Towards Augmenting Mental Health Personnel with LLM Technology to Provide More Personalized and Measurable Treatment Goals for Patients with Severe Mental Illnesses

Lorenzo J. James[1,2(✉)], Maureen Maessen[1], Laura Genga[1], Barbara Montagne[2], Muriel A. Hagenaars[3], and Pieter M. E. Van Gorp[1]

[1] Industrial Engineering and Information Systems, Eindhoven University of Technology, Eindhoven, The Netherlands
l.j.james@tue.com
[2] Treatment Center for Personality Disorders, GGZ Centraal, Center for Mental Health Care, Disorders, Amersfoort, The Netherlands
[3] Department of Clinical Psychology, Universiteit Utrecht, Utrecht, The Netherlands

Abstract. Mobile health (mHealth) tools are increasingly being used in various mental health domains to monitor patients with Severe Mental Illnesses (SMI), with the aim of potentially increasing patient engagement with their treatment. Patients with SMI who are prescribed Flexible Assertive Community Treatment (FACT) create a treatment plan together with their case manager, which serves as the leading document describing the goals that will be worked on during treatment. In order to incorporate the treatment plan goals of a patient in an mHealth application, the treatment plan goals need to be measurable. However, in previous work, we discovered that on average, only 25% of the available treatment plans include measurable goals. We have developed a protocol for making measurable goals with patients with SMI to address this issue. However, we anticipate low adoption of the protocol due to the potentially time-consuming nature of the steps involved. To mitigate this, we are exploring the use of AI to generate measurable treatment plan goals for patients with SMI and introduce a new workflow. In our exploratory study, we created a prototype of a system that may enable case managers and patients with SMI to generate measurable treatment plan goals using Large Language Models.

Keywords: SMI · mHealth · LLM · Gamification · Goals

1 Introduction

The impact of mental health is a growing concern for individuals and communities in our present-day society. Severe mental illnesses such as bipolar disorder,

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 186–200, 2024.
https://doi.org/10.1007/978-3-031-59717-6_13

PTSD, schizophrenia, and severe depression pose unique challenges for both patients and mental healthcare providers [19]. One such challenge is mental health professionals' need to dynamically scale the care for each patient diagnosed with Severe Mental Illnesses (SMI) depending on the current state of the patient. Among people with ill mental health, those diagnosed with one or more severe mental illnesses for a period of over two years, and who struggle socially from their mental illnesses, are considered patients with SMI [24]. In the last few decades, there has been a growth in the number of patients with SMI treated by mental health care institutes, increasing the workload and work pressure of healthcare professionals [11]. Many patients with SMI live independently at home or in assisted living facilities without direct help from friends or family. Due to the associated vulnerabilities experienced by patients with SMI, there is a need for long-term treatment, with the ability to scale the care to the needs of the patient. Flexible Assertive Community Treatment (FACT) is a type of treatment for patients with SMI, that treats patients within their own home environment and provides care that matches their current needs [24]. Several countries have opted to prescribe FACT to patients with SMI, in an attempt to treat patients within their home environment [22]. A case manager is a healthcare professional who is directly responsible for monitoring the effects of the treatment. During FACT treatment, patients work together with their case manager to create a treatment plan, which includes the treatment plan goals that they will work on over a one-year period [24]. The goals of the treatment plan not only target symptom recovery but also seek to empower patients, enhance their self-reliance, and provide support in addressing social issues such as employment and housing. The goals serve as guiding principles of the treatment, ensuring that the care provided to patients is in accordance with their own goals, focusing on a person-oriented and holistic approach [24].

Currently, case managers assess the functioning of their patients through direct contact with the patients or their surroundings (e.g., family, general practitioner, etc.). Outside of direct contact, case managers do not have any tools to monitor the state of their patients, which results in them not being able to assess when to scale care for patients efficiently. Previous research shows that mobile Health (mHealth) applications are promising when it comes to positively influencing and tracking the behaviors of patients with SMI [13]. Such an mHealth application could potentially also improve FACT by assisting case managers in assessing their patients' functioning and monitoring the progress that a patient is making on their treatment [13]. This could potentially be a support tool to help case managers efficiently assess when to scale care for patients.

In order to monitor patients with SMI who are prescribed FACT, an mHealth application should track the progress a patient is making on their treatment plan goals. However, in previous research in collaboration with FACT teams at a Dutch Mental Health and Addiction Care Institute, we have discovered that on average only 25% of the available treatment plans have a form of measurable goals available within them [14]. A well-defined Measurable goal is a goal that can be tracked to monitor progress [3]. The treatment plan goals were revealed to have a significant lack of structure, consistency, and difference in level of detail.

