital Library—we examined literature from 2020 to May 2024, ultimately filtering 146 relevant papers. Through bibliometric analysis, we assessed the distribution of publications across various categories, including sources, countries, institutions, and authors. Additionally, we conducted a network analysis of frequently used keywords to identify prominent themes within the literature. In the qualitative trend review, we analyzed three categories: (i) highly cited publications, (ii) publications utilizing the ESConv dataset, and (iii) publications employing LLMs. Among the top 10 most cited papers, we explored approaches that incorporate psychological knowledge in developing deep learning models, as well as the datasets, computational techniques, and evaluation metrics applied in this research area. Similarly, in reviewing ESConv, we addressed the dataset's applicability within psychological counseling systems. We found that notable computational techniques included multi-task learning and the integration of external knowledge. Both automatic and human evaluation metrics were utilized to enhance the assessment of emotional support quality. Lastly, we scrutinized the use of various LLMs based on the goals of psychological counseling and their distinct features. By demonstrating the advantages of LLMs over traditional deep learning models, we also reviewed strategies to address critical limitations in using LLMs for counseling, such as hallucinations, limited mental health-related knowledge, and the lack of comprehensive counseling datasets. In the discussion, we highlighted essential challenges and outlined future directions crucial for advancing AI counseling models.

We believe this work contributes to both the machine learning and psychology communities by offering a structured roadmap to enhance the effectiveness and applicability of AI counseling systems. Specifically, our study highlights critical areas for model development, such as incorporating psychological expertise and improving data accessibility. It also offers practical recommendations, including the application of LLMs and the refinement of evaluation methods. By analyzing current research trends and establishing a foundational framework, this work has the potential to reduce manual labor, provide research resources, and promote advancements in public health.

## Acknowledgment

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program)(RS-2024-00425354) and the Global Scholars Invitation Program(RS-2024-00459638), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

## Conflicts of Interest

None declared.

## References

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large lan- guage models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

[2] Arfan Ahmed, Asmaa Hassan, Sarah Aziz, Alaa A Abd-Alrazaq, Nashva Ali, Mahmood Alzubaidi, Dena Al-Thani, Bushra Elhusein, Mohamed Ali Siddig, Maram Ahmed, et al. Chatbot features for anxiety and depression: a scoping review. *Health informatics journal*, 29(1):14604582221146719, 2023.

[3] NIST AI. Artificial intelligence risk management framework: Generative artificial intelligence profile, 2024.

[4] American Psychological Association et al. Ethical principles of psychologists and code of conduct. *American psychologist*, 57(12):1060–1073, 2002.

[5] Alexandra Bachelor. Comparison and relationship to outcome of diverse dimensions of the helping alliance as seen by client and therapist. *Psychotherapy: Theory, Research, Practice, Training*, 28(4):534, 1991.

[6] Luke Balcombe and Diego De Leo. Digital mental health challenges and the horizon ahead for solutions. *JMIR Mental Health*, 8(3):e26811, 2021.

[7] Keqin Bao, Jizhi Zhang, Xinyu Lin, Yang Zhang, Wenjie Wang, and Fuli Feng. Large

language models for recommendation: Past, present, and future. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2993–2996, 2024.

[8] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019.

[9] Lennart Brocki, George C Dyer, Anna Gładka, and Neo Christopher Chung. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 395–398. IEEE, 2023.

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901, 2020.

[11] Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7858–7873, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C Mazziotta, and Gian Luigi Lenzi. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the national Academy of Sciences*, 100(9):5497–5502, 2003.

[13] Fabio Catania, Micol Spitale, and Franca Garzotto. Conversational agents in therapeutic interventions for neurodevelopmental disorders: a survey. *ACM Computing Surveys*, 55(10):1–34, 2023.

[14] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

[15] Keqi Chen, Huijun Lian, Yingming Gao, and Ya Li. Emotional support dialog system through recursive interactions among large language models. In *National Conference on Man-Machine Speech Communication*, pages 151–163. Springer, 2023.

[16] Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, 2023.

[17] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026, 2022.

[18] Young Min Cho, Sunny Rai, Lyle Ungar, Joa˜o Sedoc, and Sharath Guntuku. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11346–11369, Singapore, December 2023. Association for Computational Linguistics.

