

Exploring the Potential of Human-LLM Synergy in Advancing Qualitative Analysis: A Case Study on Mental-Illness Stigma

HAN MENG, YITIAN YANG, YUNAN LI, JUNGUP LEE, and YI-CHIEH LEE, National University of Singapore, Singapore

Qualitative analysis is a challenging, yet crucial aspect of advancing research in the field of Human-Computer Interaction (HCI). Recent studies show that large language models (LLMs) can perform qualitative coding within existing schemes, but their potential for collaborative human-LLM discovery and new insight generation in qualitative analysis is still underexplored. To bridge this gap and advance qualitative analysis by harnessing the power of LLMs, we propose CHALET, a novel methodology that leverages the human-LLM collaboration paradigm to facilitate conceptualization and empower qualitative research. The CHALET approach involves LLM-supported data collection, performing both human and LLM deductive coding to identify disagreements, and performing collaborative inductive coding on these disagreement cases to derive new conceptual insights. We validated the effectiveness of CHALET through its application to the attribution model of mental-illness stigma, uncovering implicit stigmatization themes on cognitive, emotional and behavioral dimensions. We discuss the implications for future research, methodology, and the transdisciplinary opportunities CHALET presents for the HCI community and beyond.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI; HCI design and evaluation methods;** • **Applied computing** → **Psychology.**

Additional Key Words and Phrases: Qualitative Analysis, Qualitative Coding, Human-LLM Collaboration, Attribution Model, Social Stigma, Mental Illness

ACM Reference Format:

Han Meng, Yitian Yang, Yunan Li, Jungup Lee, and Yi-Chieh Lee. 2024. Exploring the Potential of Human-LLM Synergy in Advancing Qualitative Analysis: A Case Study on Mental-Illness Stigma. *J. ACM* 37, 4, Article 111 (August 2024), 59 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Qualitative analysis [18] is crucial for HCI and social science, offering benefits such as conceptualizing phenomena, understanding human behavior [126], and expanding theoretical frameworks [52]. It provides a solid theoretical base to grasp the societal and ethical impacts of new technologies, guiding their responsible development [5]. The importance of qualitative data in HCI research is growing [163], driven by the need to deeply understand human-AI interactions through detailed data analysis [89]. However, qualitative analysis struggles with challenges such as the difficulty of managing large datasets and risking bias [28]. The qualitative coding approach [18, 161] is vital, although it faces issues of bias [28], replication difficulties [100], and being labor-intensive [167], especially with large datasets.

Authors' address: Han Meng, han.meng@u.nus.edu; Yitian Yang, t0931554@u.nus.edu; Yunan Li, liyunan@u.nus.edu; Jungup Lee, swklj@nus.edu.sg; Yi-Chieh Lee, ycllee@nus.edu.sg, National University of Singapore, 21 Lower Kent Ridge Road, Singapore, 119077.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 0004-5411/2024/8-ART111

<https://doi.org/XXXXXXX.XXXXXXX>

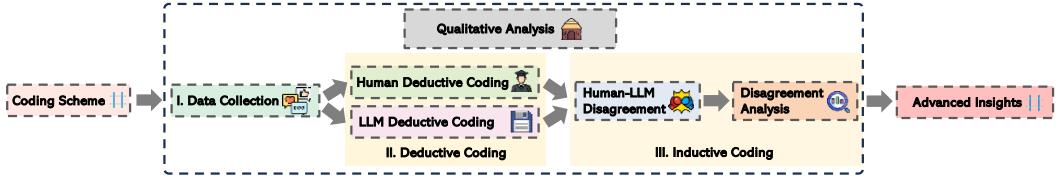


Fig. 1. Overview of the CHALET framework. In this work, we propose this framework to advance qualitative analysis.

LLMs, such as ChatGPT¹ and Llama [157], offer the potential to streamline the labor-intensive qualitative analysis process. As prominent examples of generative AI, these models can understand natural language instructions and generate human-like text outputs, making them well-suited for this task. Recognizing this potential, LLMs are increasingly used for preliminary qualitative coding tasks, facilitating both inductive [87] and deductive [167] approaches. Although recent studies highlight the capability of LLMs in deductive coding *within* existing coding schemes [82, 95], their ability to uncover novel findings and generate insights that facilitate qualitative analysis from these coding schemes remains underexplored. Existing research mainly quantifies LLM-human coder agreement [22, 167], focusing on model performance rather than investigating causes of suboptimal results, i.e., cases where the model's output diverges from the human-generated code. This emphasis on quantitative measures overshadows the importance of qualitative analysis of the discrepancies [95], leading to a gap in understanding and interpreting instances where LLMs differ from human judgment. Without a detailed examination of the inconsistencies between coders, it hinders the understanding and further derivation of new theoretical insights.

After acknowledging the gap, efforts to investigate the discordance between humans and LLMs for deductive coding have been proposed, focusing on refining existing code definitions to improve qualitative coding [36, 172]. While refining code definitions based on discrepancies is a step in the right direction, it is not sufficient to derive qualitative advancements [52]. These initiatives have not thoroughly assessed the overall effectiveness of the coding frameworks in place or fully leveraged the potential of LLMs to contribute to deeper conceptual development. LLM-assisted deductive coding has the potential to go beyond simple code refinement and open avenues for deeper integration with conceptual development through qualitative coding, as LLMs have been shown to stimulate research creativity [119] and detect linguistic subtleties as well as latent arguments [170], possibly revealing new perspectives that could enrich and deepen the qualitative analysis process. Thus, further exploration of how LLMs can contribute to qualitative analysis and conceptualization is warranted, as they might offer deeper and nuanced interpretations.

To explore the potential of LLM-assisted qualitative analysis in fostering richer conceptualization, we thereby propose a methodological framework called CHALET (Collaborative Human-LLM AnaLysis for Empowering Conceptualization in QualiTative Research), which aims to promote qualitative analysis by employing an LLM-assisted hybrid coding process, following three steps: data collection, deductive coding, and inductive coding. Figure 1 illustrates the above steps. LLMs play a crucial role in all steps of the framework: 1) facilitating the collection of multi-dimensional data through the promotion of self-disclosure and 2) enabling LLM-assisted qualitative coding. Uniquely, our framework focuses on discrepancies between human and LLM coding. By analyzing instances where LLM coding diverges from human coding, we performed in-depth inductive coding with LLM support to uncover new themes and insights. This approach, naturally possessing a synergistic human-LLM collaboration paradigm, adeptly integrates the inductive ability, social

¹<https://chat.openai.com/>

awareness and reasoning strengths of human researchers with the efficient data processing [167] and extensive knowledge retrieval capabilities of LLMs [109]. It allows us to derive new conceptual insights through qualitative analysis, leveraging the unique perspectives offered by human-LLM collaboration.

We assessed our framework with a case study on the stigma attribution model of mental illness [32]. This model, linking causes attributed to behaviors with mental illness stigma, connects responsibility perceptions to stigma-driven emotions and actions [32]. Emerging concepts such as microaggression [58] and mental illness invalidation [62] necessitate a deeper qualitative analysis of the attribution model to reflect contemporary social stigma perspectives. Our human-LLM synergistic framework, applied here, integrates these views, enhancing understanding of social stigma and its conceptual insights. We used this framework to analyze conversations from 1,002 participants, revealing new insights into cognitive, emotional, and behavioral responses, and identified linguistic features and prompting techniques that affect LLM coding of psychological constructs, demonstrating our framework's ability to advance qualitative analysis.

Our work contributes in several ways. Firstly, we introduce an LLM-assisted framework to enhance conceptualization, using an attribution model of mental illness stigma to demonstrate its effectiveness in advancing qualitative analysis and providing insights through human-LLM collaboration. This method addresses issues like bias, reproducibility, and time efficiency. Secondly, we offer a new perspective on structuring human-LLM collaboration for qualitative research, discussing future method implications, practical LLM uses, and the pros and cons of LLM adoption, potentially enriching future theoretical work. Finally, we present a systematic method for LLMs to analyze psychological constructs, exploring prompt design, and identifying performance-limiting factors. This supports the creation of psychological datasets with LLM assistance [40], merging extensive language data collection with LLM-assisted coding to produce high-quality and psychologically relevant datasets, helping future research.

2 RELATED WORK

2.1 The Importance of Qualitative Analysis

Qualitative analysis is a crucial research methodology in various fields, with qualitative coding being one of its essential techniques. Qualitative coding involves the process of categorizing and labeling data to identify patterns, themes, and hypotheses. Deductive coding [50] applies predefined codes derived from existing theories, while inductive coding [71] generates codes from the data itself; hybrid deductive-inductive coding [51] combines both approaches, starting with predefined codes and allowing new codes to emerge from the data concurrently.

Qualitative coding plays a crucial role in understanding complex human experiences, behaviors, and psychological constructs. Through the systematic examination of qualitative data, researchers can gain deep insights into the thoughts, emotions, and motivations that drive human actions [139]. Moreover, qualitative analysis allows researchers to explore the rich tapestry of human experience across various domains [18]. This approach is valuable in the field of psychology, where understanding the intricacies of human cognition and behavior is essential for developing effective interventions and treatments [146].

However, analyzing complex qualitative data is a challenging task that requires significant time, effort, and expertise [100]. The process of coding, categorizing, and interpreting qualitative data is often labor-intensive and demands a high level of analytical skills and subject matter knowledge from researchers [129]. To address these challenges, technological methods and tools are needed to assist in the qualitative analysis process. Advanced technologies like Natural Language Processing (NLP) [124] and LLMs [43] might potentially support these activities.

2.2 LLM-Assisted Data Annotation

LLMs have gained the attention of researchers, with applications in domains such as education [132] and potential for future exploration in fields such as sociology, psychology and other social sciences [40]. Recent trends focus on the potential of LLMs as text data annotators [74], as they have shown promise in text classification tasks [150]. Previous studies found that LLMs could outperform MTurk in tasks such as relevance determination and topic detection [3, 59], and perform well in zero-shot tasks, such as labeling COVID-19 vaccination-related social media data [153]. They were also used in automatic genre identification [96], outperforming fine-tuned models on unseen datasets, and demonstrated potential to label misinformation [80].

Despite the already demonstrated performance of LLMs, their ability to handle subjective tasks crucial for examine social phenomena has been explored. These tasks, known to be challenging even for humans, include data annotation in various domains such as stance detection, sentiment analysis, hate speech, and bot detection [174]. The results suggest LLMs' effectiveness in these tasks. Further research [84] has shown LLMs' potential in tasks requiring analyzing of social norms and cultural context, like detecting implicit hate speech. Similarly, LLMs have shown high accuracy and reliability in annotating political Twitter messages [156]. Amin et al. [6] also found LLMs to be versatile and high-performing in tasks like personality prediction, feelings analysis, and suicide tendency detection, without specialized training.

Current research has made significant progress in using LLMs for text classification tasks. However, these efforts often lack theoretically-grounded qualitative analysis, which can limit the depth and scope of theoretical findings derived from the data [84]. By focusing primarily on conceptualization-agnostic classifications, existing studies may not fully exploit the potential significance and intrinsic relationships within the data. Furthermore, emphasizing classification performance while neglecting the annotation process itself can hinder effective conceptualization, potentially diminishing the overall theoretical contribution of the research.

To address this limitation, qualitative coding [97] can compensate for the lack of theoretical insights and deeper analysis. Qualitative coding is typically guided by specific conceptual coding schemes, enabling a more profound and comprehensive understanding of the inherent meanings and interconnections within the data. LLMs have the potential to perform qualitative coding, such as aiding the understanding of complex psychological constructs, offering insights into nuanced relationships and latent structures through their vast knowledge. For instance, attribution models [32] that aim to explain perceived causes of behavior could benefit from the rich contextual information captured by LLMs. Consequently, it is crucial to investigate whether LLMs can be leveraged for qualitative coding, a classification approach rooted in conceptual groundings. Such an investigation could potentially empower researchers to derive more conceptually-rich findings from the vast textual data processed by LLMs.

2.3 Qualitative Analysis with Technology Support

2.3.1 LLMs Support Deductive Coding. When performing deductive coding, researchers typically adhere to a comprehensive codebook to ensure that the coding is aligned with the predefined characteristics or constructed concepts being applied [151]. Manual methods for deductive coding, such as iterating in the codebook and data coding, are time-consuming and resource-intensive, particularly for large datasets [161]. With technological advancements, computer-assisted methods like dictionary-based approaches and machine learning have emerged [124]. These methods promise efficiency, enabling the analysis of larger, diverse datasets and the detection of nuanced patterns [26]. The advent of LLMs has further enhanced the processes of deductive content analysis.

Recent studies have shown the potential of LLMs in deductive coding tasks. Xiao et al. [167] found a fair to substantial agreement between human experts and LLMs when using expert-developed codebooks. Similar conclusions have been drawn in previous studies [44, 95] indicating that for certain codebooks, LLMs are capable of effectively conducting large-scale content analysis. Tai et al. [151] demonstrated that LLMs to identify the presence of codes in a given sample text and provide supporting evidence, concluding that LLM analysis can support qualitative researchers by deductively coding transcripts in a systematic and reliable manner. Hou et al. [82] found that prompt engineering techniques can improve agreement between expert and LLM-coded data. However, Bano et al. [12] found low agreement between LLMs and human coders, attributing this to humans relying on personal experiences and LLMs on word choices. Similarly, Ashwin et al. [10] suggest that using LLMs can lead to significant bias because they overpredict many codes, which can be thought of as the LLM bringing the preconceptions.

Although prior studies have showcased the capabilities of LLMs in a range of content analysis tasks that require deductive coding, the main focus has been on the accuracy of LLMs and the degree to which they code correctly, with less emphasis on the elements that lead to disagreements. To address this, only a few studies [22, 172] have made initial efforts to refine code wording based on the differences observed between human coders and LLM coders. These studies suggested methodological structures to enhance code definitions for more accurate deductive coding. Nevertheless, depending solely on the discrepancies between humans and LLMs to refine the code wording might not be adequate. LLMs hold the potential to work in tandem with human researchers on complex research tasks, thanks to their impressive abilities, such as the capacity to identify linguistic nuances [170]. This can be beneficial in several ways, such as discovering and defining new constructs and conceptualizing new findings. Considering the potential of LLMs, it is necessary to gain a deeper understanding from the perspective of disagreements between humans and LLMs.

2.3.2 LLMs Support Inductive Coding. Inductive coding, a qualitative coding method, aids in this process by interpreting raw text-based data to identify patterns and construct ideas, independent of existing conceptualization or hypotheses [16]. Initially, inductive coding process involved manually finding observations from the data, deriving patterns, forming hypotheses, and finally conceptualizing the phenomenon [136]. Subsequently, several computer-assisted methods emerged to facilitate inductive coding for researchers, including platforms such as Atlas.ti [140], MaxQDA², nVivo³, and nCoder [172]. Additionally, various NLP-based techniques have been developed to assist in this process, such as topic modeling [101], a computer-assisted inductive coding procedure that includes the detection of code patterns [123], and clustering methods [128]. The emergence of LLMs has further advanced inductive coding.

Research has shown that LLMs can support inductive coding by generating, searching, and uncovering themes [43, 78, 87]. Their performance, evaluated comparable human-led coding on the same datasets, proved their ability to deduce the most of primary themes from previous studies [37, 39]. Furthermore, LLMs have been found to be a superior alternative to traditional qualitative coding and conventional NLP techniques [57].

While studies show that LLMs can help in proposing codes and themes from interview transcripts, their utility is limited with large datasets. Previous approaches often involved directly feeding large volumes of data into LLMs without a systematic methodology, which might not be suitable for massive data [39] due to LLMs' input length limitations. In such situations, humans still need to manually process and analyze the large amounts of data, which is laborious. Additionally, previous studies have primarily focused on distinguishing relatively obvious topics [87], but identifying

²<https://www.maxqda.com>

³<https://lumivero.com/products/nvivo/nvivo-product-tour/>

inner attributions and constructs requires a deeper level of social awareness, presenting a more challenging endeavor that has not been extensively explored. Thus, exploring new ways of human-LLM collaboration could help address both the *data quantity* burden and the increased analytical complexity when conceptualizing such nuanced constructs.

2.3.3 Hybrid Approach of Inductive and Deductive Coding. The hybrid approach to inductive and deductive coding is a powerful qualitative analysis method that combines the strengths of both top-down and bottom-up coding strategies [51]. This approach involves starting with a set of predefined codes based on existing conceptual frameworks, while also allowing new codes and themes to emerge from the data itself. By integrating these two complementary techniques, researchers can build on established theoretical foundations while remaining open to discovering novel insights grounded in the specific context of the study [68]. This hybrid method represents an indispensable and integral component of qualitative analysis and enables a more comprehensive and flexible analysis of qualitative data.

Although LLMs have been preliminarily applied in deductive and inductive coding, their application in the hybrid coding approach remains underexplored. To date, no research has incorporated LLMs or other AI techniques into the process of hybrid qualitative analysis, resulting in a research gap between human-LLM synergy and the potential for advanced conceptual insights derived from qualitative analysis using a hybrid approach. Filling this gap is crucial, as it can complement the impact of the human-LLM collaboration paradigm on advancing qualitative analysis and conceptual development using this novel approach. Our research endeavors to close this gap by introducing a harmonious human-LLM collaboration paradigm that tackles the absence of LLM assistance in hybrid approach of deductive and inductive coding, thereby advancing the technology-supported qualitative analysis. We propose a methodology that deftly combines the strengths of human researchers and LLMs, leveraging their complementary abilities to uncover profound conceptual insights. This approach could simultaneously address the limitations of previous studies, which often lacked depth in their conceptualized and qualitative findings and placed a substantial burden on human researchers.

2.4 Chatbot for Data Collection

Qualitative data collection, a crucial first step in qualitative analysis, significantly impacts data quality. Several methods exist, including interviews, focus groups, and chatbot interactions [9, 20, 46]. However, these methods may elicit social desirability bias, especially when dealing with sensitive information [76, 104, 117]. Although chatbots can induce socially desirable responses [141], people tend to disclose *more* sensitive information to them compared to humans [142]. Chatbot-assisted data collection outperforms traditional survey methods by eliciting more detailed information, enhancing engagement [17]. Compared to human interviews, this approach also offers several advantages, including increased anonymity, mitigation of interviewer bias, improved standardization, and reduced costs [4, 49, 92, 131]. Based on these advantages of conversational agent-assisted data collection, this study developed a chatbot to act as a virtual interviewer and ask questions.

Interviewees' self-disclosure during an interview is crucial for obtaining high-quality, insightful data. *Self-disclosure*, is an act of sharing personal information, thoughts, feelings, and vulnerabilities with others [29]. Recent studies have demonstrated the effectiveness of chatbots in promoting self-disclosure by emulating human-to-human interactions [99]. Chatbots equipped with self-disclosure capabilities have been shown to influence users' perceptions and emotions [79]. Drawing inspiration from previous works, our study aims to employ a chatbot with self-disclosure to understand participants' opinions. By utilizing the chatbot's *neutral* self-disclosure, we seek to

avoid manipulating participants' opinions and elicit more extensive disclosure from them. This approach helps to efficiently collect a large keystone dataset, which is essential for reducing bias [40] and facilitating deeper understanding in qualitative analysis.

3 CHALET: COLLABORATIVE HUMAN-LLM ANALYSIS FOR EMPOWERING CONCEPTUALIZATION IN QUALITATIVE RESEARCH

To bridge the gap in understanding the role of LLM-assisted qualitative coding in advancing qualitative analysis and conceptualization, we thereby made an initial foray into exploring it. We proposed a novel methodological framework,  CHALET, which leverages a hybrid qualitative analysis approach that specifically focuses on the synergistic interplay between human researchers and LLMs. We illustrate our methodological framework in Figure 2. Based on this, we elaborate on each step of the proposed  CHALET framework in the following sections. Our approach addresses the limitations of previous qualitative analysis approaches by eliminating temporal and geographical constraints in qualitative data collection, generating insights from human-LLM discrepancies, and overcoming the cost constraints of conceptualization in large datasets. To guide and introduce our proposed methodology, we have designed the following questions:

- **Q1:** How is the quality of the large-scale dataset collected through human-chatbot interaction?
- **Q2:** How can human-established deductive coding schemes be effectively integrated into LLM-assisted coding to improve the model's interpretation and application of these schemes?
- **Q3:** How can the discrepancies between human coding and LLM-assisted coding be identified and analyzed in the hybrid coding framework?
- **Q4:** What richer qualitative insights can be gained by reconciling human-LLM coding disagreements through inductive coding?

3.1 Chatbot-Assisted Data Collection

We aim to acquire qualitative data from participants and conducting content analysis. Inspired by the current trend of using chatbots to collect data and the finding that chatbots have been proven to enhance people's self-disclosure [99], we employ a human-chatbot conversation approach to simulate interviews. By utilizing an LLM-enabled chatbot to facilitate interviews, we can collect rich and informative data more efficiently, without the constraints of time and location. This LLM-powered data collection mechanism, with the chatbot's carefully curated questions aligned with a coding framework, enables the collection of data where each message corresponds to a specific code, facilitating qualitative coding. AI has been widely adopted for data collection. This paper primarily focuses on the data collection methodologies enabled by LLM-powered chatbots. Techniques like gamification [152] and humanization [135] can transform survey questions into engaging chatbot interactions. Vignettes [1] can be also employed to contextualize the questions within relevant scenarios. By employing a well-structured vignette-to-question flow, we showcase a demonstrative case of human-chatbot interaction for survey questionnaires.

3.2 Human-LLM Synergistic Deductive Coding

After data collection, we used a synergistic deductive coding approach involving both human coders and LLM coders. Initially, human deductive coding was conducted to serve as a basis for developing and obtaining a codebook. Subsequently, we devised a method to transform the human-generated codebook into a set of instructions that could be provided to the LLM. This approach facilitated the LLM's ability to identify the constructs within the content, enabling it to perform deductive coding accurately and in closer alignment with human coders. This process involved both human coders and LLM coders to code the same set of messages, facilitating an in-depth comparison and

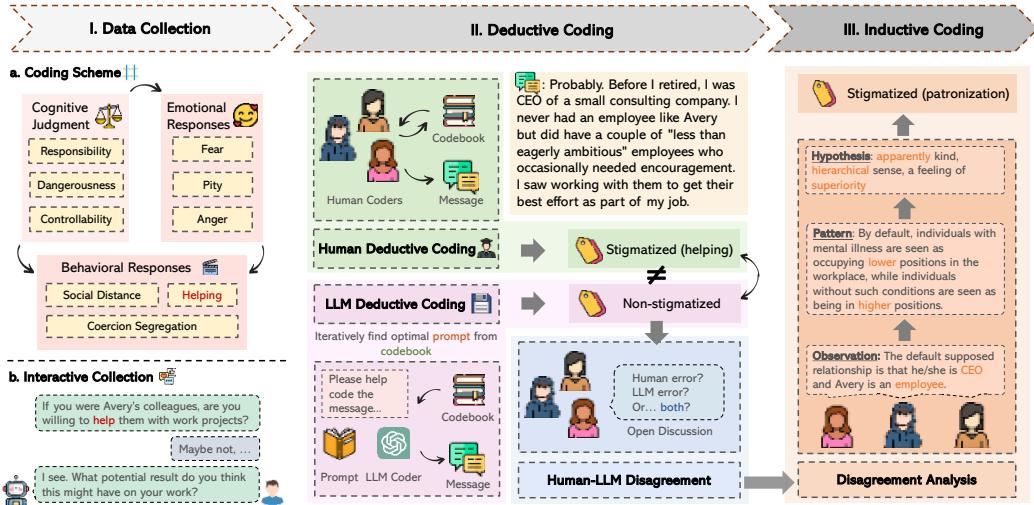


Fig. 2. The CHALET framework. **I. Chatbot-Assisted Data Collection:** (a) **Coding Scheme**. The attribution model [32] is selected as the foundational conceptual framework. This model investigates social stigma associated with mental illness, covering *cognitive judgment* (responsibility, dangerousness, controllability), *emotional responses* (fear, pity, anger), and *behavioral responses* (social distance, helping, coercion segregation). (b) **Interactive Collection**. Data is interactively collected from participants through questions that align with the proposed coding scheme. For example, participants are queried about their willingness to assist individuals with mental illness (corresponding to the *helping* attribution). **II. Human-LLM Synergistic Deductive Coding.** Human coders utilize the attribution model to develop a codebook and code the response as *Stigmatized*. This codebook is then learned by LLMs, which, in contrast, code the same response as *Non-stigmatized*. **III. Collaborative Inductive Coding.** Disagreements between human coders and LLMs lead to further qualitative analysis, revealing a perceived CEO-employee relationship and an underlying assumption of a hierarchical social structure. This is hypothesized to indicate a sense of superiority, resulting in the final coding of *Stigmatized (Patronization)*.

qualitative analysis of their coding results, ultimately achieving a synergistic collaboration between human and LLM coders.