This results in the majority of the goals currently found in the treatment plans not being suitable to be used in mHealth applications [14].

To assist case managers with making measurable treatment plan goals that can be used in mHealth applications, we introduced a protocol for a structured approach to creating measurable treatment plan goals for patients with SMI. Despite case managers being positive about the protocol, due to its time-consuming nature, it will have a low adoption rate. Given the high workload of case managers, they are not always open to new time-consuming tasks. Following all the steps in the protocol is a more time-consuming process for the case manager, compared to their usual goal-setting and tracking workflow. To reduce the time case managers would spend following the protocol when creating and tracking measurable goals, we will explore ways to leverage AI to potentially automate a part of the workflow.

The recent breakthroughs of Large Language Models (LLMs), such as BERT, GPT-3, and GPT-4 have been disruptive. These LLMs have been trained on large data sets and the models available have produced impressive results when prompted by humans. When prompted correctly, these models have the ability to respond to specific queries given to them when provided a specific context [2]. Due to this technology's ability to generate relevant text responses when given a context, we will explore the potential use of LLMs to introduce a new workflow for case managers to create treatment plan goals with their patients.

This study will aim to create a prototype that explores using LLMs to create AI-generated treatment plan goals, potentially improving the workflow of creating treatment plan goals with patients. Evaluating the prototype is outside the scope of this project. Nevertheless, we will assess the goals generated by the prototype to evaluate whether the quality of generated goals can be improved through modifications to the prototype and the prompts sent to the LLM.

2 Theoretical Background

To establish how to create measurable goals, in this section, we will review relevant work in the areas of behavior change, the use of AI in mental healthcare, and the possible risks of generating goals for patients with SMI using LLMs.

2.1 Behavior Change Theories

According to behavior change theories such as the COM-B system and the Fogg behavior model, behavior is a product of three fundamental factors: capability, opportunity, and motivation [8,18]. To successfully perform a targeted behavior at a particular time, it is essential to have the capability and opportunity, including an enabling environment. The strength of motivation to engage in the behavior must be higher than any other competing behaviors. These three factors interact to produce the desired behavior [20]. Motivation can be split into intrinsic and extrinsic motivation. Where extrinsic motivation is being motivated by external factors, intrinsic motivation is motivated by an inherent interest which leads to more persistence [20].

According to the Self-Determination Theory (SDT), in the context of an mHealth tool, a tool that satisfies the need for autonomy (i.e., the desire to have control over tasks), competence (i.e., the need to acquire new skills), and relatedness (i.e., the desire to feel connected to others) can enhance intrinsic motivation [15,20]. The proposed workflow of using LLMs to generate treatment plan goals can potentially enhance the patient's intrinsic motivation, which in turn can lead to better engagement with the treatment goals. Finally, the Goal Setting Theory states that specific, challenging goals and appropriate feedback contribute to higher and better results. For engaging a person in a target behavior, goals must follow five principles: clarity, challenge, commitment, feedback, and task complexity [17]. Clear goals are Specific, Measurable, Attainable, Realistic, and Time-oriented (SMART) [4]. We will be using the SDT and Goal Setting Theory as criteria for evaluating the quality of the goals that are generated by LLMs. The generated goals will also be structured in the format of SMART goals, to the extent that we can accomplish this.

Treatment outcome goals focus on a result (e.g., losing 2 kg of weight), while treatment behavioral goals focus on an individual's action (e.g., going for a walk) [18]. For this study, we consider outcome goals as overall treatment plan goals, and behavior goals as smaller goals patients should achieve to reach their desired outcome goals. In order to track the progress toward the treatment goals, the behavior goals need to be measurable (e.g., I will go for a walk twice a week throughout the month of September). Integrating behavioral goals in mHealth interventions could potentially lead to more successful interventions [7]. An additional useful tool to motivate users in mHealth interventions is Gamification, which is the use of game design elements in non-game contexts [5].

2.2 AI in Mental Healthcare

While there has been a growing trend toward integrating AI technology in physical health applications, adopting such technology within the mental health domain has been comparatively slow [10]. Mental health practitioners emphasize patient-centered care and a hands-on approach in their clinical practice in contrast to non-psychiatric practitioners. This approach involves softer skills, including establishing strong relationships with patients and closely observing their behaviors and emotions [9]. Mental health clinical data is frequently in qualitative and subjective patient statements and written notes. Because of this, there is still much to be gained in the field of mental health practice through the incorporation of AI technology [10]. The application of AI techniques could present the opportunity to develop more accurate pre-diagnosis screening tools and risk models to determine an individual's susceptibility or likelihood of developing a mental illness [21]. AI-based LLMs have already showcased their efficacy in diverse areas, such as explainable AI, conversational agents, education, information retrieval, and text summarizing [6]. With their remarkable capabilities, LLMs can potentially transform various industries [10]. Research on LLM technology for mental health has yielded mixed results, and the enduring effects of using LLM on mental health remain unexplored [12]. The potential of LLMs in