[19] Keum-Hyeong Choi, Wendy Buskey, and Bonita Johnson. Evaluation of counseling outcomes at a university counseling center: The impact of clinically significant change on problem resolution and academic functioning. *Journal of Counseling Psychology*, 57(3):297, 2010.

[20] Arthur J Clark. Empathy: An integral model in the counseling process. *Journal of Counseling & Development*, 88(3):348–356, 2010.

[21] Simon Coghlan, Kobi Leins, Susie Sheldrick, Marc Cheong, Piers Gooding, and Simon D'Alfonso. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health*, 9:20552076231183542, 2023.

[22] National Council. America's mental health 2018. 2018.

[23] MH Davis. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology/American Psychological Association*, 85, 1980.

[24] Jean Decety. Perspective taking as the royal avenue to empathy. *Other minds: How humans bridge the divide between self and others*, 143:157, 2005.

[25] Kerstin Denecke, Alaa Abd-Alrazaq, and Mowafa Househ. Artificial intelligence for chatbots in mental health: opportunities and challenges. *Multiple perspectives on artificial intelligence in healthcare: Opportunities and challenges*, pages 115–128, 2021.

[26] Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095, 2023.

[27] Jacob Devlin. BERT: Pre-training deep bidirectional transformers for language understanding.
*arXiv preprint arXiv:1810.04805*, 2018.

[28] Changming Duan and Clara E Hill. The current state of empathy research. *Journal of counseling psychology*, 43(3):261, 1996.

[29] Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3184–3200, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[30] James F Efstation, Michael J Patton, and CarolAnne M Kardash. Measuring the working alliance in counselor supervision. *Journal of counseling Psychology*, 37(3):322, 1990.

[31] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399, 2018.

[32] Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. Emowoz: A large-scale corpus and labelling scheme for emotion recog- nition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, 2022.

[33] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

[34] Mauzama Firdaus, Gopendra Singh, Asif Ekbal, and Pushpak Bhattacharyya. Multi-step prompting for few-shot emotion-grounded conversations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3886–3891, 2023.

[35] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[36] James Gibson, David C Atkins, Torrey A Creed, Zac Imel, Panayiotis Georgiou, and Shrikanth Narayanan. Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, 13(1):508–518, 2019.

[37] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[38] Albert Gu, Caglar Gulcehre, Thomas Paine, Matt Hoffman, and Razvan Pascanu. Improving the gating mechanism of recurrent neural networks. In *International conference on machine learning*, pages 3800–3809. PMLR, 2020.

[39] Imane Guemghar, Paula Pires de Oliveira Padilha, Amal Abdel-Baki, Didier Jutras-Aswad, Jesseca Paquette, and Marie-Pascale Pomey. Social robot interventions in mental health care and their outcomes, barriers, and facilitators: scoping review. *JMIR Mental Health*, 9(4):e36094, 2022.

[40] Suchin Gururangan, Ana Marasovic´, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

[41] MD Romael Haque and Sabirat Rubya. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR mHealth and uHealth*, 11(1):e44838, 2023.

[42] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[43] Jochen Hartmann. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/ emotion-english-distilroberta-base/, 2022.

[44] Clara E Hill. *Helping skills: Facilitating exploration, insight, and action*. American Psycholog- ical Association, 2020.

[45] Adam O Horvath and B Dianne Symonds. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139, 1991.

[46] Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[48] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, 2023.

[49] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. Neural computation, 6(2):181–214, 1994.

[50] Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee and Jinyoung Yeo. Can language models provide good emotional support? Mitigating preference bias in emotional support conversations. In Lun-Wei Ku, Andre Martins and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[51] Laura B Kasper, Clara E Hill, and Dennis M Kivlighan Jr. Therapist immediacy in brief psychotherapy: Case study i. *Psychotherapy: Theory, Research, Practice, Training*, 45(3):281, 2008.

[52] Md Nadim Kaysar and Shun Shiramatsu. Mental state-based dialogue system for mental health care by using gpt-3. In *International Congress on Information and Communication Technology*, pages 891–901. Springer, 2023.