3.2.1 Human Deductive Coding. Initially, human coders engage in qualitative retrospective analysis, *deductive coding* [139], anchored in the collected data and the conceptual framework. Methodologically, the process begins with identifying key concepts from the conceptual framework and developing a preliminary codebook to maximize *inter-rater reliability* [27] and coding coherence. The codebook includes code names, definitions, qualifiers, exclusions [50], and examples [125], serving as a guide throughout the process [16] and evolving during the early stages of the analysis. Following prior practices [25, 163], at least two primary coders systematically apply the finalized prior codes to the entire dataset, resolving disagreements through open discussion with an additional researcher and consulting a domain specialist if needed. An initial coding step on a subset of data refines the codebook, followed by iterative codebook refinement at checkpoints until inter-rater reliability is met. After coding the entire dataset, a stratified sampling recheck mechanism is applied to ensure coding consistency, reinforcing the validity, reliability, and trustworthiness of the coding process.

3.2.2 LLM Deductive Coding. To train LLMs as skillful coders capable of dissecting the constructs and contents, the curated and finalized codebook during the human deductive coding step serves

as a suitable training material. This can form the basis for the content of the LLM prompts [167]. Understanding the outcomes of different prompts and finding an optimal prompt are crucial. The closer the alignment between LLM and human deductive coding results, the more desirable, as it indicates that the LLM has learned coding methodologies akin to a trained human coder and performs them well.

Notably, unlike previous studies [22], we do not consider human-generated codes as ground truth, but we aim for LLMs to accurately learn the coding scheme set up by humans and replicate human-generated codes, thereby enhancing inter-rater reliability. Tangentially, finding the *optimal prompt* is also necessary to minimize the influence of model suboptimality on our decision and analysis of codes. We aim to reduce suboptimality to avoid errors caused by inadequate prompts that lead to the LLMs' misinterpretation, which may result in biased analyses. To this end, it is hence imperative to design an exploration aimed at searching the prompt space to find the optimal prompt that maximizes inter-rater reliability between LLM and human coding.

To determine the optimal amount of information for LLM coding, we systematically investigate the impact of incorporating various elements of the codebook, such as code definitions, keywords, coding rules, and examples, into the LLM instructions. Furthermore, we explore whether providing additional information beyond the codebook enhances the LLM's coding performance. This exploration allows us to identify the most effective combination of codebook-derived information and supplementary details to construct prompts that yield the best LLM coding results. We also fine-tune the model parameters and textual structures to achieve optimality, which allows for a more fine-grained modulation compared to previous works [134]. Finally, to ensure the interpretability of LLM coding results, we enable the LLM to output brief justifications for its assigned codes.

Once the optimal prompt is identified and high inter-rater reliability between human and LLM coding is achieved, we can proceed to code the remaining parts of the dataset not coded by humans using the LLM. This is followed by sampling to check if consistency between human and LLM coding is maintained [10], thus creating a thoroughly scrutinized dataset. Given that model predictions are substantially less time and cost prohibitive compared to human coding, this step provides the greatest relief from heavy workloads relative to traditional content analysis methods. Most of the coding work can now be done by the model rather than by human coders.

3.2.3 Human-LLM Disagreement. After the completion of both human deductive coding and LLM deductive coding (the two steps mentioned above), for each message in a subset of the data, there exists a human code along with multiple LLM codes corresponding to different prompts. Most messages may be coded with at least one LLM code that matches the human code.

We then identify the subset of messages where *all* LLM codes are mismatched with the human code. Any message with at least one LLM code matching the human code is excluded from the disagreement discussion because such messages likely indicate disagreement due to model suboptimality. Thus, we only select the messages in which none of the prompt variants enables agreement between LLM and humans.

Subsequently, for cases of disagreement between human and LLM codes (hereafter as *human-LLM disagreement*), all researchers independently review the data and engage in open discussion for further qualitative analysis, aiming to achieve more agreement. All coders are required to provide their opinions on each message involved in human-LLM disagreement:

- **Human Coding Error:** During human deductive coding, coders made incorrect decisions.
- **LLM Coding Error:** Alternatively, the human-LLM disagreement may stem from LLM error, indicating insufficient model performance leading to difficulty in coding certain messages, particularly those with complex or implicit patterns.

- **New Code:** It is also plausible that the human-LLM disagreement signals a novel theme that has not been discussed before. Both human and LLM have erred, indicating that the current coding scheme cannot cover the theme present in the message.

If the coders, after open discussion, suggest that the disagreement may shed light on a novel conceptual perspective, systematic inductive coding is conducted to determine the conceptualization of the patterns therein.

3.3 Collaborative Inductive Coding

The inductive coding process, rooted in grounded theory principles [60], involves interacting with data, conducting comparative analyses and synthesis [97]. The process facilitates the conceptualization of the empirical data. We adhere to a three-step coding procedure: *pre-coding*, *refinement of concepts*, and *grouping and concept formation*. The inductive coding protocol can be employed collaboratively with LLMs [39], leveraging their capabilities to generate codes from messages and synthesize themes from codes. To align with previous practices [89, 166], firstly, three coders independently record any piquing *observations* and recurring *patterns* from participants' messages through preliminary jotting and connecting related insights. This is followed by open discussions for collective review and conceptualization of these observations by sharing remarks about the commonalities among participants. Upon identifying an observation, a preliminary name is assigned, and this iterative process continues until all data have been examined. Subsequently, these findings and names are conceptualized into high-level concepts. Further discussions and affinity diagramming [72] are then conducted to determine the grouping of codes, conceptualizing these into sub-themes and ultimately forming high-level *hypotheses*.

These newly discovered themes and their sub-themes offer fresh insights that enrich our understanding of the concept under investigation. In sum, LLM's pivotal role enables us to pinpoint human-LLM disagreements through deductive coding and apply inductive coding to these messages to discover new themes. The proposed methodology harnesses the power of human-LLM collaboration to generate novel insights, thereby advancing qualitative analysis.

4 CASE STUDY: ADVANCING QUALITATIVE ANALYSIS OF MENTAL-ILLNESS STIGMA

We conducted a case study to perform a qualitative analysis on psychological constructs, using social stigma as a representative example, following the  CHALET framework. We structured our work into three primary steps, including chatbot-assisted data collection, human-LLM synergistic deductive coding, and collaborative inductive coding.

4.1 Attribution Model for Mental-Illness Stigma

Attribution model is one of the conceptualization used to understand human psychological constructs. Although the attribution model has had an enduring influence and frequent application over the past two decades, newer constructs have continually emerged. These constructs often providing fresh insights into social stigma toward individuals with mental illness that do not entirely overlap with the attribution model. Recent studies [71, 126, 149] demonstrate that different approaches and conceptual findings can emerge when investigating various populations. Consequently, it remains unclear whether the attribution model is controversial, inadequate, or ambiguous [121], thereby necessitating further qualitative analysis. Our work aims to obtain new revelations through the practice of hybrid qualitative coding, employing the paradigm of human-LLM synergy.

4.1.1 Attribution Model. In the mental-illness stigma context, the attribution model [32] is a well-designed and widely applied conceptual framework proposed to explore the factors contributing to the emergence of social stigma. Our case study builds on the foundation of the attribution

model, which includes three dimensions: personal responsibility belief, emotional responses, and behavioral responses. The attribution model suggests that people's perceptions of the extent to which a person with a mental illness is responsible for their condition can lead to stigmatized emotional responses, such as lack of pity, anger, and fear. These emotional responses can further elicit discriminatory behavioral responses, including coercion, segregation, social distance, and withdrawal of help.

Social stigma, originally defined by Goffman [61], is a mark or attribute that diminishes an individual, transforming them from a complete and ordinary person into one who is tainted and devalued. It is observed when an individual possesses or is perceived to possess a characteristic that implies a devalued social identity within a specific context [35]. It comprises stereotypes, prejudice, and discrimination [32]. Our qualitative analysis mainly focused on public stigma, which arises when a significant portion of the population endorses negative stereotypes [113]. Social stigma is a prevalent issue among individuals with mental illness [33], which hinders their opportunities, self-determination, and recovery [33, 127].

4.2 Step 1: Chatbot-Assisted Data Collection

The data collection procedure included a short online questionnaire, which was used to recruit participants for an interactive conversation session with an LLM-enabled *conversational agent* that represents a specific scenario in human-LLM communication. Data collection for this study ran between November 2023 and February 2024.

4.2.1 Vignette Design. *Vignettes* [107], brief evocative stories, are effective tools for eliciting opinions and attitudes toward individuals with mental illness in conversational agent-mediated social contact [94, 98, 143], as they provide detailed stimuli that resonate with participants' experiences [116] while avoiding excessively emotionally charged situations [65]. We designed a vignette based on DSM-5 [11], as shown in Figure 6, which focused on *Avery*, a person experiencing depressive symptoms. The vignette described Avery's symptoms without using technical or medical jargon, and did not explicitly specify Avery's gender or age. It illustrated how the symptoms negatively impacted various aspects of Avery's life, including academics, work, relationships, and interactions with others. To represent the most common symptoms, we avoided including more extreme symptoms such as self-harm or suicide.

4.2.2 Chatting Question Design. During the study, after the participants read the vignette, chatbot posed questions based on the vignette's content to understand their opinions. The strong capability and flexibility of LLMs could allow us to ask the questions in a vivid manner. To collect qualitative data for each attribution in the original attribution model [32] and then gain a deeper understanding, we designed multiple questions corresponding to the various attributions outlined in the model. Our chatting questions were adapted from the 27-item *Attribution Questionnaire* (AQ-27) [8, 30], which is frequently employed to measure people's stigmatization levels in both HCI and psychological practices [85, 98, 165]. The questions we used to collect qualitative data are presented in Table 1. Our materials were reviewed and collaboratively refined by a social work specialist within the research team and a professional psychiatrist. We adapt the original questionnaire to fit the context of our chatbot's questions. The detailed adaption can be found in Appendix A.1.3.

4.2.3 Strategies to Elicit Self-Disclosure. The conversation began with a small talk session to build rapport and increase the participant's willingness to disclose sensitive information. Next, the chatbot, named *Nova*, delivered a vignette about Avery's experiences with depression across multiple messages, intermittently asking participants for brief responses. After the vignette, Nova posed open-ended questions prompting participants to share their own related experiences and opinions.

Table 1. The **chatbot question scripts** designed for data collection. They were meticulously crafted, with each question corresponding to a specific attribution listed in the attribution model [32]. This approach was employed to obtain *paired data*, consisting of the participant's opinion and the associated attribution type, thereby facilitating a comprehensive analysis of the connections between the expressed opinions and their underlying attributional factors.

Question Script	Attribution Type
Just like what is mentioned in the story, Avery is currently facing difficulties in both their relationships with colleagues and their work performance. Do you believe Avery's current situation is primarily a result of their actions? Please share your thoughts.	Responsibility
Let's imagine you were one of Avery's friends at the party or other social occasions. Would you feel anger toward them due to their actions, for example, losing their temper and yelling at someone?	Anger
I'd like to know if you would feel concern and sympathy at Avery's mental state?	Pity
If you were traveling or engaging in activities with Avery, would you have any feelings of frightened or being threatened regarding the potential implications of your interaction?	Fear
I wonder if you were one of Avery's colleagues or friends, then would you be inclined to help them with tasks, such as work projects?	Helping
Apart from being a colleague or friend, I'm also curious if you are Avery's neighbor, then would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?	Coercion Segregation
If you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?	Social Distance

These core questions assessing stigmatized attitudes towards mental illness were presented with a strategically placed mid-break session halfway through to re-engage participants with lighthearted dialogue about their hobbies. This was designed to mitigate potential emotional burden and further encourage open disclosure of personal views.

Figure 3 illustrates our overall conversation flow design. During the delivery of each question, the chatbot employs several techniques [7] to facilitate disclosure, enabling better understanding and analysis. The first technique is the use of *follow-up questions* [70]. Since questions in Table 1 can be answered with a simple 'yes' or 'no', a 25-character threshold is set. If a response to emotional responses and responsibility attribution questions is under 25 characters, the chatbot asks a follow-up question to prompt for reasons. For the three behavioral response questions (coercion segregation, social distance, helping), the chatbot asks for reasons if the response is non-stigmatized. If the response is stigmatized or the participant has no idea, the chatbot asks about potential results. A 150-character threshold is also set; if the initial response plus the response to the follow-up question doesn't reach 150 characters, the chatbot asks an additional question about what characteristics led to the participant's conclusion. Figure 7 provides an example of a follow-up question.

The second technique is active listening skill [173], where the chatbot restates the points mentioned by the participants and expresses agreement or understanding. Additionally, the conversational agent engages in neutral self-disclosure, which aims to elicit participants' self-disclosure

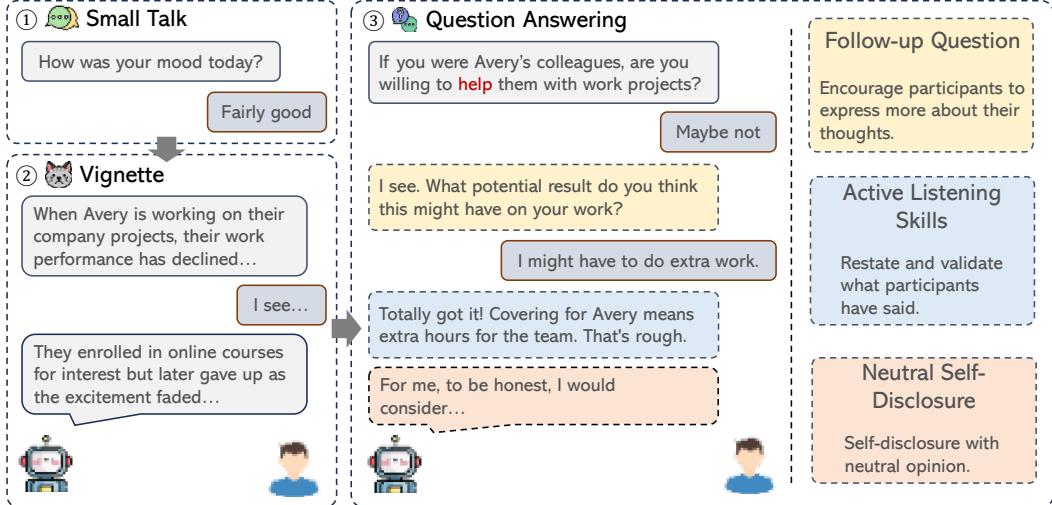


Fig. 3. Conversation flow design. The conversation consists of three main tasks: **small talk**, **vignette delivery**, and **question answering**. During the question answering task, the conversational agent employs three key techniques to facilitate effective communication and elicit informative responses from participants. First, **follow-up questions** are used to collect additional information about participants' attitudes [70]. Second, **active listening skills** are employed to engage participants and encourage them to provide high-quality responses [168]. Third, the conversational agent uses **neutral self-disclosure** to promote participant self-disclosure without influencing their attitudes [99].

[99], while ensuring that their original attitudes are not shifted. The conversational agent intentionally mentions both positive and negative aspects. For example, Figure 7 illustrates the neutral self-disclosure of *fear*, where the agent rationalizes fear on the one hand while also presenting valid reasons for not being fearful on the other. All neutral self-disclosure texts are presented in the sample conversation flow with light gray provided in Appendix A.1.1.

4.2.4 System Implementation. We utilized the UChat⁴ platform to build our LLM-powered conversational agent. We employed a hybrid approach for implementation, where the text delivered by the agent can originate from pre-written scripts or be generated by prompting an LLM. The small talk questions, vignette delivery, questions corresponding to the attribution model [32], and the chatbot's neutral self-disclosure for each question were pre-scripted and hardcoded into the system. Other text elements, such as follow-up questions and active responses to participants, were generated by LLM. Examples of our prompts for follow-up questions can be found in A.1.5. We employed gpt-4-1106-preview⁵ to generate our chatbot's text, setting the maximum tokens to 100 and the temperature to 0.2⁶. Our interface was embedded into the Qualtrics⁷ study. We ensured that the concurrency rate did not exceed 50.

⁴<https://uchat.au/>

⁵<https://platform.openai.com/docs/models/overview>

⁶We experimented with different temperature values and discovered that higher temperatures caused the model to generate a wider variety of follow-up question texts. However, this led to less controllable questions, which confused some participants in our pilot study. Therefore, we opted for a lower temperature to ensure that the model consistently generated similar and reliable questions.

⁷<https://www.qualtrics.com/>

Table 2. Participant Characteristics. **Mental illness experience** refers to whether participants had close friends or family members affected by mental illness.

		ALL (N=1,002) n (%)
Mental Illness Experience	Yes	526 (52.5)
	No	299 (29.84)
	Maybe	177 (17.66)
Age	21-24	70 (6.99)
	25-34	234 (23.35)
	35-44	211 (21.06)
	45-54	155 (15.47)
	55-64	131 (13.07)
	65+	201 (20.06)
Gender	Male	540 (53.89)
	Female	460 (45.91)
	Prefer not to say	1 (0.1)
	Prefer to self-describe: Trans Man	1 (0.1)
Ethnicity	White	695 (69.36)
	Black or African American	210 (20.96)
	Asian	47 (4.69)
	Mixed	26 (2.59)
	American Indian or Alaska Native	6 (0.6)
Education	Other (Hispanic, Chicano, etc)	18 (1.8)
	Less than Primary	2 (0.2)
	Primary	24 (2.4)
	Some Secondary	14 (1.4)
	Secondary	139 (13.87)
	Vocational or Similar	104 (10.38)
	Some University but no degree	190 (18.96)
	University - Bachelors Degree	328 (32.73)
	Graduate or professional degree (MA, MS, MBA, PhD, Law Degree, Medical Degree, etc)	197 (19.66)
Prefer not to say		4 (0.4)

4.2.5 Participants. We recruited participants from diverse backgrounds, including various age groups and educational background, through online platforms. During the recruitment process, we clearly outlined the study's general scope, duration, and participants' right to withdraw at any point. Our recruitment criteria regulated that all participants must 1) be 21 years or older, 2) have English as their first language, 3) be willing to read materials related to mental illness, and 4) participants were required not to have immediate and urgent mental health concerns. We incorporated the last criterion due to the potential risk that the vignettes about mental illness could cause distress to individuals grappling with such issues [137]. Ultimately, we recruited 1,002 participants, none of whom reported having an ongoing mental illness. The average age of the participants was 46.38 ($SD = 16.45$). Detailed participant characteristics are presented in Table 2.

4.2.6 Procedure. The *one-time* study lasted approximately 30 minutes, for which the participants received a reimbursement of about 6.3 USD for their time. Initially, we obtained participants' consent, informing them that their personal data would remain confidential. Participants were

forewarned that the study would involve a mental illness-related scenario and were given the option to quit if uncomfortable. Subsequently, the participants were asked basic demographic questions. The participants then engaged in a conversation with the conversational agent for approximately 20 minutes. After the conversation, a debriefing session was conducted, during which participants were informed about common types of stigma, and the researchers' objective of assessing participants' stigma attitudes was explained.

4.3 Step 2: Human-LLM Synergistic Deductive Coding

4.3.1 Step 2.1: Human Deductive Coding. To ensure reliable coding results and mitigate human bias, we implemented *triangulation*. We had multiple diverse researchers independently code and analyze the data, and collected data from different times, spaces, and groups. These triangulation approaches effectively reduced the impact of inherent human bias on the coding results. Firstly, two coders independently conducted *pilot coding* using the initial codebook, coding 70 messages, with 10 messages for each attribution. After independently coding, we engaged in open discussions about these 70 messages, reached agreements, and subsequently established initial and applicable rules.

We then proceeded to the *formal coding* stage, involving a third coder and providing necessary training. The codebook refinement followed, where we randomly selected 700 messages from 100 participants for coding. Two coders independently coded the messages and set up frequent checkpoints: after coding 10 participants (Cohen's $\kappa = 0.55$), then another 10 participants (Cohen's $\kappa = 0.53$), followed by four sets of 20 participants each (Cohen's $\kappa = 0.66, 0.79, 0.76$, and 0.72 , respectively). Coders openly discussed the codebook to ensure consistency, adding empirically-derived rules, revising definitions, and including keywords for each code. Finally, when the inter-rater reliability among coders reached a satisfactory level, we continued meeting with the specialist in the research team to finalize the codebook. We followed the specialist's advice and set up *longer* checkpoints after finalization: coding 50 participants (Cohen's $\kappa = 0.74$), then another 50 participants (Cohen's $\kappa = 0.66$), followed by 100 participants (Cohen's $\kappa = 0.69$), and two sets of 150 participants each (Cohen's $\kappa = 0.69$ and 0.66 , respectively). We completed the coding of 600 participants' messages, totaling 4,200 messages (Cohen's $\kappa = 0.71$). Upon completion, we consulted with the specialist in the research team to confirm the codes for messages where the three coders had not reached an agreement. This concluded the human deductive coding process, outputting both the human codes and the development of the coding scheme.

4.3.2 Step 2.2: LLM Deductive Coding. Henceforth, based on the aforementioned human deductive coding established coding scheme, we enabled the LLM to follow the proposed coding scheme for deductive coding and evaluate its performance by coding the same messages as the human coders. To demonstrate the performance and to find the prompt that maximizes the inter-rater reliability between LLM and human coders, we conducted experiments around prompt design, drawing inspiration from previous practices [44, 75, 83]. We designed the following three aspects of exploration and comparison, with all experiments using the inherently high-performing gpt-4-1106-preview. Previous research on prompt design is inspired by techniques such as role prompting [134, 158], where a role is assigned (e.g., “*Your role is a competent coder for social stigma toward mental illness*”). We applied this technique consistently in our experiments.

Which aspects of the codebook are necessary when leveraging LLM for deductive coding?

Firstly, as the prompt content originates from the human-written codebook, the components can be divided into several parts: *code names*, *keywords*, *additional rules* (refer to Section A.2.1), and *examples* for all codes. Additionally, it includes the vignette as background knowledge, the question from Table 1, and the participants' messages. We investigated whether all information, originally

provided to humans in the codebook, is useful for LLM deductive coding and if each piece of information contributes to increased inter-rater reliability between the LLM and humans, or if the presence of certain information may lead to information overload and have a negative effect. Therefore, we systematically controlled the presence or absence of each component within the codebook. Furthermore, we compared the performance under two distinct information quantity scenarios: prompts containing *all code* information (eight codes) and prompts containing only the *target code*. The latter approach was inspired by previous work [44], where we provided only the code corresponding to the question associated with the message in the prompt, along with the *Non-stigmatized* code. All other attributions of stigmatization were categorized as *Stigmatized (others)*, effectively modeling the coding process as a three-class classification task.

Is there a need for additional information, and if so, how impactful is its inclusion?

Secondly, we explored whether additional content beyond the codebook is necessary and if it can further improve performance [151]. Previous research has demonstrated the usefulness of *chain-of-thought* prompting [164], where a complex reasoning task is decomposed into sub-tasks, and the model is instructed to perform the sub-tasks sequentially. The original human-written codebook does not provide the reasoning process. We investigated whether the reasoning process can help the LLM code messages more consistent with humans. Additionally, in-context learning has proven beneficial in NLP tasks [41], so we also explored whether increasing the number of examples is helpful. Moreover, previous papers have shown that the order of examples in the prompt can affect LLM performance [111]. Therefore, we conducted this part of the experiment under the condition of only having target-code information, i.e., adjusting the order of examples for the three codes: *Stigmatized*, *Non-stigmatized*, and *Stigmatized (Others)*, to investigate the impact of example order.

Are there any additional settings of the model that are impactful? Lastly, we experimented on whether different model parameters or other LLM settings, unrelated to the amount of information, significantly affect the results. Inspired by Ziems et al. [175], we explored the impact of the order of instruction and content on the results. Following [134], we investigated the effect of the temperature settings on the results. Similarly, we explored and compared three temperature conditions: 1, 0.5, and 0. By integrating these various aspects of exploration, we searched the prompt space for the optimal prompt to achieve optimal performance and find the highest agreement between the LLM and human coders, thereby facilitating the interpretive capabilities of LLM coders and minimizing the impact of model suboptimality on the analysis.