healthcare stems from the ability to process and learn from massive amounts of free-text data. In theory, LLMs could generate measurable treatment goals based on data available on the web. Currently, much of the responsibility of creating treatment plan goals falls under the workload of the case manager [24]. In previous work, we created measurable goals for patients with SMI together with case managers [14]. Even when provided with a protocol for creating measurable goals, case managers found creating measurable goals for their patients a time-consuming and difficult task. Generating measurable goals using LLMs, could potentially alleviate some of the workload from case managers to patients, as it could empower patients to easily formulate measurable goals for their treatment. To the best of our knowledge, no literature is available that discusses the use of LLMs to generate treatment goals for patients with SMI.

The recent LLMs that have been released for public use usually strictly abide by the relevant laws and regulations of the countries in which they are released [16]. The LLMs usually do not mention anything offensive, violent or criminal in their conversations, and do not give any unethical medical advice. However, it is known that LLMs are known to hallucinate [1]. Hallucinations in the context of LLMs are when the system generates information not found in its training set [1]. This could potentially cause the generated goals to contain harmful elements. It is important to note that case managers are responsible for ensuring the safety of the goals within the patient's treatment plan. If patients are allowed to generate treatment plan goals unchecked using LLM technology there could be potential risks associated with that. Therefore it is recommended for both patients and case managers to assess the generated plan goals before adding them to the treatment plan.

Currently, electronic treatment records are utilized by mental health institutes to keep records of patients, including their treatment goals. However, these systems lack the inclusion of measurable goals. In our prototype, we explore features that can possibly add measurable goals to these systems.

3 Methods

To explore the feasibility of using LLMs to generate treatment plan goals for patients with SMI, a prototype of such a system was built. The functioning prototype was built using the GPT-3 LLM, the prototyping tool Mendix, and the gamification engine Gamebus. The architecture of the system can be seen in Fig. 1. The goals generated by the prototype were then evaluated using behavior change theory guidelines by a group of students. To evaluate a prototype with patients with SMI, a prerequisite entails securing the approval of the study by at least three ethical committees. In addition to obtaining ethical approval from the university, the mental health institute mandates an external committee to ascertain whether the study qualifies as medical-scientific research before conducting its own ethical review. Considering that the study is in the exploratory phase, evaluating the basic capabilities of LLMs and prompt engineering strategies to evaluate if generated goals could possibly be improved upon, it was deemed

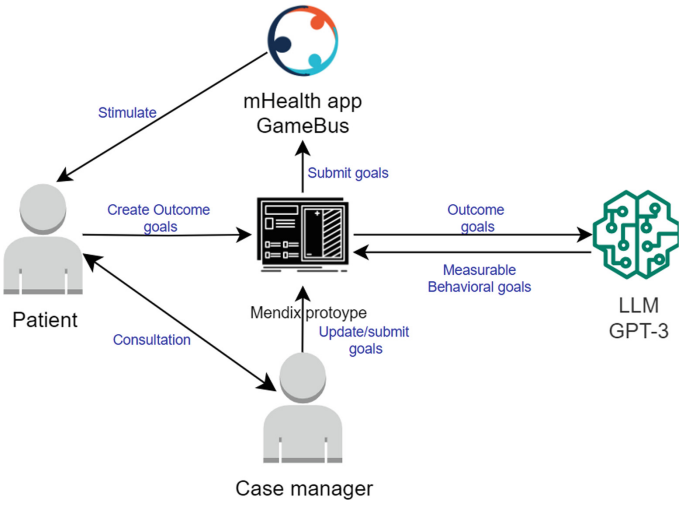

Fig. 1. The architecture of the prototype.

appropriate to first evaluate the goals with students before entering the ethical procedures. Once the basic capabilities have been established we will again involve patients and case managers in the process.

3.1 Prototype

Mendix. Mendix is a low-code development platform that was chosen to develop the prototype of the new goal-setting workflow, because of its ease of use built-in capabilities with REST API architectures, and ability to facilitate rapid prototype development. GPT-3 was accessed using the available API. The responses were formatted and directly implemented in both Mendix and the GameBus system. As the new workflow is being constructed, specific milestones were identified for evaluating the goals. For this purpose, behavior change theories were utilized. While three main theories were outlined in the theoretical background, in the scope of this study we focus on SDT and goal-setting theory.