[53] Ksenia Kharitonova, David Pe´rez-Ferna´ndez, Javier Gutie´rrez-Hernando, Asier Gutie´rrez- Fandin˜o, Zoraida Callejas, and David Griol. Incorporating evidence into mental health q&a: a novel method to use generative language models for validated clinical content extraction. *Behaviour & Information Technology*, pages 1–18, 2024.

[54] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, 2021.

[55] Jina Kim, Daeun Lee, Eunil Park, et al. Machine learning for mental health in social media: bibliometric study. *Journal of Medical Internet Research*, 23(3):e24870, 2021.

[56] D Martin Kivlighan III, Barry A Schreier, Chelsey Gates, Jung Eui Hong, Julie M Corkery, Cari L Anderson, and Paula M Keeton. The role of mental health counseling in college students' academic success: An interrupted time series analysis. *Journal of Counseling Psychology*, 68(5):562, 2021.

[57] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Supporting the demand on mental health services with ai-based conversational large language models (llms). *BioMedInformatics*, 4(1):8–33, 2023.

[58] Catherine Penny Hinson Langford, Juanita Bowsher, Joseph P Maloney, and Patricia P Lillis. Social support: a conceptual analysis. *Journal of advanced nursing*, 25(1):95–100, 1997.

[59] Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Mataric´, Daniel J McDuff, and Megan Jones Bell. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479, 2024.

[60] Pei-Chun Lee and Hsin-Ning Su. Investigating the structure of regional innovation system research through keyword co-occurrence and social network analysis. *Innovation*, 12(1):26–40, 2010.

[61] Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683, 2022.

[62] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Ku¨ttler, Mike Lewis, Wen-tau Yih, Tim Rockta¨schel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[63] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, 2017.

[64] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. EmpDG: Multi-resolution interactive empathetic dialogue generation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computa- tional Linguistics*, pages 4454–4466, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[65] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001, 2022.

[66] Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E Kraut, and David C Mohr. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236, 2023.

[67] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13622–13623, 2020.

[68] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.

[69] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, 2021.

[70] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[71] Jose Llanes-Jurado, Luc´ıa Go´mez-Zaragoza´, Maria Eleonora Minissi, Mariano Alcan˜iz, and Javier Mar´ın-Morales. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, 246:123261, 2024.

[72] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, 2017.

[73] Ellen T Luepker. *Record keeping in psychotherapy and counseling: Protecting confidentiality and the professional relationship*. Routledge, 2012.

[74] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal,

Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking emotions for empathetic re- sponse generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online, November 2020. Association for Computational Linguistics.

[75] Laura Moeseneder, Euge´nia Ribeiro, John Christopher Muran, and Franz Caspar. Impact of con- frontations by therapists on impairment and utilization of the therapeutic alliance. *Psychotherapy research*, 29(3):293–305, 2019.

[76] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[77] Jouni Petteri Moilanen. A review of proposals for improvements in evaluation of natural language generation. 2023.

[78] Vijaya Lakshmi Pavani Molli. Effectiveness of ai-based chatbots in mental health support: A systematic review. *Journal of Healthcare AI and ML*, 9(9):1–11, 2022.

[79] Hongbin Na. CBT-LLM: A Chinese large language model for cognitive behavioral therapy- based mental health question answering. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia, May 2024. ELRA and ICCL.

[80] Aure´lien Naldi, Elisabeth Remy, Denis Thieffry, and Claudine Chaouiya. Dynamically consis- tent reduction of logical regulatory graphs. *Theoretical Computer Science*, 412(21):2207–2218, 2011.

[81] Steven Adler et al. OpenAI, Josh Achiam. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[82] Barbara L Paulson, Derek Truscott, and Janice Stuart. Clients' perceptions of helpful experiences in counseling. *Journal of counseling psychology*, 46(3):317, 1999.

[83] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. Control globally, un- derstand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*, 2022.

[84] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversa- tion. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4324–4330. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[85] Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowledge-Based Systems*, 264:110340, 2023.

[86] Seamus Prior. Overcoming stigma: How young people position themselves as counselling service users. *Sociology of health & illness*, 34(5):697–713, 2012.

[87] Yushan Qian, Weinan Zhang, and Ting Liu. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements.

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, 2023.

[88] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*, 2023.