Finally, when the agreement between the LLM and human coders reaches a satisfactory level, we can utilize the LLM to code the remaining data. As previous works have found that using LLMs for qualitative coding may introduce bias [10, 174], after completing the LLM deductive coding on data not previously coded by humans, we select 25 messages from each attribution, maintaining the deductive approach outlined in Section 4.3.1, and perform human deductive coding to enhance validity.

4.3.3 Step 2.3: Human-LLM Disagreement Analysis. After identifying the optimal agreement (represented by the highest Cohen's κ) between LLM deductive coding and human deductive coding, indicating that we have achieved a sufficiently high level of agreement, we shifted our focus to the human-LLM disagreement set to further investigate the underlying causes of disagreement.

The previous LLM deductive coding resulted in various prompt designs, corresponding to different coding outputs from the LLM. For each message, there was one human-provided code and multiple codes generated by the LLM. We thereby identified the set of messages where the human code differed from *all* LLM codes across the various prompts. This approach was taken because even the optimal LLM deductive coding result (which had the highest agreement with humans) may have

instances where other coding results with lower human agreement can align with the human code, while the optimal result itself cannot. In other words, despite having the highest overall agreement with humans and undergoing prompt optimization and selection, the optimal LLM deductive coding result may still exhibit suboptimality, leading to disagreements that could be resolved by other LLM results with different prompts. By taking the *intersection* of human-LLM disagreements across all LLM deductive coding results, we minimized the impact of model suboptimality on our disagreement analysis.

Drawing upon the disagreement resolution methods proposed by Chinh et al. [23], we leveraged an open discussion approach to resolve human-LLM disagreements. We conducted the disagreement-resolving process for each attribution in sequence, engaging in an open discussion throughout to check for any discrepancies in understanding among coders. At this step, three coders independently code the messages and determine whether the disagreement arises from *human coding error*, *LLM coding error*, or the emergence of a *new code*. If the coders, after engaging in open discussion, believe that the disagreement may indicate a new code, it is retained for further exploration in the next step.

4.4 Step 3: Collaborative Inductive Coding

Following the paradigm outlined in Section 3.3, we conducted inductive coding on messages identified as potentially containing new codes within the human-LLM disagreement dataset. This process was carried out by three coders to ensure reliability. Inspired by prior research on leveraging LLMs for inductive coding [87], we explored two paradigms: 1) using LLMs to autonomously generate new themes and code, and 2) a human-LLM collaborative approach. After expert review of the new findings produced by each paradigm, we chose to present the results from the human-LLM collaborative paradigm as our final output.

To facilitate collaborative inductive coding, we first review all the data, then independently *pre-coded* by taking notes and assigning labels. We refined our notes and reached agreement on the codes through open discussions. The agreed-upon codes were incorporated into the affinity diagram (Appendix A.5.2). These sessions allowed for the development and *refinement of concepts* as we identified emerging *patterns* from our *observations* and from LLM-provided output. The iterative refinement process involved evolving patterns as more participant messages were analyzed.

After compiling codes, we engaged in *grouping* by identifying commonalities and combining similar codes. Throughout several rounds of defining themes and refining code definitions, we engaged in *concept formation* based on the codebook. Based on the data patterns and the input codes provided, our conceptualization process involved a collaborative approach that combined human researcher's ideas with LLM-generated ideas. Following the attribution model, we categorized the codes into cognitive judgment, emotional responses, and behavioral responses. Finally, we discussed our findings, derived from the human-LLM collaboration hybrid qualitative analysis, with domain specialists to validate and finalize our novel insights.

5 CASE STUDY RESULTS

5.1 Chatbot-Assisted Data Collection (Q1)

To answer Q1, we analyzed the overall situation and quality of the qualitative data we received. The average time for participants to complete the chatbot conversation was $t = 17.63$ ($SD = 8.18$) minutes. We assigned one question, along with follow-up questions (if any), to each attribution. When answering questions, participants typed an average of 43.82 ($SD = 14.68$) words for responsibility, 43.4 ($SD = 15.49$) for social distance, 40.86 ($SD = 13.45$) for anger, 41.5 ($SD = 14.2$) for helping, 39.3 ($SD = 13.98$) for pity, 40.18 ($SD = 13.7$) for coercion segregation, and 39.74 ($SD = 14.14$) for fear,

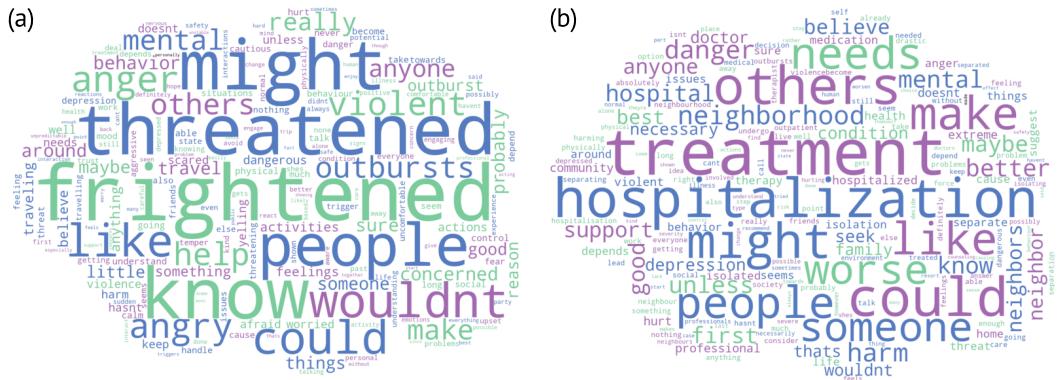


Fig. 4. Word frequency analysis on two attributions: (a) *fear* and (b) *coercion segregation*.

showing they expressed substantial opinions. The average informativeness score [169] of the data collected through our chatbot is 87.34 ($SD = 35.80$), indicating a high level of richness. Details on the calculation of the informativeness score can be found in Appendix A.1.6. The specific collected data from participants are presented in Table 6.

We further validated the quality of the collected data by examining the word frequency in each attribution. As shown in Figure 4, we found that participants' discussions with the chatbot closely revolved around the questions posed for each attribution. The frequent appearance of words directly related to the scenario and question, as well as the presence of words reflecting complex emotions and thinking, indicated that our LLM-enabled chatbot effectively facilitated rich conversations about the social stigma of mental illness. Moreover, the results revealed the participants' engagement and reflection on the topics and issues presented in the scenarios. Further analysis of word frequency is provided in Appendix A.1.7.

Overall, the data provides a rich and multifaceted exploration of social stigma toward mental illness. The insights and experiences of the participants offer valuable perspectives on this important issue. This further demonstrates that our LLM-enabled chatbot is effective for data collection and can collect diverse, multi-perspective, and rich qualitative data.

5.2 Human-LLM Synergistic Deductive Coding

5.2.1 Human Deductive Coding. Of the total 4,145 messages we coded, 46.00% ($n = 1,911$) of the responses included stigmatized attitudes. Among the stigmatized codes, the most prevalent was *social distance*, accounting for 9.14% ($n = 379$) of all codes. *Responsibility* followed closely at 9.51% ($n = 394$), and *fear* made up 8.88% ($n = 368$) of the total codes. *Anger* and *coercion segregation* constituted 7.19% ($n = 298$) and 6.56% ($n = 272$) of the codes, respectively. The least common stigmatized codes were *helping* at 3.81% ($n = 158$) and *pity* at 1.01% ($n = 42$). In addition, we achieved the following inter-rater reliability scores (Cohen's κ) for human deductive coding on each attribution: *responsibility* (0.71), *anger* (0.75), *pity* (0.36), *fear* (0.69), *helping* (0.56), *coercion segregation* (0.71), and *social distance* (0.7).

In our deductive coding process, we established rules and achieved consensus among coders through discussions with specialists. We also defined each stigma attribution according to the original model (Appendix A.2.1). Our coding scheme includes seven stigmatized and one non-stigmatized

thought, totaling eight codes. The non-stigmatized code represents texts explicitly convey positive sentiments, like sympathy and assistance. For example, P383 considered the hospitalization suggestion offensive, reflecting a non-stigmatized view against coercion and segregation.

"No, I think that is a deeply offensive and appalling suggestion. I think that is a strange question to ask. The real question is what right would anyone have to take her Avery away or hospitalise her against her will? She is harming no one." (P383)

Table 6 presents the quotes for each stigmatized code. Throughout this, we provide a detailed practice on human deductive coding in the attribution model.

5.2.2 LLM Deductive Coding (Q2). To answer Q2, we conducted three explorations, investigating the utility of the information in the codebook, whether additional information was needed, and the influence of different settings. The findings of these explorations are presented as follows to demonstrate the impact of information amount differences on LLMs' interpretation of the content to be coded, which, in turn, provide guidance on enhancing LLMs' ability to analyze the materials we aim to code effectively.

Overall, all information in codebook proves to be valuable - We present the results in Table 3. First, comparing the scenarios where all codes are provided versus only the target code, we observe that information from other codes can assist the LLM in investigating the overall context, potentially improving performance. Consequently, information from other codes is not regarded as noise for the target code.

Second, by examining the impact of different amounts of information from the codebook, we focus on scenarios where only the target code information is available. Vignettes have a minor effect, negatively correlated with the performance of codes that have lower human agreement (*pity, helping*) or lower LLM agreement (*coercion segregation*), and positively correlated with the performance of easily classifiable codes. Rules play a vital role for all codes, particularly for attributions with lower human agreement (*pity, helping*). Inclusion of rules can enable the human-LLM agreement to surpass the human-human agreement. Moreover, both keywords and examples contribute slightly to the improved performance of all codes.

However, there are two exceptions: *pity* and *coercion segregation*. Notably, for *pity*, information from other codes can act as noise. This is empirically observed due to the low sample size with the code *Stigmatized (Pity)*, accounting for only 32 out of 592 instances, and the low human-human agreement, suggesting that the LLM, like a human coder, struggled with *edge cases* and ambiguous situations, especially where implicit information was required to make a judgment. For example, "*Concern for his situation and we should all be humble to them always. Suddenly anger.*" (P523) is a participant's response to the question about whether they feel pity for Avery.

Coder 1 labeled the message as *Stigmatized (pity)*, suggesting that the pity shown was not genuine, but due to societal expectation and obligation. Coder 2, however, saw it as *Non-stigmatized*, viewing the pity as sincere. After discussion, it was agreed to classify the message as *Non-stigmatized*. Despite this, there was a sense of unease, indicating a borderline case between stigma and genuine pity. The coding notably differed from LLM results, where 21 prompts identified it as *Stigmatized (pity)* and six as other forms of stigma. This discrepancy highlights the controversial nature of this message, with both human coders and LLMs struggling to reach a consensus. This also underscores the need to revisit and refine our coding categories to better capture such complex cases.

For *coercion segregation*, integrating external code information enhances performance, but additional codebook details may add noise. Excessive information leads to incremental performance gains once again. Vignettes and rules reduce agreement levels. One possible reason is that human coders rely on more than just the codebook; their understanding benefits from collaborative codebook creation, discussions, and oral guidance as well. In contrast, LLMs lack this interactive context

Table 3. Cohen's κ between LLM codes and human codes. Each column represents the agreement between the LLM and human coders across attributions using prompts with the same rules. **All-code info** indicates that the prompt provides an equal amount of information for all codes, while **target-code info** indicates that the prompt only provides information for the corresponding attribution, with other attributions collectively coded as *Stigmatized (Others)*. The *name* column indicates that the prompt only includes the code name, *+vig* indicates the addition of the vignette as background knowledge, *+rule* indicates the inclusion of additional rules, *+keyword* represents the addition of keywords, and *+exp* indicates the inclusion of one example. The column highlighted in **red** represents the column with information that completely aligns with the human-written codebook, albeit with a different format. The **bold** grids represent the highest agreement between human and LLM for each attribution. The last row presents the combined total results across all attributions.

	All-code info					Target-code info				
	name	+vig	+rule	+keyword	+exp	name	+vig	+rule	+keyword	+exp
Responsibility	0.68	0.67	0.63	0.64	0.69	0.61	0.63	0.62	0.63	0.64
Anger	0.57	0.48	0.6	0.62	0.65	0.3	0.31	0.51	0.52	0.54
Pity	-0.04	-0.03	0.42	0.45	0.46	0.13	0.12	0.44	0.48	0.5
Fear	0.69	0.66	0.63	0.67	0.71	0.58	0.61	0.68	0.69	0.69
Helping	0.15	0.22	0.68	0.69	0.69	-0.03	0.01	0.58	0.62	0.64
Coercion	0.58	0.53	0.49	0.51	0.54	0.26	0.24	0.35	0.38	0.35
Segregation	0.59	0.67	0.73	0.74	0.76	0.25	0.34	0.59	0.54	0.61
Total	0.27	0.52	0.71	0.72	0.75	0.28	0.3	0.58	0.59	0.6

and rely solely on written descriptions. Some implicit knowledge obvious to the original coders might be missing from these descriptions [44].

The inclusion of extra information beyond human codebook minimally affects LLM-human coding agreement - Table 4 shows changes in LLM-human coding agreement when additional information is used with the human-written codebook versus the exact codebook. Each cell value indicates the difference in agreement for each setting compared to the baseline in Table 3 (the red column). Generally, adding examples, extra reasoning, or changing example order shows minimal impact on LLM performance. However, for easily coded attributions like *responsibility*, *anger*, *fear*, *coercion* *segregation*, and *social distance*, more examples slightly enhance performance.

An anomaly in *helping* shows that adding more examples or detailed reasoning decreases performance. This may occur because additional examples extend the input, overwhelming the model with too much information, making it hard to identify crucial details. Additionally, the model's need to handle multiple tasks like natural language understanding (NLU), reasoning chain planning, and natural language generation (NLG) simultaneously complicates learning.

Table 4 shows that example permutation minimally affects coder-LLM agreement, yet we explored its effect on prediction distribution. This further analysis focuses on the effect of permutation under only target-code information. We quantified the number of predicted samples for each code when providing examples in different orders in prompt. Figure 12 reveals that codes at the *beginning* are predicted *more* frequently. For example, positioning the *Stigmatized* example first results in 278.5 predicted *Stigmatized* samples, compared to 240.5 when it's last. This pattern holds for other attributions too.

Temperature has little effect; placing instructions before content when prompting LLMs improves LLM-human coding consistency - Figure 5 displays the key insights include: temperature has a minimal impact on LLM performance, with lower temperatures slightly enhancing

Table 4. Differences in Cohen’s κ . The differences are between conditions with different additional information and the condition with information equivalent to the human-written codebook. **CoT** represents chain-of-thought, while **NoCoT** indicates the absence of chain-of-thought. **S** represents *Stigmatized*, **NS** represents *Non-stigmatized*, and **O** represents *Stigmatized (Others)*. **S_NS_O** denotes the order of the three examples as *Stigmatized*, *Non-stigmatized*, and *Stigmatized (Others)*. The other notations similarly indicate other possible permutations. **Purple** indicates that the information content represented by the cell leads to higher LLM-human coding agreement compared to using the human-written codebook as instructions, while **green** indicates lower agreement. Darker colors signify larger differences.

		+ CoT & + More Exps													
		S_NS_O		S_O_NS		NS_S_O		NS_O_S		O_S_NS		O_NS_S			
		CoT	NoCoT	CoT	NoCoT	CoT	NoCoT	CoT	NoCoT	CoT	NoCoT	CoT	NoCoT		
Responsibility		0.03	0.06	0.03	0.04	0.04	-0.02	-0.02	-0.03	0	0.02	0.02	-0.01	0.02	
Anger		0.03	0.11	0.09	0.13	0.1	0.08	0.08	0.08	0.09	0.1	0.1	0.07	0.08	
Pity		0.05	0.01	-0.02	-0.01	-0.07	0	-0.08	0.03	-0.06	-0.05	-0.1	-0.01	-0.08	
Fear		0.02	0.02	0.01	0.02	0.03	0.02	0.03	0.02	0.01	0.02	0.01	0.02	0.02	
Helping		-0.02	-0.12	-0.09	-0.18	-0.12	-0.15	-0.13	-0.14	-0.09	-0.15	-0.09	-0.16	-0.09	
Coercion		-0.04	0.05	0.04	0.06	0.08	0.02	0	0	0	0.03	0.02	0.01	0.05	
Segregation		0.03	0.07	0.06	0.03	0.06	0.03	0.05	0.04	0.05	0.02	0.07	0.07	0	
Social Distance		Total	0.01	0.02	0.03	0.01	0.02	0	-0.01	0	0	0	0.01	-0.01	0.01

outcomes. Additionally, introducing *instructions* at the beginning of prompts, followed by the user *content*, improves performance. These findings highlight the need for tailored instruction design in LLMs, guided by human-created codebooks. Further details on prompt design are in Appendix A.3.

Our studies show that coding complexity varies for both humans and LLMs across different qualitative analysis codes. Providing LLMs with the full codebook information, including additional rules and examples, etc., is crucial for LLM coding. For challenging codes, extra rules are necessary and can improve performance beyond human coding. Conversely, simpler codes may only need basic definitions. All code information serves as valuable data that aids the LLM in analyzing what it should *not* learn and classify as the target code. Supplementary information beyond codebook is generally not essential, although a slight impact is observed for codes that are easy for humans to code. Multiple examples alter sample predictions but do not significantly change overall results. Temperature shows no significant effect, and initial clear instructions are advised for LLMs. Ultimately, our research suggests prompt designs that enhance LLMs’ ability to interpret human psychological constructs, leading to more accurate LLM-assisted coding aligned with human interpretations and enriching our understanding of societal and psychological dynamics.

Ultimately, after optimizing the LLMs’ ability to interpret and apply the coding scheme, we achieved results that exhibit the highest agreement with human deductive coding. By consolidating all the aforementioned findings, we identified the optimal agreement (Cohen’s κ) between humans and LLMs for each attribution: responsibility (0.7), anger (0.65), pity (0.55), fear (0.71), helping (0.69), coercion segregation (0.58), and social distance (0.76). Upon discovering the optimal prompt for each attribution, we utilized these prompts to code all messages in the dataset that had not been previously coded by humans. Finally, we randomly selected 25 messages from each attribution for validation, resulting in a human-LLM agreement (Cohen’s κ) of 0.87, demonstrating the reliability of LLM coding.

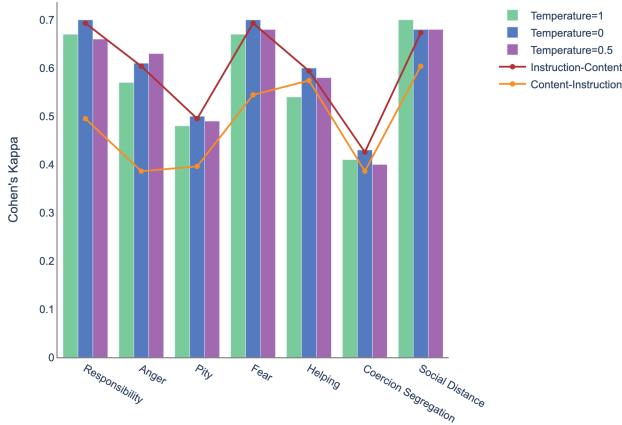


Fig. 5. The impact of different temperatures and the order of content and instructions on the agreement between humans and LLMs (Cohen's κ).

5.3 Human-LLM Disagreement Analysis (Q3)

To address Q3, which focuses on the selection of the human-LLM disagreement set, we minimized the selection range to reduce the impact of LLMs' suboptimal performance. The aforementioned exploration (Section 5.2.2) involved 23 different prompts, corresponding to 23 different LLM coding results. For each message, there is one human code and 23 codes provided by the LLM. We thus identified the set of messages in which the human code differs from all the codes generated by the LLM across various prompts. Of the 4,145 messages, 273 (6.5%) have human codes that are inconsistent with all LLM-generated codes under different prompt conditions. This includes 32 messages for responsibility, 51 for anger, 18 for pity, 47 for fear, 13 for helping, 88 for coercion segregation, and 24 for social distance.

After analyzing the human-LLM disagreement set, we identified the reasons for these discrepancies. Initially, we observed that most disagreements arose when the LLM consistently labeled messages as *Non-stigmatized* and humans as *Stigmatized*. Subsequently, three coders independently classified these into human coding errors, LLM coding errors, and new code emergence. Discussions were held post-coding to finalize attributions.

During the open discussion, we uncover valuable insights through qualitative analysis. As illustrated in the following example, P242's response highlights the complexity of identifying stigma:

"I would feel embarrassment more than anger. But I would stand up for them and explain to the others of Avery's condition so they would have some sympathy. I would also like to have a chat to Avery after they cool down and explain how their emotions might cause a damper on the mood of a social gathering." (P242)

P242 mentioned feeling more embarrassment than anger. Our human code indicated *anger* attribution related to stigmatization, and we initially agreed that the latter part expressed pity and willingness to help. However, embarrassment remains an uncomfortable description, possibly indicating the participant's unease with the situation occurring during the gathering. Considering this implicit aversion, we coded it as *Stigmatized (anger)*.

On the other hand, eleven prompt variants led the LLM to classify the message as *Stigmatized (pity)*, interpreting embarrassment as a lack of pity, while ten variants resulted in a *Non-stigmatized*

classification, focusing on willingness to help and pity. No consensus emerged between human and LLM codes. In discussions, we rejected the notion that embarrassment equates to anger or a simple lack of pity, challenging the LLM's classifications. We also dismissed *Non-stigmatized* as an inaccurate code, acknowledging the imprecision of our initial judgments. We concluded the message expressed a stigmatized emotion not captured by existing models, prompting further analysis and solution development.

We examined messages labeled as *Non-stigmatized* by LLMs and *Stigmatized* by humans to identify discrepancies. Our analysis revealed mutual linguistic features, detailed in Table 5, that explain divergences in LLM and human coding. LLM coders may have overlooked that language like *someone else might think* could represent personal views, a nuance human coders recognized due to their social awareness and interpersonal experiences. In addition, LLMs struggled to differentiate between descriptions of a *potential outcome* and perceptions of the *current state*. They also found it challenging to gauge the intensity differences between *need* and *suggestion*, which is crucial for judgment. Furthermore, distinguishing whether stigma was aimed at the vignette's *characters* or the entire *group* was also difficult.

These findings reveal language nuances that LLMs, with limited social awareness, might miss in qualitative coding of psychological constructs. Our research advances this analysis by emphasizing the need to consider subtle linguistic differences overlooked by LLMs.

After careful analysis, we identified a total of 273 messages that fell into three distinct categories: 51 messages indicating human coding errors, 41 messages revealing LLM coding errors, and 181 messages suggesting the need for new codes.

5.4 Collaborative Inductive Coding (Q4)

To address Q4, we conducted an inductive coding process to identify new codes absent in the initial model, aiming to resolve human-LLM disagreements and enhance our understanding. We explored two approaches⁸: 1) autonomous LLM coding, and 2) collaborative human-LLM coding, with humans leading and integrating LLM ideas. We assessed these paradigms by comparing results and expert evaluations. Our primary findings stem from the collaborative approach, while autonomous results are detailed in A.5.1.

Our findings reveal that the codes generated by autonomous LLM inductive coding provide a superficial summarization of the literal meaning, lacking the awareness of multi-dimensional sub-theme conceptualization. Moreover, the LLM failed to comprehend the influence of cultural background and social norms, while also exhibiting a deficiency in critical thinking and the ability to perform multiple rounds of scrutinization, which human coders are capable of. This comparative result demonstrates that though LLMs are capable of reviewing natural language, human researchers still need to play a leading role in inductive coding due to their life experiences, capacity to capture little nuances in messages, subjective judgment and empathy capabilities. Despite the powerful abilities of LLMs, the human element remains crucial in this qualitative analysis.

Based on the hierarchical structure of the original attribution model, we structured our new findings into three dimensions: cognitive judgment (adapted from the original term *cognitive mediator*), emotional responses, and behavioral responses. Under the cognitive judgment dimension, beyond responsibility, we discovered five sub-themes: self-neglect, vulnerability, crime-prone tendencies, minimization, as well as disconnection and functional incapability. In the emotional responses dimension, we identified four new sub-themes: apathy, frustration, embarrassment, and compassion fatigue. Lastly, within the behavioral responses dimension, our inductive coding process revealed three new, more implicit [118, 130] sub-themes alongside the explicit attributions

⁸Using gpt-4-1106-preview.

Table 5. Linguistic features and content characteristics that lead to coding discrepancies in stigma identification.