GameBus. GameBus is a digital platform that promotes a healthy lifestyle by hosting and facilitating healthy challenges and competitions. Users are motivated to continue working on the challenges available to them through the use of gamification elements such as leaderboards and points [23]. For this study, we configured the GameBus platform as an mHealth application. The GameBus application can use measurable goals as input for challenges.

3.2 Measurable Goals

To assess if it is possible to improve the quality of the measurable goals that GPT-3 generates, by prompt engineering. We recruited a group of 8 students

from the Eindhoven University of Technology to assess the generated goals, over 5 iterations. The students received a survey that included an explanation of the project and a video of the latest version of the prototype. They were asked to rate the generated behavioral goals for two outcome goals and provide tips for improving the goals. The two outcome goals chosen for evaluation were based on data on treatment plan goals from patients with SMI, provided by a Mental Health and Addiction Care Institute in the Netherlands. Participants rated the goals based on the following elements of the SDT and Goal Setting theory: autonomy, competence, relatedness, clarity, challenge, commitment, feedback, and task complexity to assess the quality of the goals. This rating used a five-point Likert scale which consists of 1 = Strongly Disagree (SD); 2 = Disagree (D); 3 = Neutral (N); 4 = Agree (S); and 5 = Strongly Agree (SA). Each aspect was analyzed, and the average score was used to evaluate the goals.

We define a change in the average score of an element, compared to the previous average score of that element, as significant when there was a difference of at least 0.5. With eight aspects, the maximum score for a goal was 35. This approach allowed participants to provide specific ratings for each aspect, ensuring an assessment of their perceptions. By implementing the dimensions within the goal-setting theory and SDT frameworks into the goals, we increase the potential intrinsic motivation a person has to complete the goals. The group of students was also given a plain text field to provide feedback.

The information gathered, along with the provided feedback, guides the areas for improvement. The prompt in the prototype was revised according to the chosen improvements. Once these changes were implemented, the group was requested to complete the updated survey again. This iterative process aims to use prompt engineering to refine the goals based on participant feedback. By incorporating this, the prompt sent to GPT-3 can be improved and the quality of the generated behavioral goals can potentially be increased.

4 Results

4.1 Prototype

Mendix. The created Mendix prototype allows users to log in as patients or case managers. Patients can create, edit, and delete treatment outcome goals. For each treatment goal, the system generates 3 behavioral goals using GPT-3's API. Patient information and goal format instructions are included in the prompt sent to GPT-3. GPT-3's response is mapped and stored in the data model created in Mendix, and the goals are assigned to the patient. Patients can view and modify their behavioral goals. They can also remove or regenerate specific behavioral goals. After finalizing their treatment outcome goals, patients submit them to their case manager for approval before adding them to the treatment plan in the Mendix prototype. Patients have an overview page displaying all their approved and non-approved treatment outcome goals along with associated behavioral goals, each with a step-by-step guide. Case managers can also generate, edit, and remove goals for their assigned patients. Case managers are able to see all

the submitted goal requests, requests for goal changes, and goal edit requests by their patients. Case managers can also accept or reject these requests, save the treatment plan in the Mendix prototype, and ultimately submit the treatment plan to the GameBus application.

GameBus. The GameBus application has been extended to receive the treatment plan goals of the Mendix prototype, through the API. The Mendix prototype sends an API request to the GameBus application, sending a patient's approved treatment plans to the GameBus application in JSON format. The GameBus application converts the treatment plan it receives into GameBus challenges and saves the treatment outcome and behavioral goals within the treatment plan, in the GameBus data model. The patients using the Mendix prototype to create the treatment plan goals, also have access to the mHealth configuration of the GameBus application. The GameBus mHealth application supports multiple modular gamification and personalization options that have been designed to increase the intrinsic motivation of GameBus users. GameBus can personalize the challenges of users by tailoring the elements within them such as adjusting the frequency of the tasks needed to be completed within the challenge, adjusting the time and date a challenge should be completed or tracked, and allowing challenges to be done in groups. The GameBus application can be configured to use gamification such as giving points to users when they complete tasks within challenges, displaying a leaderboard displaying user scores, and awarding users loot boxes. The system administrator of the GameBus application can easily configure these personalization and gamification elements to be visible to any of the GameBus users. Each of the patient accounts in the Mendix prototype has a GameBus account assigned and can log into the GameBus application. Due to the challenges in the GameBus application now deriving directly from the treatment plan of the patients the challenges present in the application are even more relevant to the patient's treatment.