[89] Stephen A Rains, Corey A Pavlich, Bethany Lutovsky, Eric Tsetsi, Anjali Ashtaputre. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: A comforting computer program was revisited. *Journal of Social and Personal Relationships* 37(2):647–666, 2020.

[90] Rashmi Rangaswamy et al. Ai-driven mental health counseling: Opportunities, challenges, and ethical implications. *Revista Electronica de Veterinaria*, 25(1S):550–558, 2024.

[91] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, 2019.

[92] General Data Protection Regulation. General data protection regulation (gdpr). *Intersoft Consulting, Accessed in October*, 24(1), 2018.

[93] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics.

[94] Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A prelimi- nary review. *Future Internet*, 15(6):192, 2023.

[95] Sahand Sabour, Chujie Zheng, and Minlie Huang. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237, 2022.

[96] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205, 2021.

[97] Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. Medical dialogue system: A survey of categories, methods, evaluation and challenges. *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861, 2024.

[98] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. Generating empathetic responses by looking ahead the user's sentiment. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7989–7993. IEEE, 2020.

[99] Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, *Proceedings of the 4th Workshop on NLP for*

*Conversational AI*, pages 77–97, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[100] Steven John Stack. Mental illness and suicide. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, pages 1618–1623, 2014.

[101] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, 2021.

[102] Adela C Timmons, Jacqueline B Duong, Natalia Simo Fiallo, Theodore Lee, Huong Phuc Quynh Vo, Matthew W Ahle, Jonathan S Comer, LaPrincess C Brewer, Stacy L Frazier, and Theodora Chaspari. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, 18(5):1062–1096, 2023.

[103] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Tim- othe´e Lacroix, Baptiste Rozie`re, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[104] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[105] Amy JC Trappey, Aislyn PC Lin, Kevin YK Hsu, Charles V Trappey, and Kevin LK Tu. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. *Processes*, 10(5):930, 2022.

[106] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, 2022.

[107] Nicholas Vacc and Larry C Loesch. *Professional orientation to counseling*. Routledge, 2013.

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

[109] Karin Verspoor. 'fighting fire with fire'—using llms to combat llm hallucinations, 2024.

[110] Wout Vossen, Maxwell Szymanski, and Katrien Verbert. The effect of personalizing a psy- chotherapy conversational agent on therapeutic bond and usage intentions. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 761–771, 2024.

[111] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[112] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31, 2021.

[113] Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu,

Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970, 2022.

[114] Rui Wang, Fei Mi, Yi Chen, Boyang Xue, Hongru Wang, Qi Zhu, Kam-Fai Wong, and Ruifeng Xu. Role prompting guided domain adaptation with general capability preserve for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2243–2255, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[115] Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Ar- maghan Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8299–8307. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Survey Track.

[116] Anuradha Welivita and Pearl Pu. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, 2020.

[117] Anuradha Welivita and Pearl Pu. Heal: A knowledge graph for distress management con- versations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11459–11467, 2022.

[118] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE, 2022.

[119] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, 2023.

[120] Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, 2021.

[121] Jeffrey N Younggren and Eric A Harris. Can you keep a secret? confidentiality in psychother- apy. *Journal of clinical psychology*, 64(5):589–600, 2008.

[122] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-alpha: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

[123] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[124] Wei-Nan Zhang, Lingzhi Li, Dongyan Cao, and Ting Liu. Exploring implicit feedback for open domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[125] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[126] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics.

[127] Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In Anna Rogers, Jordan Boyd- Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694, Toronto, Canada, July 2023. Association for Computational Linguistics.

[128] Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. CPsyCoun: A report-based multi-turn dialogue re- construction and evaluation framework for Chinese psychological counseling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 13947–13966, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[129] Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. Transesc: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, 2023.

[130] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, 2023.

[131] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[132] Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325– 11345, 2024.

[133] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731, 2024.

[134] Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729, 2023.

[135] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[136] Zheming Zuo, Matthew Watson, David Budgen, Robert Hall, Chris Kennelly, and Noura Al Moubayed. Data anonymization for pervasive health care: systematic literature mapping study. *JMIR medical informatics*, 9(10):e29871, 2021.