Linguistic Features	Fea-	Quote	Human Code
'Someone Else Might Think'		Neighbors may find it hard to understand Avery's outbursts and strange behavior if they do not know them very well. I think you jumped a step. Surely they could talk to someone regularly before hospitalization becomes necessary. (P189, paternalism)	Paternalism
Over-conjecture & Misconception	Over- conjecture	It would depend on whether he was impacting the neighborhood. I think he would mostly keep to himself. (P47)	Disconnection and Functional Incapability
	Misconception	If they are not threatening to me or others then I am comfortable. She isolates herself from others so she would not feel a need to interact with them. (P144)	Disconnection and Functional Incapability
Need & Suggestion	Need	I am not sure about that. It is very complex. Avery has a lot going on. She definitely needs to talk to someone. No weakness there. (P550)	Paternalism
	Suggestion	As long as they are not hurting anyone, I would suggest that they seek treatment to help them feel better. I doubt if they need to be hospitalized. Their behavior is upsetting to them and to others, but they have not actually tried to physically hurt anyone. (P460)	Non-stigmatized
Individual-level & Stereotypical-level	Individual-level	So far other members of the house are not at risk. Also, isolating him might worsen his situation. He could start having suicidal thoughts thereby hurting himself. (P282)	Self-Neglect
	Stereotypical-level	Certainly, people with depression tend to face such difficulties and challenges. They find it quite hard concentrating and seeing things through. They easily get distracted. And if not carefully managed, it progresses and gets worse and may lead to suicidal thoughts creeping into their subconscious. (P187)	Self-Neglect

of helping, coercion, segregation, and social distance. These implicit sub-themes include differential support, paternalism, and condescension.

The results and quotes from participants are reported in Table 6. The detailed definitions of themes, sub-themes, and codes can be found in Appendix A.5.3. These new findings deepen our understandings of the attribution model, which initially only included explicit stigmatized thoughts. Our advanced qualitative analysis facilitates a more comprehensive conceptualization of social stigma by incorporating more implicit stigmatized thoughts [48, 62], including microaggressions [13, 58, 63] and other less easily detectable attributions. By unraveling these intricate facets of stigma, we reconciled human-LLM disagreements, such as those shown in Table 5, by providing new conceptualizations. These novel concepts contribute to a more nuanced understanding of how people stigmatize mental illness.

Table 6. New attributions for mental-illness stigma. The green sub-themes represent those found in the original attribution model.

Theme Code)	(Sub-theme, Quote
Theme 1: Cognitive Judgment	
Responsibility	I really cannot say as I do not have the full details about Avery's life, but if their current situation is triggered by maybe a habit of theirs, then they need to step up and do better. (P45)
<i>Sub-theme 1.1: Self-Neglect</i>	
Code 1.1.1: Think they will suicide and do self-harm	Certainly, people with depression tend to face such difficulties and challenges. They find it quite hard concentrating and seeing things through. They easily get distracted. And if not carefully managed, it progresses and gets worse and may lead to suicidal thoughts creeping into their subconscious. (P187)
Code 1.1.2: Think they cannot take care of themselves very well	Not scared of Avery, just concern for their well-being. I might be concerned about whether or not they would show up and if they would be prepared, dressed appropriately and such. I might also take on responsibility for booking things if I were worried they would not get that kind of task done. (P333)
<i>Sub-theme 1.2: Vulnerability</i>	
Code 1.2.1: Think they will be harmed by others	I would feel worried for them regarding their safety. Others may retaliate to the anger Avery displayed. (P394)
Code 1.2.2: Think they will bring trauma	Not frightened or threatened but maybe triggered. Triggering my own depression because of being around someone who is in a completely "down" state of feeling. (P600)
<i>Sub-theme 1.3: Crime-Prone</i>	Certainly not. It is not a humane way of dealing with someone with a mental illness like depression specifically. If they are still of sound mind and can make decisions for themselves they have the ability to turn their lives around. However, if their depression was making them act violently toward others or influenced them to partake in criminal behavior then yes I agree they should be hospitalized, only in an extreme case. (P144)
<i>Sub-theme 1.4: Minimization</i>	
Code 1.4.1: Think they will deny that they have a mental illness	I would let them know I will be there to assist if needed. Help may be declined. They might not think they have a problem. (P270)
Code 1.4.2: Think they have no problem and can recover themselves	I probably would, but I would want to stress that they should feel free to talk to me about any issues. Avery seems like a good person. They enjoy learning. They currently have problems, but I feel those can be overcome. (P637)
<i>Sub-theme 1.5: Disconnection and Functional Incapability</i>	
Code 1.5.1: Think they are not well-integrated into society	While feeling concerned for well-being and the potential impact on the neighborhood, I might also be uncertain about the appropriateness of separating Avery and advocating for hospitalization without a clear understanding of their specific needs. (P435)

Theme Code)	(Sub-theme, Quote
Code 1.5.2: Think they are incapable of functioning normally	Assuming there are no issues with paying bills, I see no reason why not. Depression or depressive episodes on their own do not bother me to a point where I would deny the person to be a tenant, but if the situation will lead to an inability to fulfill payments, then that would be an issue. (P420)
Theme 2: Emotional Response	
Anger	To a certain extent, I will feel a bit angry because I want all my guests to have a good time, but I try to make people understand Avery's behavior. They need help and should be empathized with. (P64)
Pity	Unlikely. I have my own concerns and Avery may not need my input. I do not want to get involved. (P148)
Fear	Yes definitely. I will need to have more information on what triggers her episodes. When she is in a crowded place or if she sees joyful people. (P154)
<i>Sub-theme 2.1: Apathy</i>	It is not about separation or not, but what the doctors deem necessary. Nothing, we all have our own issues to deal with. (P165)
<i>Sub-theme 2.2: Frustration</i>	I would not if they were open about their struggles. However, if they never shared then I might feel hurt. Openness is key. If they are open in sharing then I would understand the struggle and not read into their behavior too much because I would know they are having a bad day with chemical imbalance. (P533)
<i>Sub-theme 2.3: Embarrassment</i>	I would feel embarrassment more than anger. But I would stand up for them and explain to the others of Avery's condition so they would have some sympathy. I would also like to have a chat to Avery after they cool down and explain how their emotions might cause a damper on the mood of a social gathering. (P242)
<i>Sub-theme 2.4: Compassion Fatigue</i>	I will be neutral in that situation. I believe there is a limit on how much you can accommodate anger and pain from a loved one. (P434)
Theme 3: Behavioral Response	
Social Distance	Honestly, I do not think I would. I would be concerned that they would not take care of the place, or they would lose their job and not be able to pay. (P260)
Helping	No. I would feel like it would have a negative impact. Lack of assistance from her part. (P67)
Coercion Segregation	Yes. She will get worse. Her mental state might spill out into arguments with innocent people. (P136)
<i>Sub-theme 3.1: Differential Support</i>	Not frightened, but I would be aware of her sensitive position always, in order I could choose my words more wisely. When you are depressed you already feel so bad about things, that anything can make it feel even worse. That is the care you need to have. (P374)
<i>Sub-theme 3.2: Paternalism</i>	No, I feel that Avery is feeling withdrawn due to their depression. Avery needs someone to "take the reins" and listen to their needs. Avery needs friendship to support mental health and mental health referrals. (P239)
<i>Sub-theme 3.3: Condescension</i>	No, that seems too far. I am not a doctor and would not know if that is best for them. They do not seem to be physically violent or self-harming and they need to be taught how to deal with everyday life rather than being removed from it. (P130)

5.4.1 Reliability Validation. We integrated the new themes obtained through our human-LLM collaborative qualitative analysis with the attribution model to form a more nuanced and comprehensive coding scheme, which we then subjected to LLM-assisted qualitative validation. We

employed a similar approach to LLM deductive coding (Section 5.2.2) to evaluate the LLM’s performance on the new coding scheme using the same message set. Setting the integrated coding scheme as qualitative coding basis, we instructed the LLM to perform deductive coding. The results yielded a Cohen’s κ of 0.23 for the new coding scheme. Although not exceptionally high, it demonstrates an improvement compared to the LLM’s performance on the original attribution model, where the LLM was unable to assign codes matching the human coding, resulting in *zero* agreement. This suggests that our new findings enable humans and LLM to reach more agreement and reduce irreconcilable disagreements.

To further validate our newly discovered themes on new data, we conducted additional coding on a previously uncoded dataset. From the messages of the remaining 402 out of 1,002 participants, we randomly sampled 100 messages and employed both the attribution model [32] and the new integrated coding scheme for human coding. The coding manner was consistent with the previous coding, followed by an open discussion to align understanding between coders. We observed a reduction in irreconcilable disagreements after integrating our new conceptual insights into the coding scheme. The number of messages where the two human coders couldn’t reach an agreement decreased from 23 out of 100 (Cohen’s $\kappa = 0.67$) using the original attribution model to 9 out of 100 (Cohen’s $\kappa = 0.87$) when integrating our findings into attribution model. This finding demonstrates the capability of our insights in resolving disagreements, provides a more nuanced understanding of stigma attribution, and highlights the potential of our CHALET framework in advancing qualitative analysis.

6 DISCUSSION

Our proposed framework (🏡 CHALET) for advancing qualitative analysis is validated through a case study on conceptualizing social stigma based on the attribution model. Our results confirm the effectiveness of the 🏠 CHALET framework in conceptualizing mental-illness stigma, demonstrating the methodology’s feasibility. Additionally, our prompt design enhances LLMs’ grasp of psychological constructs and identifies factors influencing their interpretation. This study highlights the role of human-LLM collaboration in advancing qualitative analysis and deepening insights into stigmatized attitudes toward mental illness.

To answer Q1, we utilized a chatbot-assisted, LLM-enabled data collection method that encouraged participant self-disclosure. LLM-generated follow-up questions were particularly effective in eliciting comprehensive responses. These responses covered various content such as intentions and behaviors. Leveraging LLMs’ flexibility and language processing capabilities, we collected substantial data effectively, paving the way for a deeper understanding of the attitudes of participants.

After human deductive coding, we aimed to utilize LLMs to identify disagreements. Q2 explores the approach of integrating human-established deductive coding schemes into LLM-assisted coding to improve the model’s interpretation. During methodology validation, we found all information in the human-established codebook valuable for model performance. Additional rules and vignettes are particularly beneficial for challenging codes. Additional information beyond the human codebook (e.g., more examples, reasoning demonstrations).

Q3 investigates the discrepancies between human coding and LLM coding after identifying suboptimal results. It also explored the selection of suitable messages to advance qualitative analysis. We tackled this in our case study by pinpointing messages with different human and LLM codes and discerning the cause of disagreement: human error, LLM error, or potential new codes. We confirmed new codes through discussion and used these messages for deeper analysis.

Q4 explores the insights that can be gained by resolving human-LLM disagreements through inductive coding. This approach facilitates the conceptualization and generation of new insights using a selected set of messages. Inductive coding in our case study generated nuanced constructs

across each dimension. We identified themes such as microaggressions, improving our understanding of public attitudes towards mental illness. Preliminary validation shows that our new findings promote more human-LLM agreements and reduce irreconcilable disagreements compared to the original model, demonstrating its effectiveness and the validity of our methodology.

6.1 LLM-Assisted Qualitative Analysis

6.1.1 LLMs' Capabilities for Advancing Qualitative Analysis. Through our research, we discovered that LLMs possess the ability to help to analyze human psychological constructs and assist in the coding process, thereby contributing to the facilitation of qualitative analysis. Moreover, the introduction of the LLM trained in vast amounts of data helped mitigate human subjectivity bias, improved replicability through consistent coding of similar messages, and accelerates the analysis process compared to manual coding efforts.

Systematic Prompt Design for Superior Performance. Previous studies [3, 74, 150] explored employing simple prompts or the discussion of common techniques such as chain-of-thought prompting and few-shot/zero-shot settings. These studies showed LLMs' satisfactory performance in various text annotation tasks, including objective [3, 59] and subjective [84] tasks. Our case study resonates with these studies [156], as LLMs achieved relatively good results in identifying mental-illness stigma, surpassing the performance of human coders. The agreement between human coders and LLM coders (Cohen's $\kappa=0.75$) was higher than the agreement among human coders (Cohen's $\kappa=0.71$), underscoring the ability of LLMs to make nuanced judgments in subjective tasks. The deterministic nature of LLMs in annotating messages based on linguistic features and semantics enables them to locate messages that human coders may have misclassified or overlooked.

Furthermore, our approach to designing prompts that identify psychological constructs is distinctive. This study advances the coding of human's psychological construct by systematically investigating the impact of prompt textual structures and information content on LLMs' qualitative coding performance, while previous research has typically focused on only one or a subset of factors [95]. We found retaining all codebook information and placing instructions before content are beneficial, while more examples or reasoning prompts are less helpful. Human-established rules are equally effective for LLMs, and information redundant for humans does not substantially benefit them as well. By adopting these findings, LLMs achieve high consistency with human coders when coding psychological constructs, validating our prompt design approach. By identifying an optimal prompt, we ensure the robustness and validity of our results, laying the groundwork to the understanding of factors influencing LLMs' performance as coders and facilitating the development of more effective strategies for their utilization in qualitative analysis. It would be beneficial if future research developed more interpretable LLMs to increase trust in their outputs, addressing ethical considerations such as bias mitigation, and transparency in reporting.

Identifying LLMs' Limitation in Capturing Linguistic Nuances. This study explores LLMs' difficulties in qualitatively coding psychological constructs due to their struggles in interpreting nuanced linguistic features employed by humans, which can be derived from analyzing human-LLM disagreements, providing insights into the significance of analyzing such disagreements. Prior research [47, 102] have explored the language characteristics commonly used by humans when conveying psychological constructs. Interpreting these nuanced language differences requires capacities inherent to humans, such as rich interpersonal communication, societal experience, thinking patterns, social connections, and acute social awareness [40], which LLMs fall short, especially in handling tonal subtleties. Previous LLM studies [3, 84] have not adequately addressed these challenges. Our research suggests progressing LLM-assisted qualitative analysis by enhancing their ability to accurately interpret language intricacies. We anticipate future work incorporating

human-in-the-loop learning [120] and designing LLMs that consider social and cultural context [56] to better capture the subtle linguistic cues humans use.

Irreconcilable Human-LLM Disagreement as a Catalyst for Deepening Qualitative Analysis. Our work, aligning with previous studies [172], extended the use of human-LLM disagreements beyond refining code definitions to derive deeper conceptual insights and methodological advances in LLM-assisted qualitative research. Instead of just improving code wording, we systematically analyzed messages that cause irreconcilable disagreements to uncover new constructs and enrich concepts, thus advancing qualitative analysis. This approach treats disagreements as opportunities for deeper exploration, potentially fostering conceptual innovation and insight in qualitative research. By prioritizing the analysis of sentences where human and LLM interpretations diverge, researchers can focus on a smaller, yet potentially more conceptually valuable, subset of the data. This targeted approach enables a more efficient examination, directing attention to key information rather than analyzing substantial volumes of data that may be relatively homogeneous and less likely to yield significant theoretical contributions. We thereby promotes hybrid coding and encourages exploration of coding complexities, shifting focus from merely assessing model performance.

6.1.2 Collaborative Human-LLM Paradigm for Qualitative Analysis. While LLMs demonstrate their capabilities, we show that human involvement remains an indispensable component and that collaboration between humans and LLMs yields superior results for advancing qualitative analysis.

From Manual Hybrid Deductive/Inductive Coding to Human-LLM Collaborative Paradigm. Our work complements traditional methods of hybrid qualitative coding [51], which combine deductive and inductive approaches. Our human-LLM collaboration paradigm further offers additional benefits compared to the purely human-driven manner of hybrid coding, thereby advancing qualitative analysis. First, incorporating a consistent and rule-abiding LLM coder into the process of conceptualization can help identify ambiguous *edge cases*. During the initial stages of identifying new codes, we encountered ambiguous codes. For example, P422 mentioned “*give them pointers*” in their response, which, in our discussion with specialists, was identified as a form of *indirect help*. This type of help lies on a continuum between providing and not providing help, probably suggesting an unequal power dynamic within the interaction. We believed these codes could either be considered new or placed within existing codes, despite being on the periphery of the definition. In such cases, having a deterministic coder inform us that forcing a code into an existing category would lead to disagreement among coders allows us to revisit the ambiguous case and confidently assert that it represents a new code, as our initial uncertainty resulted in disagreement when we attempted to fit it into an existing code. It is essential for future studies to develop a more structured and deterministic process for identifying these vague edge cases. This would refine the human-LLM collaborative paradigm and enable more accurate and insightful qualitative analyses.

Second, our human-LLM collaborative methodology could help researchers become more familiar with the dataset before conducting qualitative coding. Alternating between performing deductive coding, identifying new codes, and conceptualizing them, without considering the entire dataset during coding, may result in an incomplete understanding of the data as a whole. Conceptualizing new codes without sufficient familiarity with the complete dataset can potentially result in inaccurate or misrepresentative codes. Consequently, a more methodologically sound alternative approach is to subsequently conceptualize new codes in a unified manner after gaining a comprehensive understanding of the entire dataset. To promote the human-LLM collaborative dynamics, future research could focus on developing a structured pre-coding exploration phase and investigating the optimal timing for introducing LLMs into the coding process, which provide guidance on integrating LLMs into qualitative analysis workflows.

Trade-off between Human Effort and LLM Capabilities. We conducted human deductive coding to deepen understanding, aiming to strike a balance between labor and comprehension. We stopped coding when inter-rater reliability was sufficiently high or half of the data was coded. Human interpretation of codes is enriched through codebook development, discussions, and oral instructions [44], significantly impacting the disagreement message set and subsequent collaborative inductive coding. Thus, a sufficient number of messages is required to ensure deep understanding. To conserve effort, we first conducted an exploratory analysis on the initial 4,200 messages to investigate the approach’s effectiveness. Subsequently, in the inductive coding phase, we compared human-LLM collaboration with LLM autonomous coding. The results in Appendix A.5.1 showed that LLMs alone could only identify literal codes and found it difficult to capture implicit information, confirming the essentiality of human effort and the importance of an organic human-LLM collaboration. Future studies may benefit from developing adaptive coding strategies that adjust the level of human involvement based on the specific characteristics of the data, such as those containing more implicit information or requiring deeper contextual understanding.

The importance of human involvement is also underscored by the choice of the disagreement set, which underscores the limitations of the LLM capabilities. Discrepancies are frequent when LLM-generated codes from a single prompt are compared with codes produced by humans. Nevertheless, the frequency of these discrepancies can be considerably reduced by analyzing the overlap of disagreements among human codes and various LLM codes from different prompts. This indicates that the suboptimal nature of even the best prompt significantly contributes to these discrepancies. Messages that are incorrectly coded by the LLM using the optimal prompt may get accurately coded using other, less optimal prompts. The diagram in Appendix A.4 display the significant reduction in the number of disagreements when using the intersection of disagreements, thus proving the effectiveness of this method in reducing the impact of LLM suboptimality.

6.2 Implications

6.2.1 Methodological Implications. This study proposes a novel methodology to advance qualitative analysis and facilitate conceptualization, offering implications for future researchers to explore additional HCI methods to empower qualitative research.

Firstly, field-initiated keystone datasets are crucial for realizing the full potential of LLMs in psychology and other fields. As highlighted by prior research [40], investing in the creation of such datasets is a top priority. Our methodology can inspire future research in the collection of keystone datasets. By employing LLM-enabled chatbots for data collection, we can potentially obtain rich, cross-cultural textual data, and the subsequent LLM-assisted coding process can help generate high-quality labels that meet the stringent standards for expert evaluation and impact assessment [40]. Through the extensive collection of diverse data and the implementation of human-LLM collaborative coding, we can ensure the creation of robust, well-coded datasets that adhere to the highest standards. By building upon these essential datasets, researchers can adopt our approach to engage in advanced qualitative analysis and conceptualization, paving the way for more robust and influential theoretical contributions.

Furthermore, our work not only stimulates deeper qualitative and quantitative analysis of attribution models but also has the potential for generalization to a wider range of psychological constructs. By proposing additional attributions of stigmatization through inductive coding, we inspire subsequent quantitative research to explore the relationships between these new attributions [91]. We also encourage future qualitative research to refine and extend our new findings by collecting additional empirical data and applying our proposed methodology for qualitative analysis. This may yield further conceptual insights and warrant exploration of the method’s generalizability to a broader range of domains, including other areas where attribution models are applied [66, 115].

By continually improving the attribution model for mental-illness stigma and other domains, we hope that our work will inspire researchers to adopt our methodology to other human psychological constructs, further validating the applicability of our approach.

6.2.2 Transdisciplinary Implications. Our study exemplifies interdisciplinary collaboration, using a transdisciplinary approach to perform qualitative analysis on attribution models, integrating social and computer science domains. This approach enables future human-AI interaction research to conduct more effective qualitative analysis and conceptualization, particularly through hybrid deductive-inductive coding, leading to novel insights and nuanced understanding. Achieving this requires close cross-disciplinary cooperation rather than researchers working independently in their respective fields. It enables social scientists to use LLMs for advanced qualitative research and conceptualization, fostering human-AI collaboration [12, 155]. Meanwhile, interdisciplinary collaboration provides insights into systematic prompt design and rapid, unrestricted data collection, empowering computer scientists to incorporate social science perspectives for more comprehensive and impactful human-AI interaction research, allowing for a bidirectional exchange of knowledge and facilitating development between the two disciplines.

6.3 Limitations

Our study has several limitations. Firstly, we only tested the performance of GPT-4 as our LLM coder, without evaluating the capabilities of other open-source models, such as Claude, which has been demonstrated to exhibit comparable performance to GPT-4. The exclusion of other models in the study raises uncertainty regarding the generalizability of our method to alternative LLMs. Furthermore, model parameters were not meticulously tested, and commonly used combinations were selected, potentially limiting the optimization of the approach. Future research should investigate the performance of various LLMs and the impact of different model parameters on the effectiveness of human-LLM collaborative coding.

Additionally, our dataset is constructed from human-chatbot interactions, while traditional qualitative analysis is typically based on human-human interactions. This discrepancy might potentially lead to misalignments in the application of our findings. Although several studies suggest that chatbots can elicit self-disclosure [99, 103], the extent to which self-disclosure in human-chatbot interactions compares to that in human-human interactions remains unclear and needs further investigation.

Moreover, the subjective nature of human coding renders the complete elimination of majority bias unfeasible. In qualitative coding involving multiple coders, the final coding decision is often based on the agreement of the majority, potentially leading to the loss of valuable insights from dissenting coders. This majority influence can introduce bias into the coding process, as the perspectives of the minority are overshadowed. Despite the implementation of necessary triangulation measures, the process of deductive and inductive coding in our study inevitably resulted in the potential loss of diverse perspectives due to majority bias.

Lastly, to facilitate LLMs' coding performance, we simplified the task by assigning only one attribution, i.e., corresponding code, to each message. Future work should consider the potential overlap between codes, thereby introducing greater complexity to the task and more closely reflecting real-world qualitative coding scenarios.

7 CONCLUSION

We introduce the  CHALET method, a novel approach designed to advance qualitative analysis and conceptualization. We validate the effectiveness of our method through a case study on the attribution model in the context of mental illness stigma. By applying the proposed method, we

collected a wealth of data and optimized the LLM prompt to enhance the LLM's ability to analyze human psychological constructs. We conducted an analysis of human-LLM disagreements and identified linguistic characteristics that hinder LLM's qualitative coding performance. Finally, by resolving these disagreements, we generated new understandings of social stigma across three dimensions: cognitive judgment, emotional response, and behavioral response. This qualitative analysis provides fresh insights to the conceptual illustration of mental-illness stigma. Our findings enhance LLM-assisted qualitative analysis, deriving deeper insights via human-LLM collaboration, with significant methodological and theoretical implications. This work underscores the potential of human-AI collaboration, paving the way for future research in qualitative methodologies utilizing this paradigm.

A APPENDIX

A.1 More Details about Chatbot-Assisted Data Collection

A.1.1 Sample Conversation Flow.

Table 7. Sample Conversation Flow

Small Talk

Conversational Agent (Nova): Hi, I'm Nova, nice to meet you! Could you please tell me your nickname? (Please just input your nickname!)

User: cynthia

Conversational Agent (Nova): Welcome Cynthia! I'm overjoyed to have the opportunity to work with you.

Conversational Agent (Nova): How was your mood today?

User: just ok

Conversational Agent (Nova): Hope things get better for you soon.

Conversational Agent (Nova): Btw, what is your favorite activity after finishing work or school?

User: Go out and exercise.