We extended the GameBus application to include the gamification element of levels. The relevant Mendix treatment plan goals of the patients are mapped into a level structure where each outcome goal is separated into different levels with increasing difficulty. The levels can be unlocked by completing the relevant behavior goals associated with the outcome goal. This level extension showcases the potential for patients to be more engaged with their treatment through this mHealth application, as the gamification and personalization elements are designed to maximize intrinsic motivation to the challenges in the application. The GameBus application also makes it possible to monitor which challenges each user is taking part in and track the progress users made on each of their challenges. Case managers can use this data to track which challenges their patients are making progress on, when their patients work on tasks within their challenges, and when patients have completed their challenges. The data could then potentially be used as a proxy for user engagement with their treatment.

4.2 Evaluating Generated Measurable Goals

The group of 8 students that was enlisted to fill in the 5 rounds of surveys to assess the behavioral goals rated the LLM-generated behavioral goals formulated for the following two outcome goals: 1) I want to learn how to deal with negative thoughts about events that have happened in the past. 2) I want to learn to discuss my fears, but I especially want to understand my fears.

In the following section, we describe the results of each evaluation phase and the overall changes that were made to the prompts and goals based on the results of the previous phase. An overview of the survey results can be seen in Table 1.

Table 1. Displays the average scores assigned by the participants to goals 1 (G1) and 2 (G2) based on how the goals scored on the elements of the SDT and Goal Setting Theory. Scores for evaluation phases 1–3 and 5 are presented, including total and combined average scores (AS) for both goals in each phase.

Evaluation phase	1			2			3			5		
	G1	G2	AS	G1	G2	AS	G1	G2	AS	G1	G2	AS
Autonomy	4.125	4	4.063	3.625	3.125	3.375	3.375	3.25	3.313	4	4.125	4.063
Competence	3.375	3.625	3.5	4.5	3.75	4.125	3.75	3.625	3.688	4.25	4.25	4.25
Relatedness	1.875	4.5	3.188	1.5	4.5	4	3.375	3.25	3.313	4	3.875	3.938
Clarity	3.625	4.125	3.875	4.5	3.75	4.125	3.375	3.875	3.625	4.625	4.5	4.563
Challenge	3.875	3.875	3.875	3.25	4.125	3.688	4.25	4.25	4.25	4.5	4.25	4.375
Commitment	3.75	3.5	3.625	4.25	3.5	3.875	4	3.125	3.563	4	4	4
Feedback	1.75	3.875	2.813	2	3.375	2.688	2.625	3.25	2.938	3.625	3.75	3.688
Task complexity	3.375	3.375	3.375	3.625	3.625	3.625	3.75	4	3.875	4.25	4.125	4.188
Total score	25.7	28.375	28.313	27.25	29.75	28.5	28.5	28.625	28.563	33.25	32.875	33.063

Evaluation Phase 1

The first version of the prototype features a user interface where patients can input their outcome goal and receive a corresponding behavioral goal. Additionally, the case manager also has the option to enter treatment goals for their assigned patients. The prompt used in this first iteration to generate a goal from GPT-3, only includes asking for one behavioral goal based on an outcome goal.

In the evaluation, based on the survey, participants felt a high level of autonomy in pursuing both behavioral goals, with an average score of just above 4. They clearly understood the behavioral goals and found them moderately challenging. Commitment and competence to the behavioral goals were strong across both goals. However, there were differences in relatedness and feedback. Through the open text field, participants claimed to have felt stronger social support in goal 2 than in goal 1, this may be because goal 1 is carried out alone whereas goal 2 involves others. Clarity and task complexity were consistent across both goals, indicating a similar understanding of the goals and tasks' complexity. The average total score: 28.313 out of 35 points.

Despite clarity receiving a relatively high average score of 3.875, there is still room for improvement. There were several comments that the participants did

not understand the goals well. It is important that goals are clearly defined. Therefore, the goals in the next evaluation phase will be presented in a more structured manner, in the format of SMART goals.

Instead of goals being generated in standard text form, each goal will now include the following attributes which are based on the SMART goal structure: goal, duration, frequency, frequency scale, and time period. An example of a structured goal can be seen in Fig. 2. This structured approach also facilitates an easier transformation to JSON format.

Evaluation Phase 2

In this version of the prototype, the patient is now able to modify the goals at the attribute level. Case managers now also have the ability to evaluate the goals of their patients, by allowing them to approve, modify, or reject goals. Once a goal is evaluated and approved by a case manager, the patient will receive a notification and be able to view the approved goals in their treatment plan.

The prompt has been modified for GPT-3 to respond in a more structured manner. To ensure conciseness, it is emphasized that only a subject, verb, and object are allowed in the generated goal.