[137] OECD. Suicide rates [Internet]. 2024 [Accessed on October 24, 2024]. Available from: https://www.oecd.org/en/data/indicators/suicide-rates.html

[138] Scopus. Scopus [Internet]. 2024 [Accessed on October 24, 2024]. Available from:

https://www.scopus.com

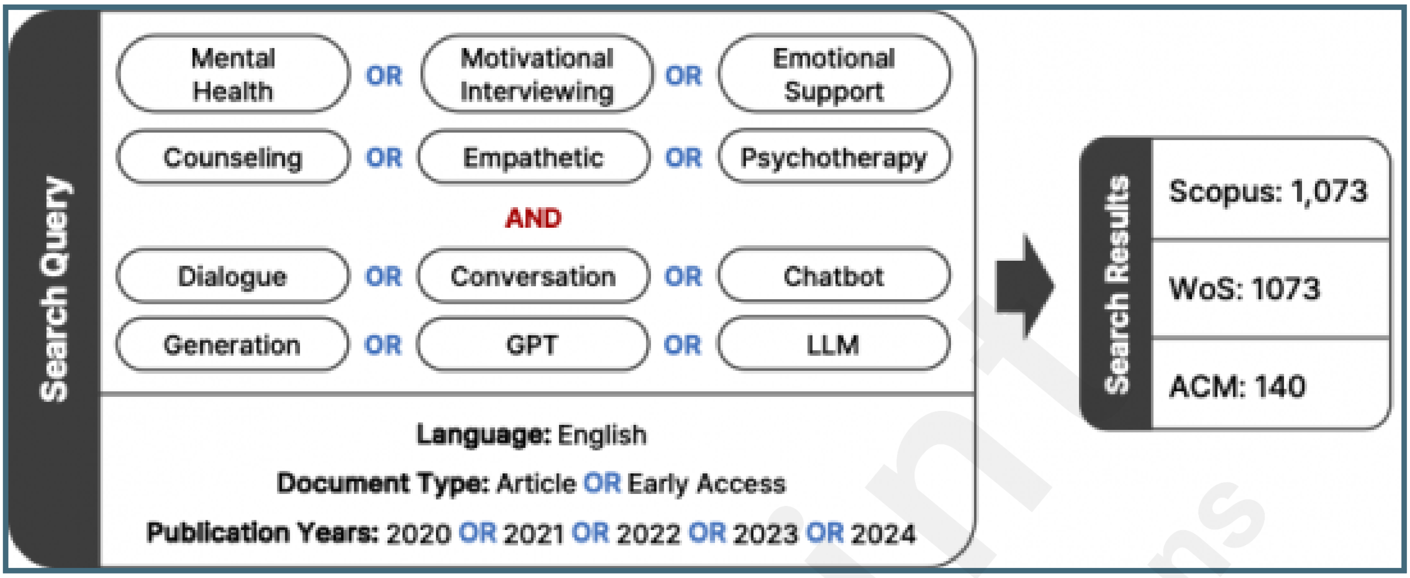
[139] Web of Science. Web of Science [Internet]. 2024 [Accessed on October 24, 2024]. Available from: https://www.webofscience.com

[140] ACM Digital Library. ACM Digital Library [Internet]. 2024 [Accessed on October 24, 2024]. Available from: https://dl.acm.org