Conversational Agent (Nova): Exercise is a great way to unwind and stay healthy.

Conversational Agent (Nova): Have you watched any movies recently? How about recommending one to me?

User: No actually. But I like The First Slam Dunk very much, and I'd like to recommend this to you.

Conversational Agent (Nova): Thanks for the recommendation! I'll check out "The First Slam Dunk".

Conversational Agent (Nova): Watching movies can be really relaxing, have you ever attended a friend's party? Do you like it?

User: yes, i like it

Conversational Agent (Nova): That's great to hear! Friend parties can be a lot of fun. Thank you for sharing!

Conversational Agent (Nova): Friend party might be enjoyable, hmm, but what do you think about a friend's party with many of your friend's friends, people you're not familiar with? Do you feel anxious or happy?

User: i think it's ok for me, but i don't think i will talk to them haha

Conversational Agent (Nova): That's totally fine. Enjoying the atmosphere without engaging in conversation can be a great way to socialize!

Conversational Agent (Nova): As for me, it depends on the stranger's personality. I can maybe learn something new from them, but also I may feel stressed if the person is not easy-going.

Vignette Delivery

Conversational Agent (Nova): May I share a story about Avery's experience during a friend's party?

User: Ok

Conversational Agent (Nova): You will hear a short story right now.

Conversational Agent (Nova): Most of Avery's colleagues think that Avery is an easygoing and happy person, but this view has changed recently.

Conversational Agent (Nova): When Avery is at a company gathering, he starts to isolate himself and becomes irritable.

Conversational Agent (Nova): Avery also feels this change: he feels angry about his surroundings, and he gets frustrated about where the fury comes from.

Conversational Agent (Nova): When he is asked by one of his colleagues to share his thoughts on the news, Avery loses his temper and jumps out of his seat.

Conversational Agent (Nova): He screams, "I hate this conversation!" as he runs out of the place.

User: Oh..

Conversational Agent (Nova): Avery can not explain why he feels the strong urge to do that.

Conversational Agent (Nova): He unexpectedly loses his temper in these instances, causing a sense of not fitting in with his friends.

Conversational Agent (Nova): Cynthia, do you ever have the similar feeling with Avery - feel uncomfortable at a friend's party?

User: No

Conversational Agent (Nova): Understood. Thanks for sharing your experience.

Conversational Agent (Nova): Let me tell you more about Avery.

Conversational Agent (Nova): Avery finds that when he is depressed, he can barely concentrate on his work.

Conversational Agent (Nova): When Avery is working on his company projects, his work performance has declined, but he won't ask for help and feels like a burden.

Conversational Agent (Nova): He's considering quitting, feeling like a failure, and his manager asks him to improve.

Conversational Agent (Nova): He has noticed that the online courses he initially enrolled in out of interest no longer excite him, so he gives them up.

Conversational Agent (Nova): At home, Avery lacks motivation and neglects household chores, causing his family members frustration.

Conversational Agent (Nova): He feels guilty and empty. He feels that his family can't comprehend his suffering.

Conversational Agent (Nova): When interacting with strangers, Avery believes that others judge him harshly and he struggles to engage comfortably in conversations.

Conversational Agent (Nova): Every time he checks Instagram and Facebook, Avery sees his friends are having fun, and he feels alone because he does not understand why he cannot be like that.

Conversational Agent (Nova): Hmm, the story is over. After reading, do you have any suggestions toward Avery?

User: I think he should go to hospital and have a checkup.

Conversational Agent (Nova): That's a reasonable suggestion, seeking professional help can provide clarity and support for Avery.

Question Answering

Q1 - Responsibility

Conversational Agent (Nova): Cool! Let's talk more about the story. Just like what is mentioned in the story, Avery is currently facing difficulties in both his relationships with colleagues and his work performance. Do you believe Avery's current situation is primarily a result of his actions?

User: no, he is just get ill and have a bad status, he can be better

Conversational Agent (Nova): So you feel that Avery's challenges are more related to his health than his behavior. Is his mental condition an indication of personal weakness? Please share your thoughts.

User: yes, it's a weakness because he is not performing well at home and at work

Conversational Agent (Nova): I see. It can be tough when mental issues create obstacles in life for sure.

Conversational Agent (Nova): For me, I think Avery's current situation is a complex interplay of his actions and his mental health challenges. His experience with mental illness does indeed have an impact on his behavior to a certain extent, but his behavior has certainly contributed to strained relationships and low work performance.

Q2 - Social Distance

Conversational Agent (Nova): What's more, may I kindly ask, if you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?

User: I probably not

Conversational Agent (Nova): Understood, your comfort is important.

Conversational Agent (Nova): What potential negative consequences do you think could arise if you were to rent your house to someone like Avery?

User: maybe he cannot manage himself, and ask for help frequently, I don't want trouble or any inconvenience so sorry

Conversational Agent (Nova): I understand your concern about Avery's ability to manage on his own. What experiences or traits does Avery have that lead you to believe he may frequently need assistance?

User: because when he at home, he barely do anything, that means something

Conversational Agent (Nova): Aye, it's tough when Avery just doesn't seem to do much at home.

Conversational Agent (Nova): I might be open to the idea of renting the room to Avery, but I do hesitate because I'm worried about how his mental condition might affect aspects like cleanliness or the overall rental situation.

Q3 - Anger

Conversational Agent (Nova): Thank you for sharing! Let's imagine you were one of Avery's friends at the party or other social occasions. Would you feel anger toward him due to his actions, for example, losing his temper and yelling at someone?

User: yes i will, he may ruin the gathering, let everyone down

Conversational Agent (Nova): I understand how Avery's actions could upset you. What specific incident caused you to feel this way about Avery's behavior?

User: like if we want to start a conversation with him, he will shout at us, that's terrible, but this's not his fault maybe

Conversational Agent (Nova): Wow, that does sound awful. No one should be shouted at for starting a conversation.

Conversational Agent (Nova): If I were one of Avery's friends at the party or other social occasions, I might feel frustrated or annoyed because his actions disrupt social events. However, I would also understand that his behavior is likely due to his mental health challenges, which would temper my anger.

Q4 - Helping

Conversational Agent (Nova): Thanks for discussing!! I wonder if you were one of Avery's colleagues or friends, then would you be inclined to help him with tasks, such as work projects?

User: no..

Conversational Agent (Nova): Fair enough, I get where you're coming from.

Conversational Agent (Nova): If Avery were to work on a project with you, what potential result do you think it could have on your study?

User: he will affect my work performance, i need to be responsible for myself first

Conversational Agent (Nova): I understand that you're concerned about your work performance with Avery's involvement. What particular behaviors of Avery make you anticipate a negative impact on your productivity?

User: for example, he have a low work performance himself, and manager has to talk to him, it's not a good info for me

Conversational Agent (Nova): Ah, Avery's poor performance and the manager's involvement might be disruptive. I got it.

Conversational Agent (Nova): For me, to be honest, I would consider giving him an opportunity to collaborate. I might also hesitate to cooperate with Avery because his condition could potentially impact his performance or reliability in the workplace. Hmm, hard to choose.

Mid-Break

Conversational Agent (Nova): Okay, let's take a brief break from the questions! I recall that you love exercise, would you like to explain why this is charming to you?

User: it helps me build muscles

Conversational Agent (Nova): Absolutely, exercise does great in muscle building, helping you maintain a fit body.

Conversational Agent (Nova): Great! Thank you so much, Cynthia! I hope you do not feel tired now, let's go back to our interesting discussion! Let me quickly recap of Avery's story:

Conversational Agent (Nova): Avery's work performance, lifelong learning, and relationships started deteriorating, then he experienced a desire to quit his job and the loss of interest in things he once enjoyed. He isolates himself from friends, experiences anger and confusion, and cannot connect with others as usual. Self-judgment feelings emerge, causing emotional distress.

Q5 - Pity

Conversational Agent (Nova): I'm really happy you are with me now. Let's continue our talk. I'd like to know if you feel concern or pity at Avery's mental state?

User: maybe

Conversational Agent (Nova): Understandable. It's a tough situation.

Conversational Agent (Nova): Let's go further. Why was it?

User: i think he is experiencing a tough time

Conversational Agent (Nova): I see, you perceive Avery is going through a hard phase. What specific events or parts of Avery's story made you feel this way?

User: because he feel confused at his condition, i think he must be very pain

Conversational Agent (Nova): Absolutely, coming to terms with his situation must be quite tough for Avery.

Conversational Agent (Nova): I'm more than glad to share my thoughts! In this scenario, I could have conflicting emotions regarding Avery's situation. On one hand, I could feel concerned for him because his condition appears to be causing significant distress and affecting his daily life. On the other hand, I might also consider his condition as vulnerabilities that he should overcome, which could lead to a lack of concern.

Q6 - Coercion Segregation

Conversational Agent (Nova): Apart from being a colleague or friend, I'm also curious if you are Avery's neighbor, then would you consider it better to separate him from the neighborhood and have him undergo hospitalization for treatment?

User: maybe yes? i'm not sure whether he will shout at home so I cannot have a good community with him

Conversational Agent (Nova): I sense your uncertainty. If Avery wasn't separated or hospitalized, what could be the potential result to your home life?

User: just as i said, he may shout, yell, and have bad feeling toward community

Conversational Agent (Nova): I totally understand, you fear Avery's behavior may disrupt your community ethos.

Conversational Agent (Nova): Well, I would probably think that placing Avery in a hospital can be an option. This could somewhat reduce the potential risk of having problems with the neighbors, and perhaps, without Avery, our neighborhood relations might become more harmonious. However, on the other hand, it's important to consider his preferences and follow medical advice. I'll also pay attention to Avery's personal well-being!

Q7 - Fear

Conversational Agent (Nova): Thank you so much for being with me for such a long time! Let's now discuss our final topic: if you were traveling or engaging in activities with Avery, would you have any feelings of frightened or being threatened regarding the potential implications of your interaction?

User: yes, he seems uncontrollable and uncertain, i don't know what he will do, i'm really afraid he shout at me

Conversational Agent (Nova): I understand your fears about Avery's unpredictability. What specific actions of Avery made you feel this way?

User: i think just like at a gathering, no one expect him to leave abruptly

Conversational Agent (Nova): Ah, I get it. Avery's sudden departures at gatherings are indeed unpredictable.

Conversational Agent (Nova): Hmmmm for me, if I were traveling or engaging in activities with Avery, I would likely have mixed feelings. On one hand, I would empathize with his struggles and try to offer support. On the other hand, I might also feel frightened about the challenges and uncertainties that could arise due to his unhappy and angry conditions.

Ending

Conversational Agent (Nova): That's all for my questions. Thank you for doing this with me!

Conversational Agent (Nova): Now, do you have any questions? Feel free to let me know!

User: no thanks

Conversational Agent (Nova): That's great! Thank you for your time and valuable responses. I'm glad to assist you. Please feel free to reach out anytime!

Conversational Agent (Nova): Congratulations, Cynthia, we have finished our fantastic discussion! I sincerely wish you all the best.

A.1.2 Vignette Design. Our vignette is demonstrated in Figure 6.

A.1.3 Question Adaption. First, some attributions have clear causal relationships or substantially similar meanings, such as fear and dangerousness (dangerousness leads to fear), as shown in Figure 2. Although they represent different attributions, the question stems in the original Attribution Questionnaire are considerably similar, which may give participants a repetitive feeling during the conversation session. Asking similar questions in a conversational scenario is inappropriate, as reflected in the feedback from our pilot study. We therefore reduced the number of nearly redundant questions, and we decided to temporarily exclude the discussion of *dangerousness*.

Second, in accordance with previous research [98], we observed that a significant number of prior studies have substituted the term *social distance* for *avoidance* in the original Attribution Questionnaire [14, 31, 106], while retaining the same definition. Consequently, we have also adopted the term *social distance* to represent the stigma associated with the desire to stay away from individuals with mental illness, thereby aligning our terminology with the established literature.

Last and most importantly, as the original Attribution Questionnaire is self-reported and the question stems are greatly straightforward without any scenario, lacking vividness, it may lead to the presence of social desirability bias [55, 159]. To address this, we further fine-tuned the question settings by randomizing the order of the questions presented to participants and incorporating a vivid scenario for each question. This approach allows participants to respond more easily and provides them with a specific as well as concrete basis upon which to express their opinions. By turning the questions into more *indirect* ones, based on previous research [53], indirect questions can reduce *social desirability bias*, assisting us obtain more genuine and honest responses from participants.

A.1.4 Conversational Agent User Interface. The user interface is shown in Figure 7.

A.1.5 Prompts in Conversational Agent. For instance, we prompt the LLM to generate an open-ended follow-up question for coercion segregation leveraging:

"Nova: Apart from being a colleague or friend, I'm also curious if you are Avery's neighbor, then would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?"

User: answer_text."

First, restate what the user has said. Then, your task is to ask a question based on their response.

If the user said they do not want to put Avery in the hospital, kindly ask them which part of Avery's story led them to think so.

If the user said they agree that it's better to put Avery in the hospital or have no idea, ask what the potential result could be if Avery is not separated from the neighborhood and does not undergo hospitalization. DON'T ASK WHY, focus on asking about the "potential result" (include these two words in your question).

Keep your response between 10 and 40 words, easy to read. Don't start with "Nova:."

Your question MUST start with 'how' or 'what'. DON'T use phrases like 'can you...' or 'could you...'.

A.1.6 Informativeness Computation. We used informativeness to quantify the richness of the data collected through our chatbot. Informativeness, a concept from information theory, measures

Vignette - Detailed:

Most of Avery's colleagues think that Avery is an easy-going and happy person, but this view has changed recently. When Avery is at a company gathering, he starts to isolate himself and becomes irritable. Avery also feels this change: he feels angry about his surroundings, and he gets frustrated about where the fury comes from. When he is asked by one of his colleagues to share his thoughts on the news, Avery loses his temper and jumps out of his seat. He screams, "I hate this conversation!" as he runs out of the place. Avery cannot explain why he feels the strong urge to do that. He unexpectedly loses his temper in these instances, causing a sense of not fitting in with his colleagues.

Let me tell you more about Avery. Avery finds that when he is depressed, he can barely concentrate on his work. When Avery is working on his company projects, his work performance has declined, but he won't ask for help and feels like a burden. He's considering quitting, feeling like a failure, and his manager asks him to improve. He has noticed that the online courses he initially enrolled in out of interest no longer excite him, so he gives them up. At home, Avery lacks motivation and neglects household chores, causing his family members frustration. He feels guilty and empty. He feels that his family can't comprehend his suffering. When interacting with strangers, Avery believes that others judge him harshly and he struggles to engage comfortably in conversations. Every time he checks Instagram and Facebook, Avery sees his friends are having fun, and he feels alone because he does not understand why he cannot be like that.

Vignette - Recap:

Avery's work performance, lifelong learning, and relationships started deteriorating, then he experienced a desire to quit his job and the loss of interest in things he once enjoyed. He isolates himself from friends, experiences anger and confusion, and cannot connect with others as usual. Self-judgment feelings emerge, causing emotional distress.

Fig. 6. Vignette design. There are two versions: a **detailed** version and a **recap** version. The detailed version is presented to participants before we deliver the questions and provides a comprehensive illustration of the vignette. In contrast, the recap version is a concise summary of the vignette, which is provided to participants after they have answered half of the questions. The purpose of the recap version is to remind participants of the essential elements within the vignette. Both draft vignettes underwent a review process and were subsequently approved by a psychiatrist who is a member of our research team.

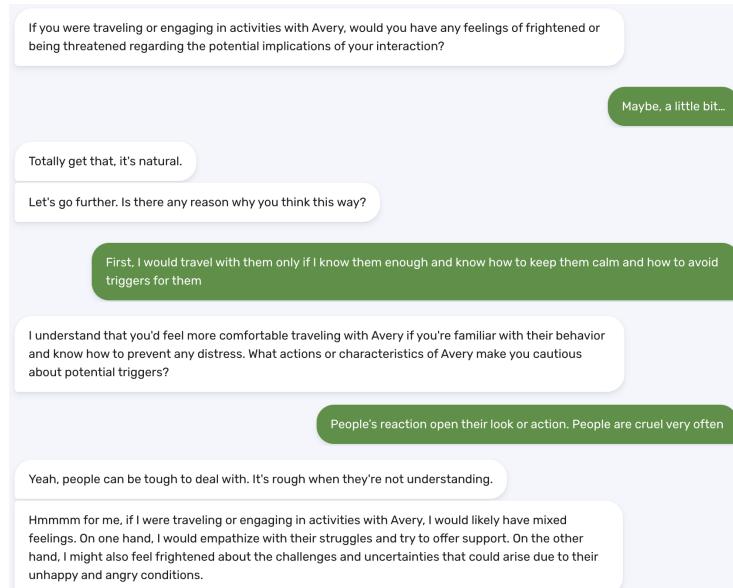


Fig. 7. Conversational agent interface. This figure demonstrates the conversational agent and participants engaging in the question answering task, using the attribution [32] of *fear* as an example. The figure presents a segment of the conversation history with Participant 981, showcasing the application of designed conversation techniques.

the surprise or uniqueness of the information contained in a message [169]. Specifically, the informativeness of a response is calculated using the following formula:

$$I(\text{Response}) = \sum_{n=1}^N \log_2 \frac{1}{F(\text{word}_n)} \quad (1)$$

where $I(\text{Response})$ is the informativeness of the response, and $F(\text{word}_n)$ is the frequency of occurrence of the word n in a large, representative corpus of modern English (the Brown Corpus, Webtext, and NPS Chat Corpus). This calculation essentially quantifies the unexpectedness or rarity of each word in a response, with the overall informativeness being the sum of these values. Words less common in everyday language use contribute more to the informativeness score, reflecting the depth and specificity of the information provided.

Our obtained high informativeness value indicates that, on average, the responses contained a substantial amount of unique information, suggesting that the participants provided detailed and varied perspectives. The relatively wide standard deviation further underscores the diversity in the depth of information across different responses. Such variability indicates the effectiveness of our chatbot in guiding participants to provide rich, nuanced insights, ranging from highly informative and detailed accounts to more straightforward responses.

A.1.7 Word Frequency Analysis on Collected Data. To analyze the quality of the collected data, we first clustered the text according to attribution, followed by preprocessing the text, which included converting to lowercase, removing punctuation, numbers, and stop words. Additionally, based on observations, we also excluded words that were three characters or shorter to eliminate the interference of auxiliary words, prepositions, and conjunctions. Ultimately, we generated word clouds shown in Figure 4 and Figure 8 for each attribution containing the top 200 high-frequency words from the cleaned text.

We found that participants' responses closely revolved around the questions we posed. For example, in the word cloud for the *Responsibility* attribution, the frequent appearance of words such as "action", "weakness", "personal" and "cause" indicated that participants were discussing whether Avery's situation was the result of their actions or reflected personal weaknesses. Similarly, the high-frequency words "help", "support", and "work" in the *Helping* attribution word cloud directly corresponded to the question of whether participants were willing to help Avery complete a work project. Additionally, words like "hospitalization", "treatment", and "neighborhood" from the *Coercion Segregation* word cloud, and "tenant", "rent" from the *Social Distance* word cloud showed that participants were seriously answering these two more controversial questions.

While ensuring that participants' responses were closely related to the topic, we also discovered through the word clouds that participants engaged in deep thinking around the scenarios we presented. This further proved the effectiveness of our LLM-enabled chatbot in facilitating rich, nuanced conversations about the stigma of mental illness. For instance, in the *Anger* and *Pity* attributions, characteristic words such as "concern", "understand", "sympathy", and "depression" indicated that participants were reflecting on the complex emotions and reactions triggered by the scenarios. This deep reflection was also evident in the *Fear* attribution word cloud, where words like "threatened", "frightened", and "outbursts" reflected participants' predictions and perceptions of Avery's potential behavior. Similarly, high-frequency words such as "work", "property" and "comfortable" in the *Social Distance* word cloud showed that participants also engaged in discussions on topics related to the issue.

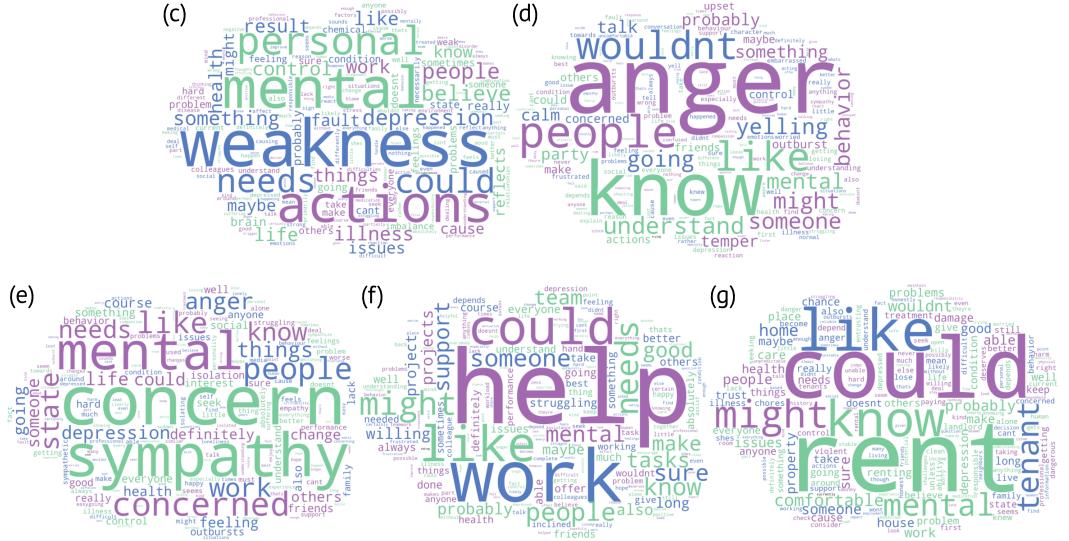


Fig. 8. Word frequency analysis across all attributions: (c) responsibility, (d) anger, (e) pity, (f) helping, and (g) social distance.

A.2 More Details about Human Deductive Coding

A.2.1 Additional Rules in Codebook. During the process mentioned in Section 4.3.1, we developed several additional rules that everyone could follow. First, if the message mentions both stigmatized and non-stigmatized thoughts, code it as *Stigmatized*. Similarly, if the message only conveys a stigmatized conclusion but with non-stigmatized reasoning, code it as *Stigmatized*. Furthermore, we discovered some messages that were difficult to categorize into our existing codes. In such cases, we introduced the N/A code. The text is considered N/A if it is completely off-topic, on-topic but does not take a stance and provides completely objective reasons or facts, on-topic and has a stance but the reason is irrelevant to mental health, or if the data quality is low. An example of a message categorized as N/A is “I would never share my house with a roommate as I live with my family, so I feel I cannot answer that question.” This statement does not express any views regarding mental illness. We eliminated 55 out of 4,200 codes of N/A, resulting in the final dataset comprising 4,145 messages.

The deductive coding scheme we derived from pre-existing conceptual frameworks is presented below, corresponding to Table 1.

Responsibility. A common stigmatized thought is that people have *control* over and are responsible for their mental illness and related symptoms. Moreover, additional rules are as follows. If the message only conveys a suggestion, code it as *Non-stigmatized*, whereas if the message not only conveys a suggestion but also highlights a potential negative outcome if the suggestion is not followed (e.g., ‘otherwise’) or expresses *disappointment*, code it as *Stigmatized*.

Anger. The stigmatized thought is to be irritated or annoyed because people are blamed for their mental illness. Furthermore, during the coding process, an agreement was reached that an enriched definition of anger includes the belief that certain actions are unacceptable or the experience of discomfort. Additionally, if the message conveys feelings of dislike or other forms of *aversion*, it should be coded as *Stigmatized*. If the message merely expresses feelings of confusion, bemusement, or bewilderment, it should be coded as *Non-stigmatized*. If the message solely refers to expressions related to ‘out of character,’ it should be coded as *Non-stigmatized*.

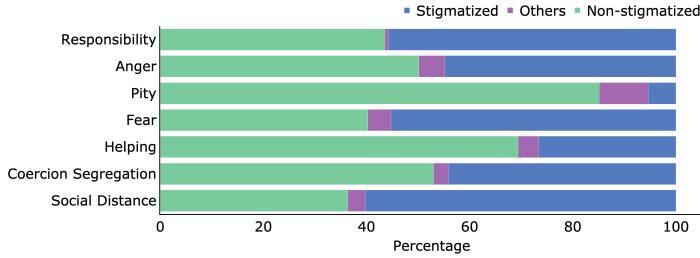


Fig. 9. The distribution of raw human codes across various attributions. For each attribution, the human deductive coding process categorized each message into three codes: *Stigmatized*, *Non-stigmatized*, and *Stigmatized (Others)*.