The main goal of this evaluation phase was to improve clarity, which increased significantly for goal 1 but decreased for goal 2. Overall clarity increased by 0.25. For goal 1 the dimensions of competence and commitment significantly increased, whereas the dimensions of autonomy and challenge saw a significant decrease. For goal 2, autonomy and feedback significantly decreased. For the average score, autonomy was significantly reduced, but competence significantly increased. The average total score: is 28.5 out of 35 points. This is a slight increase of 0.187 compared to the previous evaluation phase.

As autonomy is currently one of the lower-scoring aspects, this improvement area will be investigated. Providing three behavioral goals instead of one has the potential to increase autonomy. Combined with the ability to modify attributes, this can lead to greater personalization and foster a stronger sense of ownership. This decision also may indicate that a single behavioral goal as a response per outcome goal may not be sufficient to achieve the desired outcome.

Evaluation Phase 3

In this version of the prototype, goals have the same structure as in evaluation phase 2, however, patients now receive three behavioral goals per outcome goal, instead of one. Initially, for outcome goal 2, three identical behavioral goals were generated. Although, the goal attributes varied. This occurred despite configuring GPT-3 to maximize the randomness of its responses.

The main goal of this evaluation phase was to improve autonomy. This slightly increased for goal 2, but decreased for goal 1. Overall there was a slight decrease in autonomy. The other results also varied from the last round. The biggest changes were found for goal 1. The dimensions of relatedness, challenge,

and feedback significantly increased, whereas competence and clarity significantly decreased. For goal 2, there was a significant decrease in relatedness. For the average score, clarity significantly decreased. Average total score: 28.563 out of 35 points. A slight increase compared to the previous evaluation phase.

So far, the prompt has remained general without specifying the target group for whom the goals are intended. The next direction is to explore if giving the context that the generated behavioral goals are for patients with SMI would impact the response. Additionally, it would be interesting to investigate how including specific diagnoses would influence behavioral goals. This could potentially lead to more tailored results.

Evaluation Phase 4

The prompt has undergone two modifications. Firstly, by explicitly stating in the initial sentence that the treatment goals are intended for patients with a severe mental illness and emphasizing the goal's suitability for this individual. Secondly, an additional sentence has been included to specify the patient's diagnosis with an emphasis that the goal should be attainable for someone with this diagnosis.

Tests were conducted on these two scenarios, however, the behavioral goals did not show a significant change compared to the results from the previous evaluation phase. Modifying the attributes of the goal also yielded similar responses as before. As a result, this version will not undergo further surveys since the behavioral goals did not exhibit significant changes. It remains uncertain whether this new prompt would impact other goals.

In the earlier surveys, respondents expressed uncertainty about how to carry out the goal effectively and desired additional support through guidelines or a step-by-step plan that would enable them to pursue the goal effectively. As a result, addressing this issue will be the focus of the next evaluation phase.

Evaluation Phase 5

Attributes were added to the prompt to also include a step-by-step plan on how to achieve the behavioral goals, supportive tips to aid progress, and an explanation of the goal.

The behavioral goals have the same structure as in evaluation phase 4, with these three attributes added. As occurred in evaluation phase 3 the generated goals appeared to be very similar. In the case of goal 1, three separate behavioral goals were initially generated, but two were highly familiar, and therefore one was modified by the researcher using the edit function. The main goal was to improve clarity, which increased significantly in both goals. The other survey results also varied quite from the last evaluation phase. The most important changes were for goal 1. The dimensions of autonomy, competence, relatedness, feedback, and task complexity significantly increased. For goal 2, there was a significant increase in autonomy, competence, relatedness, commitment, and feedback. For the average

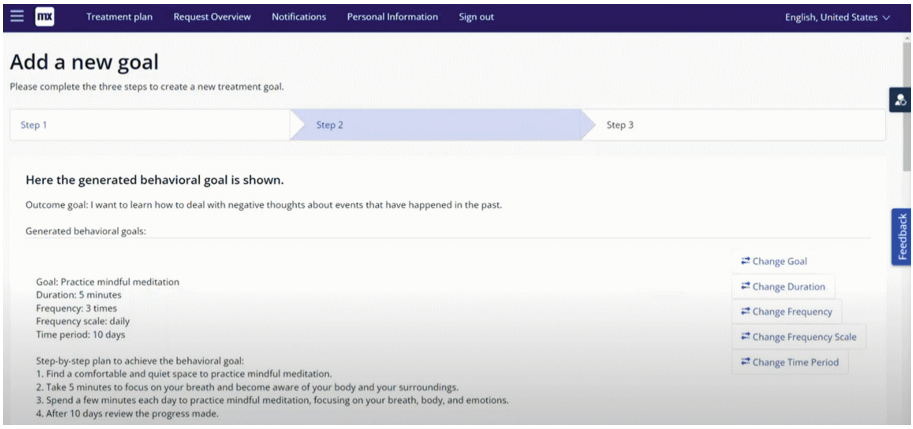

Fig. 2. Screenshot of the working prototype, showing one of the generated treatment plan goals and the option to edit the goal and its attributes.

score, autonomy, competence, relatedness, clarity, and feedback increased. Average total score: 33.063 out of 35 points. This is an increase of 4.5 compared to the previous evaluation phase.