# Supplementary Files

# Figures

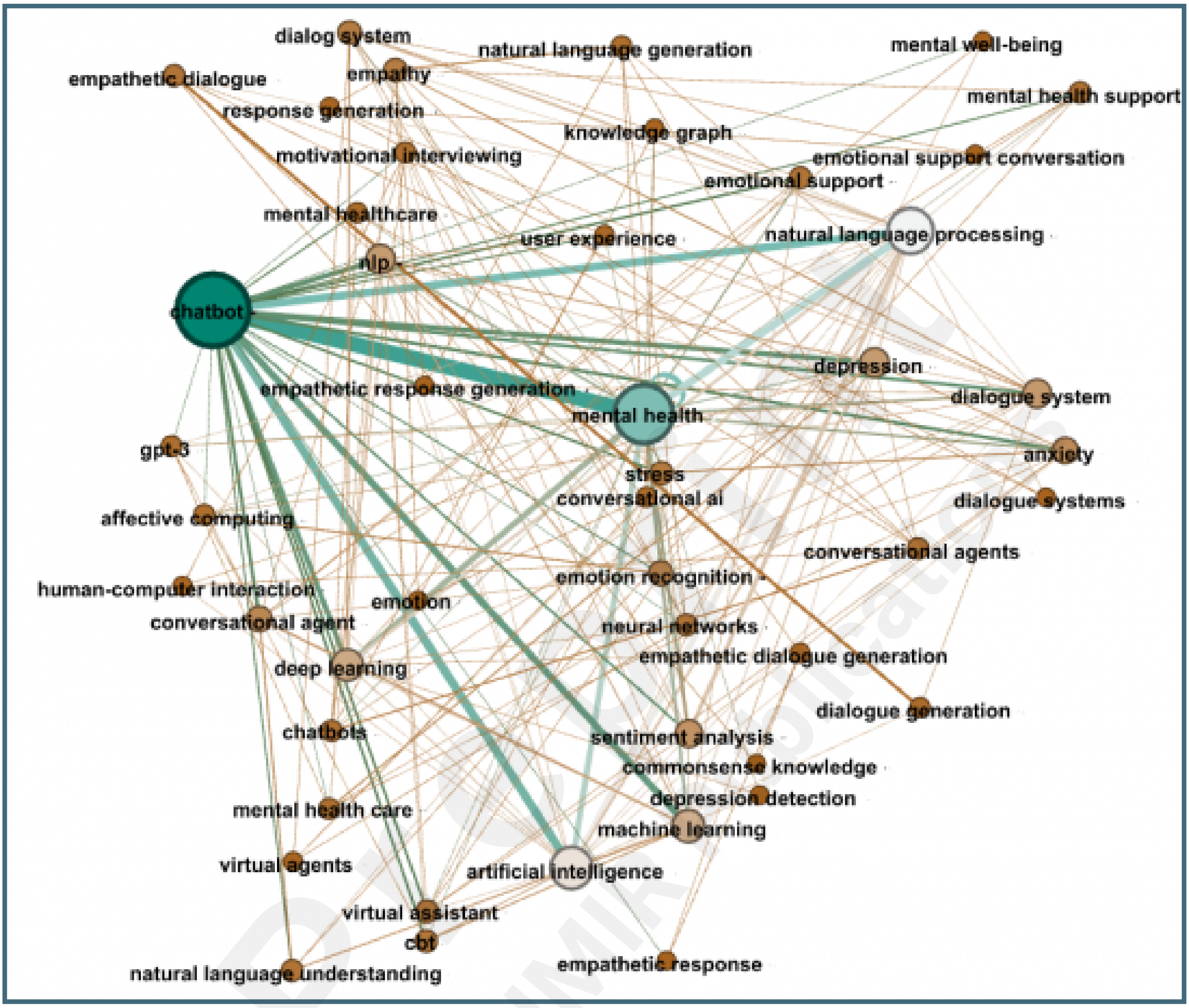
Search Query Categories with Results.