Pity. A common stigmatized thought is to be *unsympathetic* toward individuals with mental illness, and place blame on them for their mental health challenges.

Fear. The stigmatized thought is to believe individuals with mental illness are not *safe* or feel *frightened* since individuals with mental illness are dangerous or *unpredictable*. Additionally, if the message conveys feelings of caution, it should be coded as *Stigmatized*.

Helping. The stigmatized behavior is withholding support toward them because of biased thoughts. If the message suggests that helping individuals with mental illness is conditional, such as being contingent upon their willingness to seek help, it should be coded as *Stigmatized*. On the other hand, if the message indicates that being a friend of someone with mental illness or understanding their condition is a reason to offer help, it should be coded as *Non-stigmatized*.

Coercion Segregation. The stigmatized behavior is to send individuals with mental illness to institutions away from their community and force individuals with mental illness to participate in medication management or other treatments. In addition, if the message simply suggests going to treatment or counseling, it would be coded it as *Non-stigmatized*.

Social Distance. The constant stigmatized behavior is to stay away from individuals with mental illness.

A.2.2 Human Deductive Coding Statistics. During our human deductive coding process, we utilized three codes: *Stigmatized*, *Non-stigmatized*, and *Stigmatized (Others)*. The distribution information of the human deductive coding is presented in Figure 9.

A.3 More Details about LLM Deductive Coding

A.3.1 Prompt Format. Our prompt structure is shown in Figure 10.

A.3.2 Fine-Tuning Code Wording. Inspired by the findings that prompt wording can have a significant impact on the results due to subtle variations [44], we followed the advice of a specialist in our research team and modified the code definition for *pity* by replacing *sympathy* with *empathy*. The results showed that when providing only the code name, the agreement between human and LLM (Cohen's κ) increased from 0.12 with *sympathy* to 0.14 with *empathy*. This finding confirms that subtle differences in prompt wording can indeed influence the results. Consequently, we consistently used *empathy* in our prompts throughout the study.

A.3.3 Exploring the Impact of Multiple Output. Reiss [134] mentions that LLMs are non-deterministic, meaning that identical input can lead to different outputs. To mitigate the result bias caused by non-determinism, they employed a multiple output approach. We conducted a similar experiment,

[Role Assignment]	Task	Role
You will be given a question and the corresponding participant's opinion. Your role is a competent annotator for social stigma towards mental illness. The question and opinion is based on a vignette's plot. Answer the following question based on the constraints given to you:		
[Multiple-Choice Question]		
Which of the following describes “[Participant’s Opinion]”?		
A. Stigmatized (responsibility/blame). ... H. Non-stigmatized.		
[Code Definition & Keywords]		
1. Responsibility: A common stigmatized thought is that people have control over and are responsible for their mental illness and related symptoms.		Code Definition
Keywords: responsible, responsibility, weakness, stem from, blame, etc.		Keywords
... 8. Non-stigmatized: The text explicitly conveys non-stigmatized content, displaying positive attitudes such as sympathy, assistance, friendliness, etc.		
[Examples]		
[participant’s opinion] It could be about how Avery is viewing things. Instead of just focusing on themselves, they could think about others’ feelings. Not weakness, just needs a different point of view.		Complete Example of One Code
label: A		
reason: The participant's opinion suggests that people with mental illness need to change the way they view things and should think about others' feelings. This thought implies that ...		
... [participant’s opinion] Yes, and a friend needs to be there for him. To know that friends can help, just being there. label: H		
[Additional Rules]		
1. If a message mentions both stigmatized and non-stigmatized statements, annotate it as Stigmatized.		Additional Rules on Stigmatised and Non-stigmatised
2. When discussing ‘responsibility/blame,’ common stigmatized thoughts are:		
i. People have control over and are responsible for their mental illness and related symptoms.		
... v. Other statements suggest responsibility.		Additional Rules on Specific Attribution
If a message contains the stigmatized thought mentioned in i. - v., choose A.		
... 9. If it doesn't have any stigmatized thought, choose H.		
[Output Format]		
Please output your choice in the format of: label: [A/B/C/D/E/F/G/H]		
reason: explain your reason for coding.		
[Question]		
Do you believe Avery’s current situation is primarily a result of their actions? Is their mental condition an indication of personal weakness? Please share your thoughts.		
[Participant’s Opinion]		
I believe the outcome is not a result of their actions, but rather a manifestation of their depression. The depression they experience may lead to others not being fond of them.		

Fig. 10. Prompt template when we ask LLM to do deductive coding based on the attribution model. This figure is a case where we leveraged all messages in the human-written codebook. We also control the addition or reduction of fields such as additional rules to explore the impact of different amounts of information in the prompt on the agreement of deductive coding.

comparing the results of single output and majority voting from three outputs, with the results shown in Figure 11(a).

Our conclusion demonstrates that enabling multiple outputs and majority voting yields slightly worse results compared to single output, which aligns with our expectations. Multiple outputs also

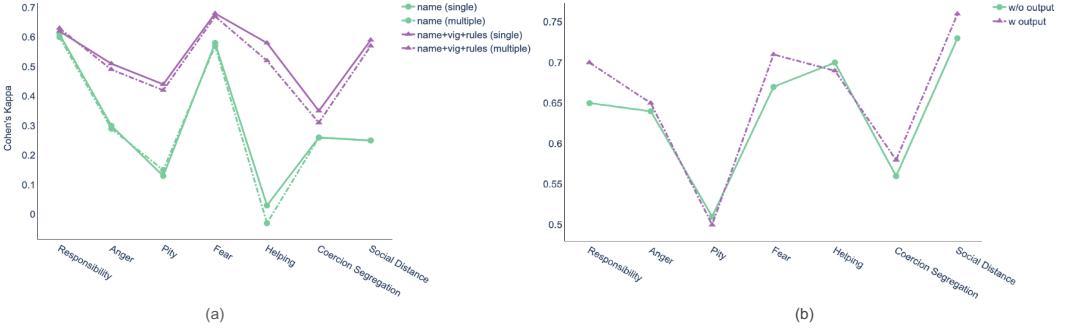


Fig. 11. Cohen's κ . (a) Results comparison between the majority voting of multiple outputs and single output across different attributions under two conditions: providing only the code name in the prompt, and providing code name, vignette and additional rules in the prompt. (b) Results comparison between when LLMs are enabled to provide coding rationales and when they are not permitted to do so, under the condition that the complete human-generated codebook is supplied as the informational resource.

imply the inclusion of content with lower probabilities, potentially affecting the LLM's performance. Conversely, single output represents the result with the highest probability, rendering it more rational.

A.3.4 Enabling LLMs to Provide Reasoning. We additionally compared two conditions: one where we allowed the LLM to output its reasons for coding and another where we did not require the LLM to provide its reasoning. The results, as shown in Figure 11(b), indicate that overall, allowing the model to output the reasons for its codes can enhance its performance.

A.3.5 Effect of Code Permutation on Predicted Codes. We found that changing the order of examples in the prompt can alter the distribution of predicted code, as illustrated in Figure 12.

A.4 More Details about Human-LLM Disagreement

We quantified the number of human-LLM disagreements across different prompts for each attribution, as illustrated in Figure 13.

A.5 More Details about Collaborative Inductive Coding

A.5.1 LLM Inductive Coding Results. Drawing from the prompts and structure employed in a previous study [39], we conducted an experiment to evaluate LLMs' autonomous inductive coding capabilities. First, we prompted the gpt-4-1106-preview to generate initial codes, which were then grouped into themes by the model itself. The resulting themes and their associated codes are presented in Table 8. Evidently, the LLM's performance fell short of the results achieved when human researchers are actively involved in guiding the inductive coding process. Notably, we observed instances of overlapping codes being erroneously assigned to different themes.

A.5.2 Affinity Diagram for Human Inductive Coding. During the human-LLM collaborative inductive coding process, we documented the notes produced by the human researchers, as well as select ideas generated by the LLM. These are presented in Figures 14 and 15.

A.5.3 Codes Definition. This section outlines the comprehensive definitions for each newly discovered theme.

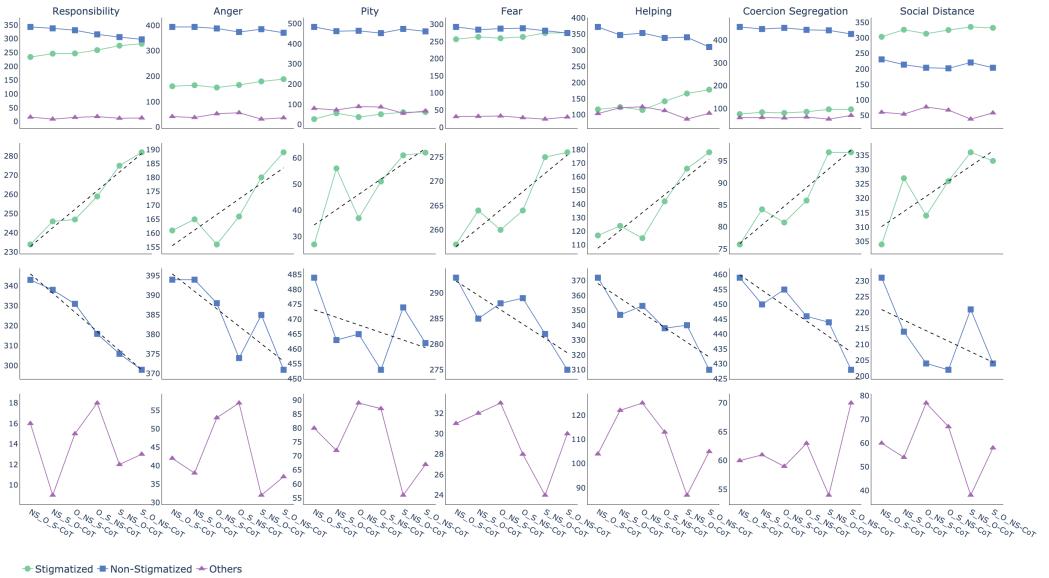


Fig. 12. Under the conditions with CoT, the number of predicted codes. The dashed line represents the linear fit obtained throughout linear regression analysis. The fourth row, the code of *Stigmatized (Others)*, does not exhibit a specific pattern that can be fitted due to the relatively low number of predictions.

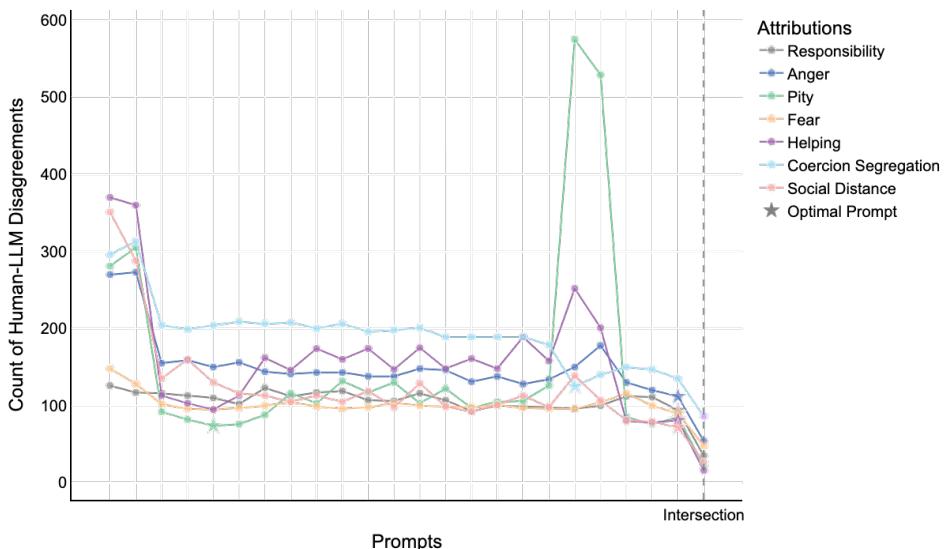


Fig. 13. The count of human-LLM disagreements varies across prompt conditions and attributions. The rightmost column shows the count after taking the intersection of all disagreement messages. While the optimal prompt yields the fewest disagreements among single prompts, the count remains substantially higher than the intersection-based approach, highlighting the suboptimality of the optimal prompt still exists.

Table 8. Results of autonomous inductive coding by LLMs on new attributions for mental illness stigma.

Theme	Descriptions & Codes
Theme 1: Health Management	Topics related to managing health, whether through professional intervention, self-care, or community support. Code 1.1: Hospitalization Debate Code 1.2: Home Treatment Preference Code 1.3: Professional Assessment Advocacy Code 1.4: Anger Management
Theme 2: Support Systems	Topics focusing on the various forms of support available or needed for individuals, including workplace, community, and emotional support. Code 2.1: Community Support Importance Code 2.2: Workplace Support Code 2.3: Sympathy and Concern
Theme 3: Mental Health Focus	Topics that specifically relate to mental health issues, including management, misconceptions, and the importance of professional assessments. Code 3.1: Mental State Misinterpretation Code 3.2: Anger Management Code 3.3: Professional Assessment Advocacy
Theme 4: Social Acceptance and Integration	Topics that deal with the acceptance and integration of individuals into various social settings, including housing and the workplace. Code 4.1: Tenant Acceptance Code 4.2: Workplace Support
Theme 5: Emotional and Psychological Well-being	Topics that touch on the emotional and psychological aspects of well-being, including the management of emotions and the importance of empathy. Code 5.1: Sympathy and Concern Code 5.2: Anger Management Code 5.3: Mental State Misinterpretation
Theme 6: Advocacy and Debate	Topics that involve discussions, debates, or advocacy for certain approaches or understandings related to health and social integration. Code 6.1: Hospitalization Debate Code 6.2: Professional Assessment Advocacy
Theme 7: Safety and Concerns	Topics that are centered around concerns for safety, whether during travel or in the context of mental health. Code 7.1: Traveling Concerns Code 7.2: Mental State Misinterpretation
Theme 8: Preference and Choice in Care	Topics that highlight the preference for where and how care and treatment should be received. Code 8.1: Hospitalization Debate Code 8.2: Home Treatment Preference
Theme 9: Community and Social Support	Topics emphasizing the importance of community and social support in various aspects of life and well-being. Code 9.1: Community Support Importance Code 9.2: Workplace Support
Theme 10: Conflict and Management	Topics related to managing conflicts, whether internal (such as anger) or external (such as tenant acceptance). Code 10.1: Anger Management Code 10.2: Tenant Acceptance
Theme 11: Mobility and Accessibility	Topics that deal with concerns related to mobility, including traveling concerns and the broader implications for individuals needing support. Code 11.1: Traveling Concerns

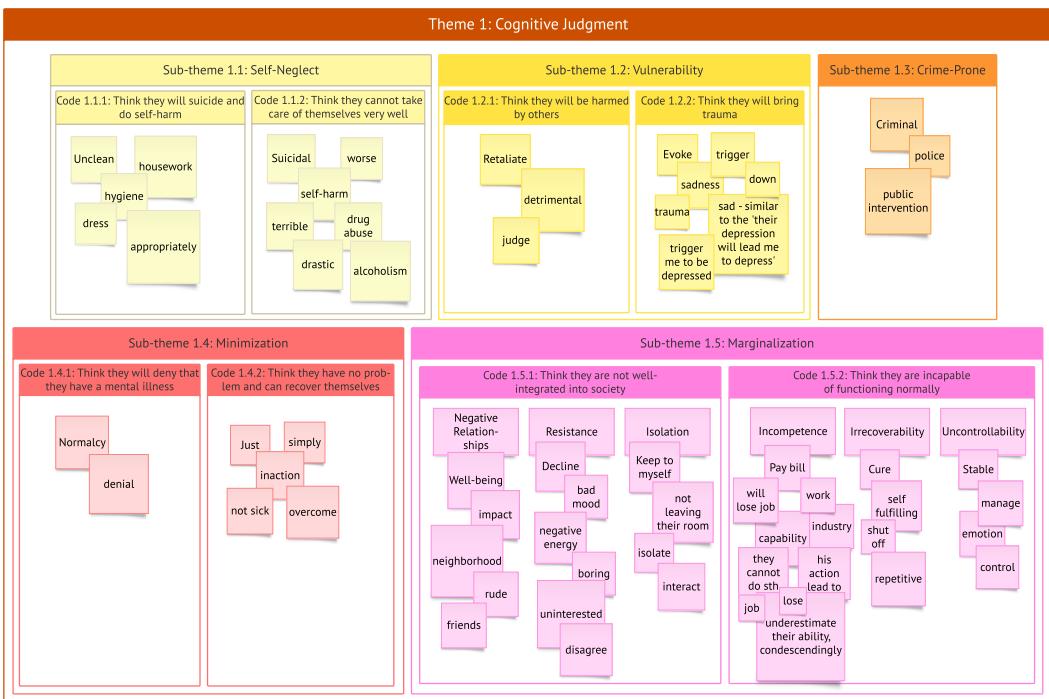


Fig. 14. Affinity diagram of theme 1.

Theme 1: Cognitive Judgments. Cognitive judgment [114] refers to perceptions toward individuals with mental illness, including mental processes and assessments involved in making decisions, forming opinions, and reaching conclusions based on available information, reasoning, and past experiences.

Sub-theme 1.1: Self-neglect. Think individuals with mental illness will neglect themselves and are unwilling or unable to do necessary self-care.

Code 1.1.1: Think individuals with mental illness will suicide and do self-harm. Many participants believe that individuals with mental illness are prone to engaging in suicidal behavior or self-harm, which is a common misconception among participants. In the given scenario, we intentionally omitted the portrayal of extreme behaviors such as self-harm. The tendency to equate or associate mental illness with self-destructive behaviors reflects a deep-rooted, stigmatized societal perception of mental illness, an excessive association that overlooks the complexity and diversity of mental illness and oversimplifies the symptomatology of patients. Previous research [93] shows that not all individuals with mental illness exhibit suicidal or self-injurious tendencies, and equating the two not only constitutes an inaccurate generalization but also reinforces the public's negative prejudices and fears about mental illness.

Code 1.1.2: Think individuals with mental illness cannot take care of themselves very well. Mental illness is often associated with self-neglect [38, 67]. This refers to concerns that individuals with mental illness will be unable, or unwilling, to care for their own essential needs. It can cover a wide range of behavior including neglecting personal hygiene, health or surroundings. This association deepens society's negative perceptions and prejudices toward those with mental illness, potentially exacerbating their feelings of inferiority.

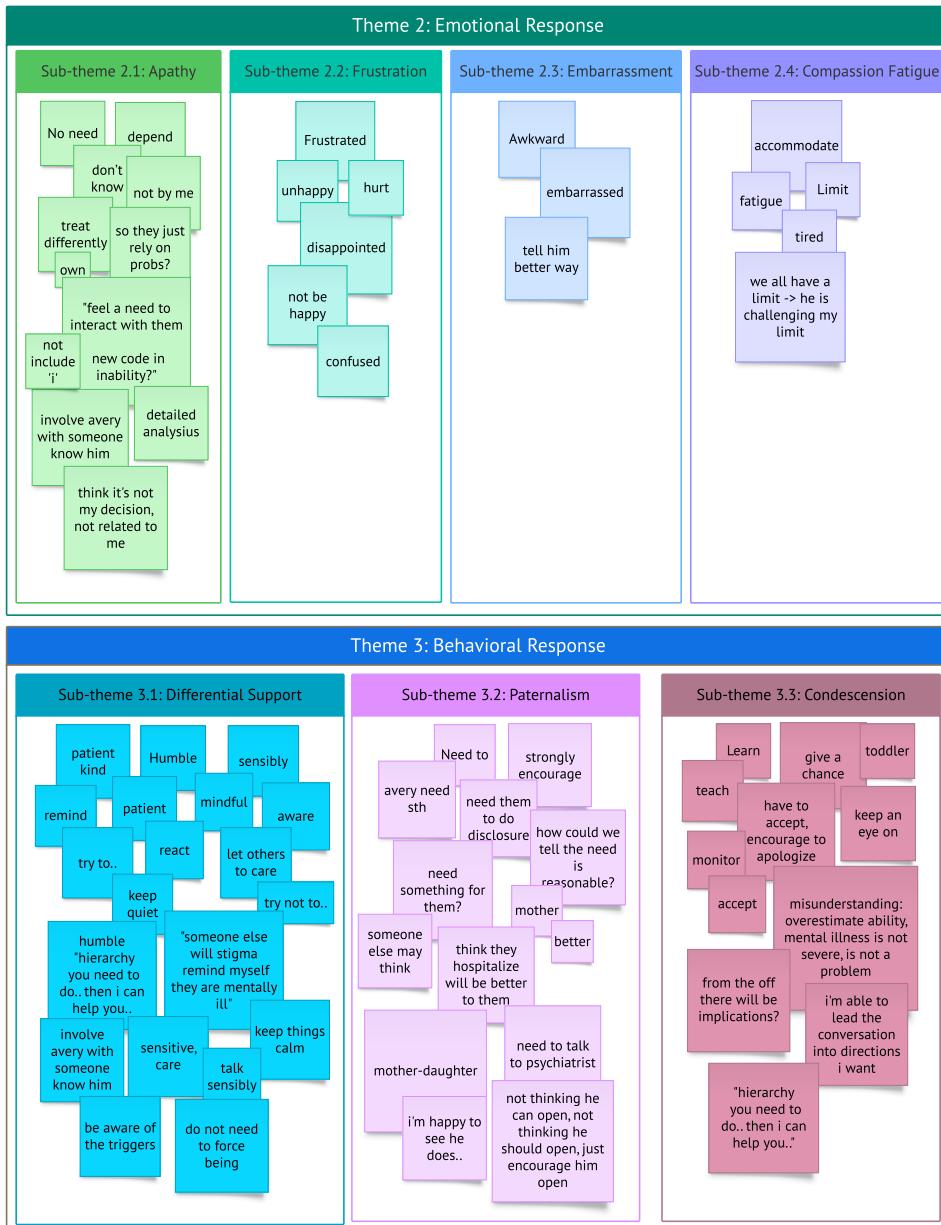


Fig. 15. Affinity diagram of theme 2 and 3.

Sub-theme 1.2: Victimization. Some participants equate mental illness with weakness and vulnerability, believing those with mental disorders are prone to victimization.

Code 1.2.1: Think they will be harmed by others. Some participants equated mental illness with weakness, helplessness, and a lack of ability to protect oneself, believing that individuals with mental disorders are susceptible to being victimized by others. Although previous research has indicated that individuals with mental illness face a higher risk of victimization compared to the

general population [64, 73, 77, 145, 154], applying this concern indiscriminately to every individual with a mental disorder while ignoring individual differences may constitute stigma. We should not broadly categorize all individuals with mental illness as potential victims or label them as such. While we respect the statistics provided in previous research [73], the participant's concern overlooks the individual differences among individuals with mental illness. Many patients are able to effectively manage their emotions and maintain normal lives and social interactions without triggering negative reactions from others. Applying the label of a potential victim to all patients is an inappropriate generalization.

Code 1.2.2: Think they will bring trauma. This is the belief that individuals with mental illness are likely to cause traumatic experiences for themselves or others, and may trigger distressing memories or thoughts in people. It suggests that interacting with those who have mental illnesses can elicit negative emotions or unease in those around them. This assumption is unfounded and can exacerbate fear and avoidance of individuals with mental illness. Generally, like everyone else, they do not inherently cause trauma by their mere presence. Every person has their own experiences, and interacting with others may evoke certain resonances to varying degrees; however, this stigmatization explicitly attributes this solely to those with mental illness.

Sub-theme 1.3: Crime-Prone: Belief that individuals with mental illness are prone to criminal behavior. A prevalent stigmatized belief is that individuals with mental illness are more likely to engage in criminal activities compared to the general population. This attribution differs from the perception that individuals with mental illness are dangerous (a stigmatization previously mentioned in the original attribution model [32]). It emphasizes that the actions of those with mental illness often necessitate public intervention, such as involvement from law enforcement. This stigma frequently stems from portrayals on social media [110] and past experiences with crime [122].

Sub-theme 1.4: Minimization. Minimization of mental illness is the act of downplaying, dismissing or trivializing the severity and impact of mental health conditions.

Code 1.4.1: Think individuals with mental illness will deny that they have mental illness. This notion suggests that individuals with mental illness lack insight into their own condition, and it reflects certain misconceptions about mental illness, such as the idea that denial is a common form of avoidance. However, previous research has indicated that individuals with mental illness are capable of recognizing their condition, and that refusing to acknowledge the illness is only one possible stage in the process of acceptance [144].