5 Discussion

Study Limitations

In this study, it is important to note that the evaluation of the prototype itself fell outside the scope of the study. Consequently, aspects related to user-friendliness, UI, and UX elements were also outside the scope of the study. It is also worth highlighting that no testing was conducted with case managers or patients with SMI. Our primary focus was centered on determining whether modifications to the prototype and prompts could lead to improvements in the quality of the generated goals. In the evaluation phase, only the behavioral goals of two outcome goals were evaluated by students, the quality of the generated goals may differ depending on how outcome goals are structured and the type of outcome goals provided. The goals were also not evaluated by case managers on their quality, it is possible that they could consider other elements.

A major risk of using LLMs to generate goals for patients with SMI is that LLMs may hallucinate and could potentially generate goals that may not align with standard FACT treatment which may result in further complications for the patient. Accordingly, we performed some exploratory adversarial attacks on the prototype to expose potential vulnerabilities in the system when it comes to generating goals that may potentially not be in line with FACT treatment. A list of 26 treatment outcome goals that were deemed potentially harmful by the researchers, was used as input in the latest version of the prototype. The treatment behavioral goals given as output by the prototype were then inspected

by the researchers. Upon analysis of the generated behavioral goals resulting from the adversarial attacks performed. We did not experience any hallucinations, however, more extensive adversarial attacks need to be conducted in order to assess the potential harm hallucinations may cause. Therefore, the evaluation of goals by the case manager is crucial before adding a goal to the treatment plan. In the exploratory adversarial attacks, we only tested the prototype with GPT-3 as its LLM, if another LLM were to be used it is unknown what the responses and how safe the generated goals would be. More elaborate and in-depth adversarial attacks need to be done in further research, to assess the safety of using LLMs to generate treatment plan goals. Lastly, this feasibility study specifically focused on setting goals for a specific target group, namely Dutch patients with SMI who are prescribed FACT treatment. Therefore, conclusions drawn from this study cannot be generalized to other (non) SMI groups without further research.

Future Work

In future work, we should recruit case managers, and patients with SMI to evaluate the goal-setting workflow and the quality of goals, as the results from healthy participants cannot be generalized to people with SMI. It is worth evaluating and tracking the progress patients with SMI are making toward the generated goal, through an mHealth application like GameBus, compared to non-LLM generated goals. The integration of SDT and Goal-setting has proven valuable in this research, but including other behavior theories, such as the COM-B, could enhance the assessment. COM-B is a valuable tool to address instances where individuals lack the necessary preconditions for certain behaviors. For example, if the generated goal says to 'ride a bicycle' it becomes an unfeasible goal if the patient does not possess a bicycle. Incorporating the COM-B model into an interactive goal-setting approach can address these situations more effectively and take the prerequisites needed to complete a goal into consideration. Mental healthcare professionals could also be given a hands-on session with the system in a usability study. An addition to the user interface that could potentially make it easier for case managers and patients to select treatment plan goals that are relevant to the treatment is to create an interface that allows users to easily swipe irrelevant goals away, and save relevant ones. Another interesting direction to explore is to investigate the willingness and privacy concerns of patients with SMI regarding using such technology in their treatment. Given the current level of distrust among patients with SMI, understanding their attitudes and perceptions toward AI-based tools and data privacy implications would provide valuable insights. Such research is essential for ensuring the ethical and effective integration of such technology in the treatment journey of patients with SMI.

6 Conclusion

This feasibility study aimed to explore the use of LLMs in enabling patients with SMI and case managers to easily create measurable treatment goals for

treatment plans. The study suggests that it is feasible to develop a system that utilizes LLMs to allow users to generate measurable treatment plan goals. Using the prototype to create measurable goals, does make it possible to directly add meaningful goals to gamified mHealth systems like GameBus, which could potentially intrinsically motivate patients further to work on their treatment goals. Although the quality of the generated goals was not assessed by healthcare professionals or patients with SMI, the evaluation process with students indicated that incremental improvements in behavior goals were attainable. Furthermore, the study revealed that the application of SDT and Goal-Setting theory could enhance the quality of behavioral goals. However, improving a certain aspect and maintaining a consistent appreciation of all other aspects can be challenging. Now that it has been established that the proposed workflow has the potential to improve the process of case managers creating goals with their patients with SMI. We can now evaluate the system within the treatment process, and gather patient and case manager feedback. It is important to acknowledge that the results of this research should be considered preliminary, as there are currently no comparable findings in the existing literature for appropriate comparisons. These preliminary findings provide a foundation for further investigation and highlight the need for future research in this area.