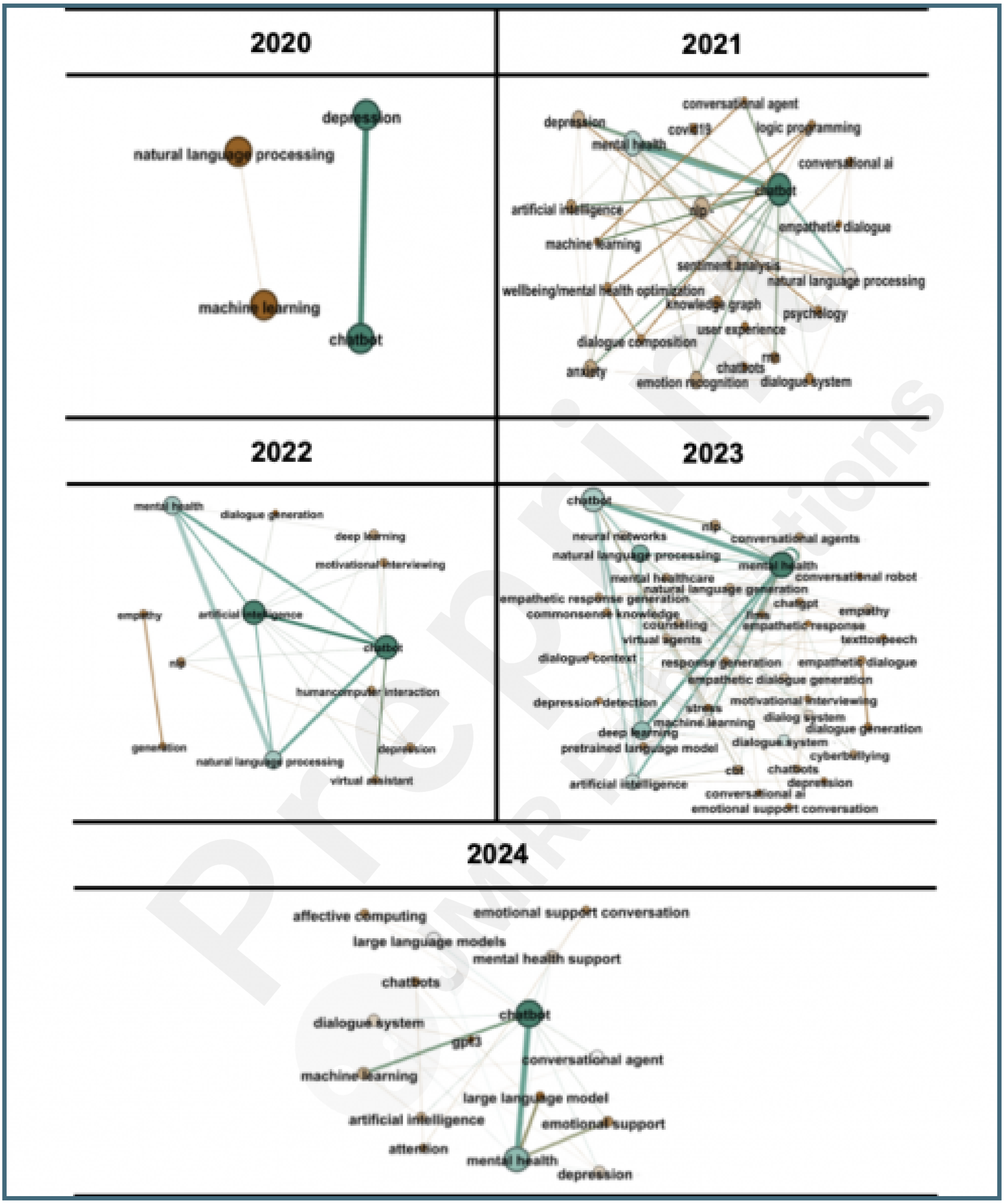
Selection Criteria Overview.

| Selection Criteria | | | |
|---|---|---|---|
| | Overall | 1. | Research proposing the chatbot that **aims to improve user's mental health** |
| | | 2. | Research addressing **deep learning model** |
| | Specific | a. | If a chatbot was developed for the purpose of **evaluating its effectiveness**, exclude it. |
| | | b. | Include if the model structure or methodology is clearly described. Exclude if the development process is **not outlined**. |
| | | c. | Exclude cases where the conversation goal is **not aimed at improving mental health** (MH), even if empathy is present (e.g., negotiations). |
| | | d. | If a module is **not chatbot but related to chatbots** (e.g., knowledge graphs) is the main focus of the study, exclude it. |
| | | e. | Exclude cases where the **chatbot is used for purposes like emotion detection or suicide prevention**. |
| | | f. | Exclude **conversations that support MH improvement** (e.g., emotional episode generation) |
| | | g. | Include studies on counseling robots if the algorithm or architecture is described, Exclude if the focus is on **hardware** or if **algorithm descriptions are absent**. |
| | | h. | Exclude **applications that are actively in service**. |