Code 1.4.2: Think individuals with mental illness have no problem and can recover themselves. This cognitive judgment underestimates the severity and impact of mental illness by suggesting that individuals with mental disorders do not actually have a problem or can recover on their own without proper treatment or support [81]. This view trivializes and dismisses the legitimate struggles and challenges faced by those living with mental health conditions.

Sub-theme 1.5: Disconnection and Functional Incapability. Disconnection encapsulates the stigmatizing beliefs that individuals with mental illness are not well-integrated into society, are incapable of functioning normally, and will ultimately face isolation due to their condition.

Code 1.5.1: Think individuals with mental illness are not well-integrated into society. This conveys a concern that individuals with mental illness face significant challenges in integrating effectively and positively into societal structures.

Think individuals with mental illness will have bad relationships with surroundings. There is a belief that individuals with mental illness will experience negative relationships with their environment or community [171]. This notion suggests that individuals with mental illnesses face difficulties integrating into normal society and establishing positive relationships with others.

Think individuals with mental illness are resistant to kindness. This stigmatization, unlike *social distance*, does not indicate an unwillingness to socialize with individuals with mental illness, but rather a belief that they do not wish to socialize with us. From another perspective, for the patients themselves, this notion may induce feelings of guilt and self-blame, leading them to believe that they “should” be more receptive to acts of kindness, thereby exacerbating their mental burden.

Think individuals with mental illness will be isolated finally. This view reflects the belief that individuals with mental illness will ultimately become completely socially isolated. However, this perspective underestimates the recovery potential of patients, exaggerates the risk of being marginalized, and disregards the fact that through social support, many patients can reintegrate into society. This stance equates mental illness with permanent social isolation [45, 108], which in itself constitutes an unfair stigmatization and labeling behavior.

Code 1.5.2: Think individuals with mental illness are incapable of functioning normally. This represents a belief in the functional impairment and inability of individuals with mental illness to function effectively.

Think individuals with mental illness are unable to work. The belief that individuals with mental illness are unable to maintain employment is a substantially common form of stigmatization [112, 147, 162]. Employment can promote broader social inclusion and is an important way for individuals with mental illness to meaningfully participate in the wider community. Stigma is both a proximate and distal cause of employment inequity for individuals with mental illness who experience direct discrimination due to prejudicial attitudes of employers and colleagues, as well as indirect discrimination resulting from historical patterns of disadvantage, structural disincentives against competitive employment, and generalized policy neglect. This is an issue that needs to be addressed urgently.

Think individuals with mental illness will never achieve self-recovery. A common misconception about depression is the belief that individuals with mental illness will never improve on their own and will be repeatedly affected by their condition [24, 105]. This view denies the possibility that patients can experience improvement or recovery through self-management and external support, equating mental illness with an incurable, lifelong state. Such a perspective potentially disparages the entire group of those affected. It may lead patients to develop thoughts of self-abandonment and helplessness, impacting their motivation to actively seek treatment and engage in self-management.

Think individuals with mental illness are weak because they cannot control themselves. This claim conveys a belief that individuals with mental illness lack self-control and are weak as a result, which is corresponding to the *controllability* dimension in the original attribution model [32] and serves as a complementary enrichment. The attribution model revealed that if individuals with mental illness are perceived as controllable, i.e., able to regulate their behaviors and emotions, they are deemed responsible for their actions. This sub-theme provides a complementary conclusion: if participants believe that the behaviors and emotions of individuals with mental illness are uncontrollable [21, 54], although it may not lead to stigmatization based on responsibility, it can result in the perception of weakness due to their inability to self-manage, disregarding potential pathological causes.

Theme 2: Emotional Responses. Emotional responses [32] refer to emotional reactions toward individuals with mental illness, representing a person’s affective state in response to internal or external stimuli.

Sub-theme 2.1: Apathy. This represents a feeling of lack of interest, enthusiasm, or concern about things that others find exciting or important. When individuals display cold, unsympathetic attitudes lacking compassion for individuals with mental illness, these attitudes themselves negatively label the mentally ill, deeming them unimportant and unworthy of attention, defaulting to viewing them as strangers [15].

Sub-theme 2.2 Frustration. Frustration is a feeling of dissatisfaction, annoyance, or disappointment resulting from the actions of individuals with mental illness. This frustration emphasizes a sense of powerlessness and despair [160], which differs from the *anger* described in the attribution model [32]. The primary reasons for the frustration may include the complex etiology and long, arduous treatment process of mental illnesses, which might lead to relapse and make family and friends feel helpless and powerless while accompanying the patient on the road to recovery. Furthermore, the required long-term care and companionship and the unwillingness of individuals with mental illness to seek help as well as their self-stigma [34] can create a sense of powerlessness due to the burden and psychological stress placed on family members [90]. Additionally, as outsiders, family and friends cannot fully understand their inner world and may find it difficult to discern and respond to their behaviors, leading to feelings of frustration, disappointment, and helplessness.

Sub-theme 2.3 Embarrassment. Embarrassment is a feeling of self-consciousness, discomfort, or shame caused by one's actions, circumstances, or appearance. Numerous participants report experiencing embarrassment at times when socializing with individuals with mental illness [2, 19]. Certain mental illnesses may possibly affect people's communication and social skills, rendering interactions more challenging. This barrier can lead to awkwardness for both parties involved. Conversely, some people worry about inadvertently reacting in a manner that might hurt or offend the other person, resulting in feelings of unease and discomfort.

Sub-theme 2.4: Compassion Fatigue. Compassion fatigue, a condition characterized by a state of tension and preoccupation, manifests itself through intrusive imagery, emotional numbness or avoidance, anxiety, hypervigilance, re-experiencing of traumatic events, irritability, and angry outbursts [138]. This phenomenon is not limited to caregivers [133], but can also occur in the context of social interactions. Several participants expressed that they had exercised considerable patience and tolerance, investing substantial emotional resources, which ultimately led to emotional exhaustion.

Theme 3: Behavioral Responses. Behavioral responses refer to observable reactions or actions in response to a specific situation, including special care, paternalism, and patronization.

Sub-theme 3.1: Differential Support. Some participants demonstrated an inclination to provide excessive attention or treatment to individuals with mental illness, a behavior that often implies the perceived incapacity or inferiority of these individuals. In an attempt to display kindness, empathy, or mindfulness in the presence of individuals with mental illness, certain participants reminded themselves to be cautious in their words and actions, to react sensibly to the behaviors of these individuals, and to humble themselves [86] during interactions. They expressed the need to be patient, mindful, and aware of potential triggers. Such conduct often stems from a desire to embrace a humbler mindset to accommodate those with mental illness or to subconsciously restrict oneself behaviorally to maintain normal social interactions. However, these actions ultimately reflect an attitude of excessive caution.

Sub-theme 3.2: Paternalism. Paternalism is a behavior characterized by adopting an overbearing or controlling attitude toward individuals with mental illness out of good intentions, often making decisions for them without their input or consent for their benefit [69, 148]. Some participants explicitly mentioned that their dynamic with individuals with mental illness was akin to a mother-daughter relationship, which is inherently unequal.

The key distinction between *paternalism* and *altruism* lies in their approach to autonomy. Paternalism involves imposing decisions or actions on others without necessarily respecting their autonomy, whereas altruism entails selflessly assisting others while respecting their agency and well-being. Paternalistic behavior often arises from a desire to implement measures that the participant believes to be correct, without considering the actual needs of individuals with mental illness.

Sub-theme 3.3: Condescension. Condescension toward individuals with mental illness involves adopting a condescending, controlling, or superior attitude, often rooted in the belief that one knows what is best for the person and can fix their problems [42]. This mindset is characterized by treating individuals with mental illness in a condescending or demeaning manner, assuming a position of authority, such as taking on a leadership or teaching role, and emphasizing the need to monitor, control, or change the person's behavior. Participants who exhibit patronizing attitudes may downplay the individual's ability [88], prioritizing their own comfort or perceived hierarchy over the person's needs. They may also imply that they are better equipped to handle the situation due to personal qualities, experience, or education. Moreover, patronization focuses on the perceived superiority of the “helper” rather than the needs and desires of the person being helped.

REFERENCES

- [1] Tahirah Abdullah and Tamara L. Brown. 2020. Diagnostic labeling and mental illness stigma among Black Americans: An experimental vignette study. *Stigma and Health* 5, 1 (2020), 11.
- [2] Shea J Addison and Susan J Thorpe. 2004. Factors involved in the formation of attitudes towards those who are mentally ill. *Social psychiatry and psychiatric epidemiology* 39 (2004), 228–234.
- [3] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. (2023). <https://doi.org/10.48550/arXiv.2307.02179>
- [4] Ramez M. Alkoudmani, Guat See Ooi, and Mei Lan Tan. 2023. Implementing a chatbot on Facebook to reach and collect data from thousands of health care providers: PharmindBot as a case. *Journal of the American Pharmacists Association* 63, 5 (2023), 1634–1642.e3. <https://doi.org/10.1016/j.japh.2023.06.007>
- [5] Nisreen Ameen, Gagan Deep Sharma, Shlomo Tarba, Amar Rao, and Ritika Chopra. 2022. Toward advancing theory on creativity in marketing and artificial intelligence. *Psychology & Marketing* 39, 9 (2022), 1802–1825. <https://doi.org/10.1002/mar.21699>
- [6] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT. *IEEE Intelligent Systems* 38, 2 (2023), 15–23. <https://doi.org/10.1109/MIS.2023.3254179>
- [7] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120, 41 (2023), e2311627120. <https://doi.org/10.1073/pnas.2311627120>
- [8] Moe Armstrong. 2003. Don't Call Me Nuts! Coping With the Stigma of Mental Illness. *Psychiatric Services* 54, 1 (2003), 116–116. <https://doi.org/10.1176/appi.ps.54.1.116>
- [9] Sabina Asensio-Cuesta, Vicent Blanes-Selva, J Alberto Conejero, Ana Frigola, Manuel G Portolés, Juan Francisco Merino-Torres, Matilde Rubio Almanza, Shabbir Syed-Abdul, Yu-Chuan (Jack) Li, Ruth Vilar-Mateo, Luis Fernandez-Luque, and Juan M García-Gómez. 2021. A User-Centered Chatbot (Wakamola) to Collect Linked Data in Population Networks to Support Studies of Overweight and Obesity Causes: Design and Pilot Study. *JMIR Med Inform* 9, 4 (14 Apr 2021), e17503. <https://doi.org/10.2196/17503>
- [10] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. 2023. *Using Large Language Models for Qualitative Analysis can Introduce Serious Bias*. Policy Research Working Paper Series 10597. The World Bank. <https://ideas.repec.org/p/wbk/wbrwps/10597.html>
- [11] American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. Washington, D.C. <https://doi.org/10.1176/appi.books.9780890425596>
- [12] Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring Qualitative Research Using LLMs. (2023). <https://doi.org/10.48550/arXiv.2306.13298>
- [13] S. Barber, P. C. Gronholm, S. Ahuja, N. Rüsch, and G. Thornicroft. 2020. Microaggressions towards people affected by mental health problems: a scoping review. *Epidemiology and Psychiatric Sciences* 29 (2020), e82. <https://doi.org/10.1017/S2045796019000763>
- [14] Anja Esther Baumann. 2007. Stigmatization, social distance and exclusion because of mental illness: The individual with mental illness as a ‘stranger’. *International Review of Psychiatry* 19, 2 (2007), 131–135. <https://doi.org/10.1080/09540260701278739>
- [15] Anja Esther Baumann. 2007. Stigmatization, social distance and exclusion because of mental illness: The individual with mental illness as a ‘stranger’. *International review of psychiatry* 19, 2 (2007), 131–135.

- [16] Richard E. Boyatzis. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Sage Publications, Inc, Thousand Oaks.
- [17] Jakob Ohme Brahim Zarouali, Theo Araujo and Claes de Vreese. 2023. Comparing Chatbots and Online Surveys for (Longitudinal) Data Collection: An Investigation of Response Characteristics, Data Quality, and User Evaluation. *Communication Methods and Measures* 0, 0 (2023), 1–20. <https://doi.org/10.1080/19312458.2022.2156489>
- [18] Loraine Busetto, Wolfgang Wick, and Christoph Gumbinger. 2020. How to use and assess qualitative research methods. *Neurological Research and Practice* 2, 14 (2020). <https://doi.org/10.1186/s42466-020-00059-z>
- [19] Peter Byrne. 2000. Stigma of mental illness and ways of diminishing it. *Advances in Psychiatric treatment* 6, 1 (2000), 65–72.
- [20] Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2018. Food Diary Coaching Chatbot. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore). Association for Computing Machinery, 1676–1680. <https://doi.org/10.1145/3267305.3274191>
- [21] Randolph CH Chan and Winnie WS Mak. 2016. Common sense model of mental illness: Understanding the impact of cognitive and emotional representations of mental illness on recovery through the mediation of self-stigma. *Psychiatry Research* 246 (2016), 16–24.
- [22] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. (2023). <https://doi.org/10.48550/arXiv.2306.14924>
- [23] Bonnie Chinh, Himanshu Zade, Abbas Ganji, and Cecilia Aragon. 2019. Ways of Qualitative Coding: A Case Study of Four Strategies for Resolving Disagreements. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI EA '19). Association for Computing Machinery, 1–6. <https://doi.org/10.1145/3290607.3312879>
- [24] Mi-Kyoung Cho and Mi Young Kim. 2024. Effects of an empathy enhancement program using patient stories on attitudes and stigma toward mental illness among nursing students. *Frontiers in Psychiatry* 14 (2024), 1304947.
- [25] Shaan Chopra, Rachael Zehrung, Tamil Arasu Shanmugam, and Eun Kyoung Choe. 2021. Living with Uncertainty and Stigma: Self-Experimentation and Support-Seeking around Polycystic Ovary Syndrome. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 202, 18 pages. <https://doi.org/10.1145/3411764.3445706>
- [26] Neta Kligler-Vilenchik Christian Baden and Moran Yarchi. 2020. Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora. *Communication Methods and Measures* 14, 3 (2020), 165–183. <https://doi.org/10.1080/19312458.2020.1803247>
- [27] Sherice N. Clarke, S Sushil, Katherine Dennis, Ung-Sang Lee, Andrea Gomoll, and Zaynab Gates. 2023. Developing Shared Ways of Seeing Data: The Perils and Possibilities of Achieving Intercoder Agreement. *International Journal of Qualitative Methods* 22 (2023), 16094069231160973. <https://doi.org/10.1177/16094069231160973>
- [28] David Collier and James Mahoney. 1996. Insights and Pitfalls: Selection Bias in Qualitative Research. *World Politics* 49, 1 (1996), 56–91. <https://doi.org/10.1353/wp.1996.0023>
- [29] Nancy L Collins and Lynn Carol Miller. 1994. Self-disclosure and liking: A meta-analytic review. *Psychological Bulletin* 116, 3 (1994), 457–475. <https://doi.org/10.1037/0033-2909.116.3.457>
- [30] Patrick W Corrigan. 2012. A toolkit for evaluating programs meant to erase the stigma of mental illness. *Illinois Institute of Technology* 10 (2012).
- [31] Patrick W. Corrigan, Annette Backs Edwards, Amy Green, Sarah Lickey Diwan, and David L. Penn. 2001. Prejudice, Social Distance, and Familiarity with Mental Illness. *Schizophrenia Bulletin* 27, 2 (01 2001), 219–225. <https://doi.org/10.1093/oxfordjournals.schbul.a006868>
- [32] Patrick W. Corrigan, Fred E. Markowitz, Amy Watson, David Rowan, and Mary Ann Kubiak. 2003. An Attribution Model of Public Discrimination Towards Persons with Mental Illness. *Journal of Health and Social Behavior* 44, 2 (2003), 162–179. <http://www.jstor.org/stable/1519806>
- [33] Patrick W. Corrigan and Jenessa R. Shapiro. 2010. Measuring the impact of programs that challenge the public stigma of mental illness. *Clinical Psychology Review* 30, 8 (2010), 907–922. <https://doi.org/10.1016/j.cpr.2010.06.004>
- [34] Patrick W. Corrigan, Amy C. Watson, and Leah Barr. 2006. The Self-Stigma of Mental Illness: Implications for Self-Esteem and Self-Efficacy. *Journal of Social and Clinical Psychology* 25, 8 (2006), 875–884. <https://doi.org/10.1521/jscp.2006.25.8.875>
- [35] Jennifer Crocker, Brenda Major, and Claude Steele. 1998. *Social Stigma* (4th ed.). McGraw-Hill. 504–553 pages.
- [36] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9993–10001. <https://doi.org/10.18653/v1/2023.findings-emnlp.669>

- [37] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9993–10001. <https://doi.org/10.18653/v1/2023.findings-emnlp.69>
- [38] Mary Rose Day, Helen Mulcahy, Patricia Leahy-Warren, and Johanna Downey. 2015. Self-neglect: A case study and implications for clinical practice. *British Journal of Community Nursing* 20, 3 (2015), 110–115.
- [39] Stefano De Paoli. 2023. Performing an inductive thematic analysis of semi-structured interviews with a large language model: an exploration and provocation on the limits of the approach. *Social Science Computer Review* (7 Dec. 2023). <https://doi.org/10.1177/08944393231220483>
- [40] Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margaret Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Danielle Krettek-Cobb, Leslie Lai, Nirel Jones-Mitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology* 2 (2023), 688–701. <https://doi.org/10.1038/s44159-023-00241-5>
- [41] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. (2022). <https://doi.org/10.48550/arXiv.2301.00234>
- [42] Karen M Douglas and Robbie M Sutton. 2011. Constructive or cruel? Positive or patronizing? Reactions to expressions of positive and negative stereotypes of the mentally ill. *British Journal of Psychology* 102, 1 (2011), 97–107.
- [43] Jakub Drápal, Hannes Westermann, and Jaromír Savelka. 2023. Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies. (2023). <https://doi.org/10.48550/arXiv.2310.18729>
- [44] Zackary Okun Dunivin. 2024. Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. (2024). <https://doi.org/10.48550/arXiv.2401.15170>
- [45] Danielle Elisha, David Castle, and Barbara Hocking. 2006. Reducing social isolation in people with mental illness: the role of the psychiatrist. *Australasian Psychiatry* 14, 3 (2006), 281–284.
- [46] Wilmer Stalin Erazo, Germán Patricio Guerrero, Carlos Carrión Betancourt, and Iván Sánchez Salazar. 2020. Chatbot Implementation to Collect Data on Possible COVID-19 Cases and Release the Pressure on the Primary Health Care System. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. 0302–0307. <https://doi.org/10.1109/IEMCON51383.2020.9284846>
- [47] Sindhu Kiranmai Ernala, Asra F. Rizvi, Michael L. Birnbaum, John M. Kane, and Munmun De Choudhury. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 43 (dec 2017). <https://doi.org/10.1145/3134678>
- [48] Melike Yönder Ertem and Mehmet Karakaş. 2022. Validity and reliability study of the mental illness micro-aggression scale-perpetrator version. *Archives of Psychiatric Nursing* 41 (2022), 234–240. <https://doi.org/10.1016/j.apnu.2022.09.005>
- [49] Giuseppe Fenza, Francesco Orciuoli, Angela Peduto, and Alberto Postiglione. 2023. Healthcare Conversational Agents: Chatbot for Improving Patient-Reported Outcomes. In *Advanced Information Networking and Applications*, Leonard Barolli (Ed.). Springer International Publishing, Cham, 137–148.
- [50] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. <https://doi.org/10.1177/160940690600500107>
- [51] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. <https://doi.org/10.1177/160940690600500107>
- [52] Greg Fisher and Herman Aguinis. 2017. Using Theory Elaboration to Make Theoretical Advancements. *Organizational Research Methods* 20, 3 (2017), 438–464. <https://doi.org/10.1177/1094428116689707>
- [53] Robert J. Fisher. 1993. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research* 20, 2 (09 1993), 303–315. <https://doi.org/10.1086/209351>
- [54] Stephen Foster and Molly O’Mealey. 2022. Socioeconomic status and mental illness stigma: the impact of mental illness controllability attributions and personal responsibility judgments. *Journal of Mental Health* 31, 1 (2022), 58–65.
- [55] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and Individual Differences* 7, 3 (1986), 385–400. [https://doi.org/10.1016/0191-8869\(86\)90014-0](https://doi.org/10.1016/0191-8869(86)90014-0)
- [56] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.
- [57] Yasir Gamieldien, Jennifer M. Case, and Andrew Katz. 2023. Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding. (June 21 2023). <http://dx.doi.org/10.2139/ssrn.4487768>
- [58] Arianna M. Gibson, Brittany L. Lindsay, and Andrew C. H. Szeto. 2023. Mental illness stigma and microaggressions: An experimental study of familiarity and relationship quality. *Stigma and Health* (2023). <https://doi.org/10.1037/sah0000471> Advance online publication.
- [59] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120. <https://doi.org/10.1073/pnas>.