References

1. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus **15**(2) (2023)
2. Arora, A., Arora, A.: The promise of large language models in health care. Lancet **401**(10377), 641 (2023)
3. Bovend'Eerdt, T.J., Botell, R.E., Wade, D.T.: Writing smart rehabilitation goals and achieving goal attainment scaling: a practical guide. Clin. Rehabil. **23**(4), 352–361 (2009)
4. Chase, J.A., Houmanfar, R., Hayes, S.C., Ward, T.A., Vilardaga, J.P., Follette, V.: Values are not just goals: online act-based values training adds to goal setting in improving undergraduate college student performance. J. Contextual Behav. Sci. **2**(3–4), 79–84 (2013)
5. Cheng, V.W.S., Davenport, T., Johnson, D., Vella, K., Hickie, I.B.: Gamification in apps and technologies for improving mental health and well-being: systematic review. JMIR Ment. Health **6**(6), e13717 (2019)
6. Dale, R.: GPT-3: what's it good for? Nat. Lang. Eng. **27**(1), 113–118 (2021)
7. Eckerstorfer, L.V., et al.: Key elements of mhealth interventions to successfully increase physical activity: meta-regression. JMIR Mhealth Uhealth **6**(11), e10076 (2018)
8. Fogg, B.J.: A behavior model for persuasive design. In: Proceedings of the 4th international Conference on Persuasive Technology, pp. 1–7 (2009)
9. Gabbard, G.O., Crisp-Han, H.: The early career psychiatrist and the psychotherapeutic identity. Acad. Psychiatry **41**, 30–34 (2017)
10. Graham, S., et al.: Artificial intelligence for mental health and mental illnesses: an overview. Curr. Psychiatry Rep. **21**, 1–18 (2019)

11. van Greuningen, M., Borgs, B.: Feiten en cijfers over mensen met een ernstige psychiatrische aandoening (2022). https://www.vektis.nl/intelligence/publicaties/factsheet-ernstige-psychiatrische-aandoeningen
12. Hamdoun, S., Monteleone, R., Bookman, T., Michael, K.: AI-based and digital mental health apps: balancing need and risk. IEEE Technol. Soc. Mag. **42**(1), 25–36 (2023)
13. Jameel, L., Valmaggia, L., Barnes, G., Cella, M.: mHealth technology to assess, monitor and treat daily functioning difficulties in people with severe mental illness: a systematic review. J. Psychiatric Res. **145**(2021), 35–49 (2022). https://doi.org/10.1016/j.jpsychires.2021.11.033
14. James, L.J., et al.: Evaluation of personalized treatment goals on engagement of smi patients with an mhealth app. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1568–1573. IEEE (2022)
15. Litvin, S., Saunders, R., Maier, M.A., Lüttke, S.: Gamification as an approach to improve resilience and reduce attrition in mobile mental health interventions: a randomized controlled trial. PLoS ONE **15**(9), e0237220 (2020)
16. Liu, B., et al.: Adversarial attacks on large language model-based system and mitigating strategies: a case study on ChatGPT. Secur. Commun. Netw. **2023** (2023)
17. Locke, E.A., Latham, G.P.: Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. Am. Psychol. **57**(9), 705 (2002)
18. Michie, S., et al.: The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann. Behav. Med. **46**(1), 81–95 (2013)
19. Organization, W.H., et al.: World mental health report: transforming mental health for all (2022)
20. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. **55**(1), 68 (2000)
21. Shatte, A.B., Hutchinson, D.M., Teague, S.J.: Machine learning in mental health: a scoping review of methods and applications. Psychol. Med. **49**(9), 1426–1448 (2019)
22. Svensson, B., Hansson, L., Markström, U., Lexén, A.: What matters when implementing flexible assertive community treatment in a Swedish healthcare context: a two-year implementation study. Int. J. Ment. Health **46**(4), 284–298 (2017)
23. Van Gorp, P., Nuijten, R.: 8-year evaluation of GameBus: status quo in aiming for an open access platform to prototype and test digital health apps. Proc. ACM Hum.-Comput. Interact. **7**(EICS), 1–24 (2023)
24. Van Veldhuizen, J.R.: FACT: a Dutch version of ACT. Community Ment. Health J. **43**(4), 421–433 (2007). https://doi.org/10.1007/s10597-007-9089-4