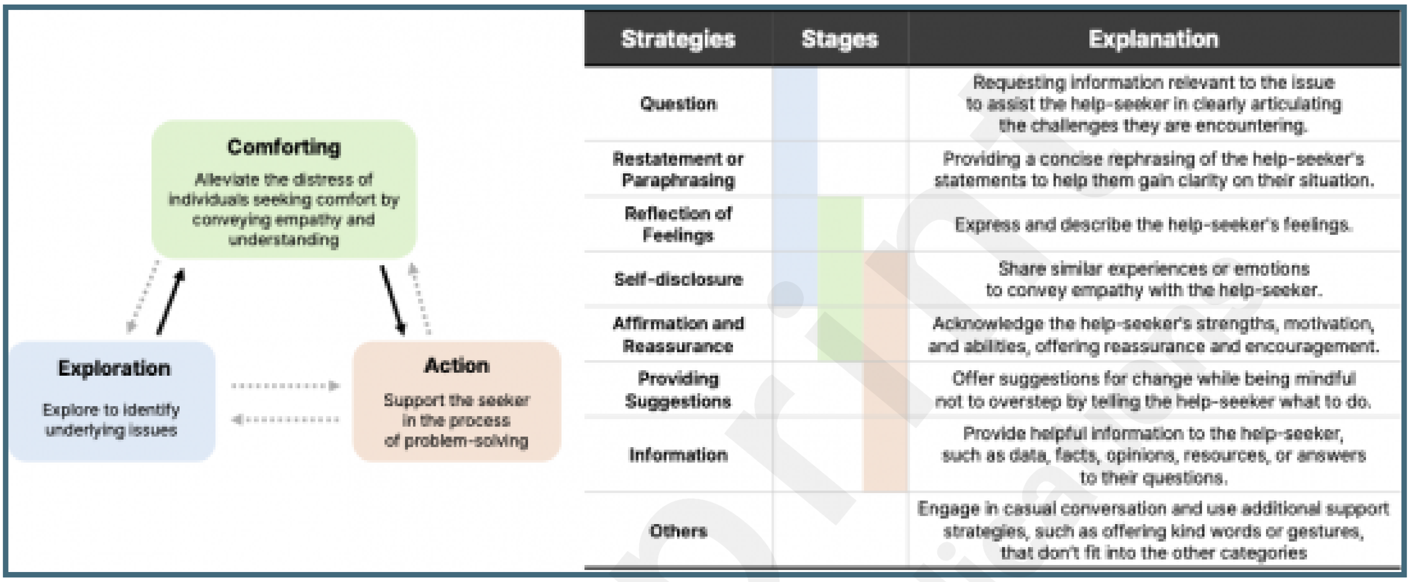
Keyword Co-occurrence Network Graph from 2020 to May 2024.



Yearly Keyword Co-occurrence Network Graphs (2020–2024).



We present figures from the original paper [69] to illustrate three stages of the ESConv framework and the corresponding eight counseling strategies. According to Liu et al. [69], this framework comprises three stages, each with specific support strategies. The exploration stage aims to help individuals identify underlying issues; the comforting stage focuses on providing empathy and understanding; and the action stage involves offering practical information or suggestions. Typically, the emotional support process follows a sequential order from 1. Exploration ? 2. Comforting ? 3. Action, as indicated by black arrows, can also be adjusted to suit the conversation’s needs, as represented by dashed gray arrows.




| Strategies | Stages | Explanation |
|---|---|---|
| Question | | Requesting information relevant to the issue to assist the help-seeker in clearly articulating the challenges they are encountering. |
| Restatement or Paraphrasing | | Providing a concise rephrasing of the help-seeker's statements to help them gain clarity on their situation. |
| Reflection of Feelings | | Express and describe the help-seeker's feelings. |
| Self-disclosure | | Share similar experiences or emotions to convey empathy with the help-seeker. |
| Affirmation and Reassurance | | Acknowledge the help-seeker's strengths, motivation, and abilities, offering reassurance and encouragement. |
| Providing Suggestions | | Offer suggestions for change while being mindful not to overstep by telling the help-seeker what to do. |
| Information | | Provide helpful information to the help-seeker, such as data, facts, opinions, resources, or answers to their questions. |
| Others | | Engage in casual conversation and use additional support strategies, such as offering kind words or gestures, that don't fit into the other categories |

# Multimedia Appendixes

The complete list of 146 papers.
URL: http://asset.jmir.pub/assets/7f7f7ebb34dd6f6815cebd9f31ddad2d.xlsx