2305016120

- [60] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.
- [61] Erving Goffman. 1964. Stigma: Notes on the Management of Spoiled Identity. *Social Forces* 43, 1 (10 1964), 127–128. <https://doi.org/10.1093/sf/43.1.127>
- [62] Lauren Gonzales, Kristin C. Davidoff, Joseph S. DeLuca, and Philip T. Yanos. 2015. The mental illness microaggressions scale-perpetrator version (MIMS-P): Reliability and validity. *Psychiatry Research* 229, 1 (2015), 120–125. <https://doi.org/10.1016/j.psychres.2015.07.061>
- [63] Lauren Gonzales, Kristin C. Davidoff, Kevin L. Nadal, and Philip T. Yanos. 2015. Microaggressions experienced by persons with mental illnesses: An exploratory study. *Psychiatric Rehabilitation Journal* 38, 3 (2015), 234–241. <https://doi.org/10.1037/prj0000096>
- [64] Lisa A. Goodman, Michelle P. Salyers, Kim T. Mueser, Stanley D. Rosenberg, Marvin Swartz, Susan M. Essock, Fred C. Osher, Marian I. Butterfield, Jeffrey Swanson, Jerilynn Lamb-Pagone, Barbara J. Burns, Keith G. Meador, Jeffrey W. Swanson, Mary E. Becker, Hayden B. Bosworth, Richard Frothingham, Ronnie D. Horner, Lauren M. McIntyre, Patricia M. Spivey, Karen M. Stechuchak, Lisa J. Garber, Jean S. Gearon, Richard W. Goldberg, John D. Herron, Raymond S. Hoffman, Corina L. Riismandel, Patricia C. Auciello, Robert E. Drake, Mark C. Iber, Ravindra Luckoor, Gemma R. Skillman, Rosemarie S. Wolfe, and Robert M. Vidaver. 2001. Recent victimization in women and men with severe mental illness: Prevalence and correlates. *Journal of Traumatic Stress* 14, 4 (2001), 615–632. <https://doi.org/10.1023/A:1013026318450>
- [65] Annabelle Gourlay, Gerry Mshana, Isolde Birdthistle, Grace Bulugu, Basia Zaba, and Mark Urassa. 2014. Using vignettes in qualitative research to explore barriers and facilitating factors to the uptake of prevention of mother-to-child transmission services in rural Tanzania: a critical analysis. *BMC medical research methodology* 14, 1 (2014), 1–11.
- [66] Sandra Graham, Bernard Weiner, and Gail Sahar Zucker. 1997. An Attributional Analysis of Punishment Goals and Public Reactions to O. J. Simpson. *Personality and Social Psychology Bulletin* 23, 4 (1997), 331–346. <https://doi.org/10.1177/0146167297234001>
- [67] S Gunstone. 2003. Risk assessment and management of patients whom self-neglect: a ‘grey area’ for mental health workers. *Journal of Psychiatric and Mental Health Nursing* 10, 3 (2003), 287–296.
- [68] Eradah O Hamad, Marie Y Savundranayagam, Jeffrey D Holmes, Elizabeth Anne Kinsella, and Andrew M Johnson. 2016. Toward a Mixed-Methods Research Approach to Content Analysis in The Digital Age: The Combined Content-Analysis Model and its Applications to Health Care Twitter Feeds. *J Med Internet Res* 18, 3 (08 Mar 2016), e60. <https://doi.org/10.2196/jmir.5391>
- [69] Johannes Hamann, Markus Büchner, and Nicolas Rüsch. 2017. Self-Stigma and Consumer Participation in Shared Decision Making in Mental Health Services. *Psychiatric Services* 68, 8 (2017), 783–788. <https://doi.org/10.1176/appi.ps.201600282>
- [70] Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. 2021. Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI ’21*). Association for Computing Machinery, New York, NY, USA, Article 389, 15 pages. <https://doi.org/10.1145/3411764.3445569>
- [71] Ainul Nadhirah Hanafiah and Tine Van Bortel. 2015. A qualitative exploration of the perspectives of mental health professionals on stigma and discrimination of mental illness in Malaysia. *International Journal of Mental Health Systems* 9 (2015), 10. <https://doi.org/10.1186/s13033-015-0002-1>
- [72] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI ’15*). Association for Computing Machinery, New York, NY, USA, 95–104. <https://doi.org/10.1145/2702123.2702561>
- [73] Michelle N. Harris, Miranda L. Baumann, Brent Teasdale, and Bruce G. Link. 2022. Estimating the Relationship Between Perceived Stigma and Victimization of People With Mental Illness. *Journal of Interpersonal Violence* 37, 3–4 (2022), NP1226–NP1252. <https://doi.org/10.1177/0886260520926326>
- [74] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. (2023). <https://doi.org/10.48550/arXiv.2303.16854>
- [75] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao‘Kenneth’ Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. (2024). <https://doi.org/10.48550/arXiv.2402.16795>
- [76] Claire Henderson, Sara Evans-Lacko, Clare Flach, and Graham Thornicroft. 2012. Responses to Mental Health Stigma Questions: The Importance of Social Desirability and Data Collection Method. *The Canadian Journal of Psychiatry* 57, 3 (2012), 152–160. <https://doi.org/10.1177/070674371205700304>

- [77] Virginia Aldigé Hiday, Marvin S. Swartz, Jeffrey W. Swanson, Randy Borum, and H. Ryan Wagner. 1999. Criminal Victimization of Persons With Severe Mental Illness. *Psychiatric Services* 50, 1 (1999), 62–68. <https://doi.org/10.1176/ps.50.1.62>
- [78] Ashish Hingle, Andrew Katz, and Aditya Johri. 2023. Exploring NLP-Based Methods for Generating Engineering Ethics Assessment Qualitative Codebooks. In *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–8. <https://doi.org/10.1109/FIE58773.2023.10342985>
- [79] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (05 2018), 712–733. <https://doi.org/10.1093/joc/jqy026>
- [80] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Using ChatGPT to fight misinformation: ChatGPT nails 72% of 12,000 verified claims. *PsyArXiv. April* 3 (2023).
- [81] Robert W Hopkins, Michela M David, and Lindy A Kilik. 2014. A re-examination of behaviour in depression: Have we grossly underestimated the extent and impact of the behavioural suffering? *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 46, 3 (2014), 456.
- [82] Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker. 2024. Prompt-based and Fine-tuned GPT Models for Context-Dependent and-Independent Deductive Coding in Social Annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 518–528.
- [83] Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker. 2024. Prompt-based and Fine-tuned GPT Models for Context-Dependent and-Independent Deductive Coding in Social Annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 518–528.
- [84] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA). Association for Computing Machinery, New York, NY, USA, 294–297. <https://doi.org/10.1145/3543873.3587368>
- [85] Ninni Ihälainen-Tamlander, Anu Vähäniemi, Eliisa Löytynniemi, Tarja Suominen, and Maritta Välimäki. 2016. Stigmatizing attitudes in nurses towards people with mental illness: A cross-sectional study in primary settings in Finland. *Journal of Psychiatric and Mental Health Nursing* 23, 6–7 (2016), 427–437.
- [86] Masunga K. Iseselo, Lusajo Kajula, and Khadija I. Yahya-Malima. 2016. The psychosocial problems of families caring for relatives with mental illnesses and their coping strategies: a qualitative urban based study in Dar es Salaam, Tanzania. *BMC Psychiatry* 16, 1 (2016), 146.
- [87] Tunazzina Islam and Dan Goldwasser. 2024. Uncovering Latent Themes of Messaging on Social Media by Integrating LLMs: A Case Study on Climate Campaigns. (2024). <https://doi.org/10.48550/arXiv.2403.10707>
- [88] Jake Jackson. 2017. Patronizing depression: epistemic injustice, stigmatizing attitudes, and the need for empathy. *Journal of Social Philosophy* 48, 3 (2017), 359–376.
- [89] Jacinta Jardine, Camille Nadal, Sarah Robinson, Angel Enrique, Marcus Hanratty, and Gavin Doherty. 2024. Between Rhetoric and Reality: Real-world Barriers to Uptake and Early Engagement in Digital Mental Health Interventions. *ACM Trans. Comput.-Hum. Interact.* 31, 2, Article 27 (feb 2024), 59 pages. <https://doi.org/10.1145/3635472>
- [90] Anita Johansson, Agneta Anderzen-Carlsson, Arne Åhlén, and Birgitta Andershed. 2012. Fathers' everyday experiences of having an adult child who suffers from long-term mental illness. *Issues in Mental Health Nursing* 33, 2 (2012), 109–117.
- [91] Annalee Johnson-Kwochka, Kyle S. Minor, Leslie Ashburn-Nardo, Wei Wu, Laura G. Stull, and Michelle P. Salyers. 2024. A new look at the attribution model: Considerations for the measurement of public mental illness stigma. *Stigma and Health* 9, 1 (2024), 1–8. <https://doi.org/10.1037/sah0000288>
- [92] Payam Kaywan, Khandakar Ahmed, Yuan Miao, Ayman Ibaida, and Bruce Gu. 2021. DEPRA: An Early Depression Detection Analysis Chatbot. In *Health Information Science*, Siuly Siuly, Hua Wang, Lu Chen, Yanhui Guo, and Chunxiao Xing (Eds.). Springer International Publishing, Cham, 193–204.
- [93] John G. Keip, Michael F. Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K. Burke, Hanga Galfalvy, Maria A. Oquendo, and J. John Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of Affective Disorders* 140, 1 (2012), 75–81. <https://doi.org/10.1016/j.jad.2012.01.045>
- [94] Taewan Kim, Mintra Ruensuk, and Hwajung Hong. 2020. In Helping a Vulnerable Bot, You Help Yourself: Designing a Social Bot as a Care-Receiver to Promote Mental Health and Reduce Stigma. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, 1–13. <https://doi.org/10.1145/3313831.3376743>
- [95] Elisabeth Kirsten, Annalina Buckmann, Abraham Mhaidli, and Steffen Becker. 2024. Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding. (2024). <https://doi.org/10.48550/arXiv.2403.06607>
- [96] Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. (2023). <https://doi.org/10.48550/arXiv.2303.03953>

- [97] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Chapter 11 - Analyzing qualitative data. In *Research Methods in Human Computer Interaction (Second Edition)* (second edition ed.), Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser (Eds.). Morgan Kaufmann, Boston, 299–327. <https://doi.org/10.1016/B978-0-12-805390-4.00011-X>
- [98] Yi-Chieh Lee, Yichao Cui, Jack Jamieson, Wayne Fu, and Naomi Yamashita. 2023. Exploring Effects of Chatbot-based Social Contact on Reducing Mental Illness Stigma. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI ’23*). Association for Computing Machinery, New York, NY, USA, Article 488, 16 pages. <https://doi.org/10.1145/3544548.3581384>
- [99] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI ’20*). 1–12. <https://doi.org/10.1145/3313831.3376175>
- [100] William Leeson, Adam Resnick, Daniel Alexander, and John Rovers. 2019. Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study. *International Journal of Qualitative Methods* 18 (2019), 1609406919887021. <https://doi.org/10.1177/1609406919887021>
- [101] Robert P Lennon, Robbie Fraleigh, Lauren J Van Scy, Aparna Keshaviah, Xindi C Hu, Bethany L Snyder, Erin L Miller, William A Calo, Aleksandra E Zgierska, and Christopher Griffin. 2021. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family Medicine and Community Health* 9, Suppl 1 (2021). <https://doi.org/10.1136/fmch-2021-001287>
- [102] Ang Li, Xiaoxiao Huang, Dongdong Jiao, Bridianne O’Dea, Tingshao Zhu, and Helen Christensen. 2018. An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. *Asia-Pacific Psychiatry* 10, 1 (2018), e12314. <https://doi.org/10.1111/appy.12314>
- [103] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI ’20*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376209>
- [104] Xiaomiao Li, Brittany L. Lindsay, Andrew C. H. Szeto, and Keith S. Dobson. 2023. An Examination of the Effect of Feedback on Meta-ignorance of Mental Illness Public Stigma. *Metacognition Learning* 18 (2023), 165–181. <https://doi.org/10.1007/s11409-022-09325-8>
- [105] Laura Limiñana-Bravo and Francisco José Eiroa-Orosa. 2017. Training for mental health professionals: towards rightsbased care. *Avances en Psicología Clínica. Asociación Española de Psicología Conductual* (2017).
- [106] Bruce G. Link, Francis T. Cullen, James Frank, and John F. Wozniak. 1987. The Social Rejection of Former Mental Patients: Understanding Why Labels Matter. *Amer. J. Sociology* 92, 6 (1987), 1461–1500. <https://doi.org/10.1086/228672>
- [107] Bruce G. Link, Lawrence H. Yang, Jo C. Phelan, and Pamela Y. Collins. 2004. Measuring Mental Illness Stigma. *Schizophrenia Bulletin* 30, 3 (01 2004), 511–541. <https://doi.org/10.1093/oxfordjournals.schbul.a007098>
- [108] Sheila J Linz and Bonnie A Sturm. 2013. The phenomenon of social isolation in the severely mentally ill. *Perspectives in psychiatric care* 49, 4 (2013), 243–254.
- [109] Stephan Linzbach, Tim Tressel, Laura Kallmeyer, Stefan Dietze, and Hajira Jabeen. 2023. Decoding Prompt Syntax: Analysing its Impact on Knowledge Retrieval in Large Language Models. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX USA) (*WWW ’23 Companion*). Association for Computing Machinery, New York, NY, USA, 1145–1149. <https://doi.org/10.1145/3543873.3587655>
- [110] Christopher Ryan Locke. 2010. *Public attitudes toward mental illness: An experimental design examining the media’s impact of crime on stigma*. Ph. D. Dissertation. The Ohio State University.
- [111] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8086–8098. <https://doi.org/10.18653/v1/2022.acl-long.556>
- [112] Alison Luciano and Ellen Meara. 2014. Employment status of people with mental illness: national survey data from 2009 and 2010. *Psychiatric Services* 65, 10 (2014), 1201–1209.
- [113] Jasmine R Marcellin, Dawd S Siraj, Robert Victor, Shaila Kotadia, and Yvonne A Maldonado. 2019. The Impact of Unconscious Bias in Healthcare: How to Recognize and Mitigate It. *The Journal of Infectious Diseases* 220, Supplement_2 (08 2019), S62–S73. <https://doi.org/10.1093/infdis/jiz214>
- [114] Howard Margolis. 1987. *Patterns, thinking, and cognition: A theory of judgment*. University of Chicago Press.
- [115] Mark J. Martinko, Paul Harvey, and Scott C. Douglas. 2007. The role, function, and contribution of attribution theory to leadership: A review. *The Leadership Quarterly* 18, 6 (2007), 561–585. <https://doi.org/10.1016/j.lequa.2007.09.004>
- [116] Steven Marwaha and Gill Livingston. 2002. Stigma, racism or choice. Why do depressed ethnic elders avoid psychiatrists? *Journal of Affective Disorders* 72, 3 (2002), 257–265. [https://doi.org/10.1016/S0165-0327\(01\)00470-0](https://doi.org/10.1016/S0165-0327(01)00470-0)

- [117] Patrick J. Michaels and Patrick W. Corrigan. 2013. Measuring mental illness stigma with diminished social desirability effects. *Journal of Mental Health* 22, 3 (2013), 218–226. <https://doi.org/10.3109/09638237.2012.734652>
- [118] II Middleton, Stewart D. 2020. *Understanding the Effects of Mental Illness Stigma on the Attitudes of Undergraduate and Graduate Students from a Micro-Aggression Theoretical Perspective*. Ph.D. Dissertation. <https://www.proquest.com/dissertations-theses/understanding-effects-mental-illness-stigma-on/docview/2410112585/se-2>
- [119] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. <https://doi.org/10.1145/3544548.3581225>
- [120] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.
- [121] Ann A Murphy. 2014. *Does attribution theory explain mental illness stigma?* ProQuest Information & Learning.
- [122] Claire Nee and Clare Witt. 2013. Public perceptions of risk in criminality: The effects of mental illness and social disadvantage. *Psychiatry Research* 209, 3 (2013), 675–683. <https://doi.org/10.1016/j.psychres.2013.02.013>
- [123] Laura K. Nelson. 2020. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research* 49, 1 (2020), 3–42. <https://doi.org/10.1177/0049124117729703>
- [124] Laura K. Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research* 50, 1 (2021), 202–237. <https://doi.org/10.1177/0049124118769114>
- [125] Kimberly A. Neuendorf. 2017. *An Integrative Approach to Content Analysis* (second ed.). SAGE Publications, Inc., 36–69. <https://doi.org/10.4135/9781071802878>
- [126] Wei Jie Ong, Shazana Shahwan, Chong Min Janrius Goh, Gregory Tee Hng Tan, Siow Ann Chong, and Mythily Subramaniam. 2020. Daily Encounters of Mental Illness Stigma and Individual Strategies to Reduce Stigma – Perspectives of People With Mental Illness. *Frontiers in Psychology* 11 (2020). <https://doi.org/10.3389/fpsyg.2020.590844>
- [127] Stacy L. Overton and Sondra L. Medina. 2008. The Stigma of Mental Illness. *Journal of Counseling & Development* 86, 2 (2008), 143–151. <https://doi.org/10.1002/j.1556-6678.2008.tb00491.x>
- [128] Angelina Parfenova. 2024. Automating the Information Extraction from Semi-Structured Interview Transcripts. (2024). <https://doi.org/10.48550/arXiv.2403.04819>
- [129] Giorgianna L. Passerello, James E. Hazelwood, and Stephen Lawrie. 2019. Using Twitter to assess attitudes to schizophrenia and psychosis. *BJP Psych Bulletin* 43, 4 (2019), 158–166. <https://doi.org/10.1192/bjb.2018.115>
- [130] Heather J. Peters, Haley N. Schwenk, Zachary R. Ahlstrom, and Lyndzie N. McLalwain. 2017. Microaggressions: The experience of individuals with mental illness. *Counselling Psychology Quarterly* 30, 1 (2017), 86–112. <https://doi.org/10.1080/09515070.2016.1164666>
- [131] Ava Podrazhansky, Hao Zhang, Meng Han, and Selena He. 2020. A Chatbot-Based Mobile Application to Predict and Early-Prevent Human Mental Illness. In *Proceedings of the 2020 ACM Southeast Conference* (Tampa, FL, USA). Association for Computing Machinery, 311–312. <https://doi.org/10.1145/3374135.3385319>
- [132] James Prather, Brent N. Reeves, Paul Denny, Brett A. Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. “It’s Weird That it Knows What I Want”: Usability and Interactions with Copilot for Novice Programmers. *ACM Trans. Comput.-Hum. Interact.* 31, 1, Article 4 (nov 2023), 31 pages. <https://doi.org/10.1145/3617367>
- [133] Susan L Ray, Carol Wong, Dawn White, and Kimberly Heaslip. 2013. Compassion satisfaction, compassion fatigue, work life conditions, and burnout among frontline mental health care professionals. *Traumatology* 19, 4 (2013), 255–267.
- [134] Michael V. Reiss. 2023. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. (2023). <https://doi.org/10.48550/arXiv.2304.11085>
- [135] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. 2022. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior* 126 (2022), 107034. <https://doi.org/10.1016/j.chb.2021.107034>
- [136] K. Andrew R. Richards and Michael A. Hemphill. 2022. A Practical Guide to Collaborative Qualitative Data Analysis. *Journal of Teaching in Physical Education* 37, 2 (2022), 225 – 231. <https://doi.org/10.1123/jtpe.2017-0084>
- [137] Laura Weiss Roberts. 2002. Ethics and Mental Illness Research. *Psychiatric Clinics of North America* 25, 3 (2002), A525–A545.
- [138] Alberto Rossi, Gaia Cetrano, Riccardo Pertile, Laura Rabbi, Valeria Donisi, Laura Grigoletti, Cristina Curtolo, Michele Tansella, Graham Thornicroft, and Francesco Amaddeo. 2012. Burnout, compassion fatigue, and compassion satisfaction among staff in community-based mental health services. *Psychiatry research* 200, 2-3 (2012), 933–938.
- [139] Johnny Saldana. 2016. *The Coding Manual for Qualitative Researchers* (3rd ed.). Sage Publications, Inc, London.

- [140] Corina Sas, Steve Whittaker, and John Zimmerman. 2016. Design for Rituals of Letting Go: An Embodiment Perspective on Disposal Practices Informed by Grief Therapy. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article 21 (aug 2016), 37 pages. <https://doi.org/10.1145/2926714>
- [141] Anita Schick, Jasper Feine, Stefan Morana, Alexander Maedche, and Ulrich Reininghaus. 2022. Validity of Chatbot Use for Mental Health Assessment: Experimental Study. *JMIR Mhealth Uhealth* 10, 10 (31 Oct 2022), e28082. <https://doi.org/10.2196/28082>
- [142] Ryan M. Schuetzler, G. Mark Grimes, Justin Scott Giboney, and Jay F. Jr. Nunamaker. 2018. The Influence of Conversational Agents on Socially Desirable Responding. *Information Systems and Quantitative Analysis Faculty Publications* 61 (2018). <https://digitalcommons.unomaha.edu/isqafacpub/61>
- [143] Joel Sebastian and Deborah Richards. 2017. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. *Computers in Human Behavior* 73 (2017), 479–488. <https://doi.org/10.1016/j.chb.2017.03.071>
- [144] Zindel V. Segal and Doreen D. Vella. 1990. Self-schema in major depression: Replication and extension of a priming methodology. *Cognitive Therapy and Research* 14 (1990), 161–176. <https://doi.org/10.1007/BF01176207>
- [145] ERIC SILVER. 2002. MENTAL DISORDER AND VIOLENT VICTIMIZATION: THE MEDIATING ROLE OF INVOLVEMENT IN CONFLICTED SOCIAL RELATIONSHIPS*. *Criminology* 40, 1 (2002), 191–212. <https://doi.org/10.1111/j.1745-9125.2002.tb00954.x>
- [146] Jonathan A Smith and Mike Osborn. 2015. Interpretative phenomenological analysis as a useful methodology for research on the lived experience of pain. *British Journal of Pain* 9, 1 (2015), 41–42. <https://doi.org/10.1177/2049463714541642>
- [147] Heather Stuart. 2006. Mental illness and employment discrimination. *Current opinion in psychiatry* 19, 5 (2006), 522–526.
- [148] Laura G Stull, John H McGrew, Michelle P Salyers, and Leslie Ashburn-Nardo. 2013. Implicit and explicit stigma of mental illness: Attitudes in an evidence-based practice. *The Journal of nervous and mental disease* 201, 12 (2013), 1072–1079.
- [149] Muhammad Arsyad Subu, Del Fatma Wati, Netrida Netrida, Vetty Priscilla, Jacqueline Maria Dias, Mini Sarah Abraham, Shameran Slewa-Younan, and Nabeel Al-Yateem. 2021. Types of stigma experienced by patients with mental illness and mental health nurses in Indonesia: a qualitative content analysis. *International Journal of Mental Health Systems* 15 (2021), 77. <https://doi.org/10.1186/s13033-021-00502-x>
- [150] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. (2023). <https://doi.org/10.48550/arXiv.2305.08377>
- [151] Robert H Tai, Lillian R Bentley, Xin Xia, Jason M Sitt, Sarah C Fankhauser, Ana M Chicas-Mosier, and Barnas G Monteith. 2024. An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods* 23 (2024), 16094069241231168.
- [152] Mariska E te Pas, Werner G M M Rutten, R Arthur Bouwman, and Marc P Buise. 2020. User Experience of a Chatbot Questionnaire Versus a Regular Computer Questionnaire: Prospective Comparative Study. *JMIR Med Inform* 8, 12 (7 Dec 2020), e21982. <https://doi.org/10.2196/21982>
- [153] Ramya Tekumalla and Juan M. Banda. 2023. Leveraging Large Language Models and Weak Supervision for Social Media data annotation: an evaluation using COVID-19 self-reported vaccination tweets. (2023). <https://doi.org/10.48550/arXiv.2309.06503>
- [154] Linda A. Teplin, Gary M. McClelland, Karen M. Abram, and Dana A. Weiner. 2005. Crime Victimization in Adults With Severe Mental Illness: Comparison With the National Crime Victimization Survey. *Archives of General Psychiatry* 62, 8 (08 2005), 911–921. <https://doi.org/10.1001/archpsyc.62.8.911>
- [155] Maya Grace Torii, Takahito Murakami, and Yoichi Ochiai. 2024. Expanding Horizons in HCI Research Through LLM-Driven Qualitative Analysis. (2024). <https://doi.org/10.48550/arXiv.2401.04138>
- [156] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. (2023). <https://doi.org/10.48550/arXiv.2304.06588>
- [157] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yunling Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poult, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Bin Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

<https://doi.org/10.48550/arXiv.2307.09288>

- [158] Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. (2024). <https://doi.org/10.48550/arXiv.2402.05129>
- [159] Thea F van de Mortel. 2008. Faking It: Social Desirability Response Bias in Self-report Research. *The Australian Journal of Advanced Nursing* 25, 4 (2008), 40–48. <https://search.informit.org/doi/10.3316/informit.210155003844269>
- [160] Susan Van Den Tillaart, Donna Kurtz, and Penny Cash. 2009. Powerlessness, marginalized identity, and silencing of health concerns: Voiced realities of women living with a mental health diagnosis. *International journal of mental health nursing* 18, 3 (2009), 153–163.
- [161] Lucy L. Purgason W. Bradley McKibben, Rochelle Cade and Edward Wahesh. 2022. How to Conduct a Deductive Content Analysis in Counseling Research. *Counseling Outcome Research and Evaluation* 13, 2 (2022), 156–168. <https://doi.org/10.1080/21501378.2020.1846992>
- [162] Geoff Waghorn and Chris Lloyd. 2005. The employment of people with mental illness. *Australian e-journal for the Advancement of Mental Health* 4, 2 (2005), 129–171.
- [163] Mark Warner, Juan F. Maestre, Jo Gibbs, Chia-Fang Chung, and Ann Blandford. 2019. Signal Appropriation of Explicit HIV Status Disclosure Fields in Sex-Social Apps used by Gay and Bisexual Men. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3290605.3300922>
- [164] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [165] Shirli Werner, Patrick Corrigan, Nicole Ditchman, and Kristin Sokol. 2012. Stigma and intellectual disability: A review of related measures and future directions. *Research in developmental disabilities* 33, 2 (2012), 748–765.
- [166] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 83 (sep 2023), 36 pages. <https://doi.org/10.1145/3603625>
- [167] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23 Companion). Association for Computing Machinery, 75–78. <https://doi.org/10.1145/3581754.3584136>
- [168] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, 1–14. <https://doi.org/10.1145/3313831.3376131>
- [169] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. *ACM Trans. Comput.-Hum. Interact.* 27, 3, Article 15 (jun 2020). <https://doi.org/10.1145/3381804>
- [170] Yeqing Yan, Peng Zheng, and Yongjun Wang. 2024. Enhancing large language model capabilities for rumor detection with Knowledge-Powered Prompting. *Engineering Applications of Artificial Intelligence* 133 (2024), 108259. <https://doi.org/10.1016/j.engappai.2024.108259>
- [171] Philip T. Yanos, Sarah Rosenfield, and Allan V. Horwitz. 2001. Negative and supportive social interactions and quality of life among persons diagnosed with severe mental illness. *Community Mental Health Journal* 37 (2001), 405–419.
- [172] Andres Felipe Zambrano, Xiner Liu, Amanda Barany, Ryan S Baker, Juhan Kim, and Nidhi Nasiar. 2023. From ncoder to chatgpt: From automated coding to refining human coding. In *International conference on quantitative ethnography*. Springer, 470–485.
- [173] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building Emotional Support Chatbots in the Era of LLMs. (2023). <https://doi.org/10.48550/arXiv.2308.11584>
- [174] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. (2023). <https://doi.org/10.48550/arXiv.2304.10145>
- [175] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* (03 2024), 1–55. https://doi.org/10.1162/coli_a_00502