

A TRUSTworthy speech-based AI monitorING system for the prediction of relapse in individuals with schizophrenia
 GA N° 101080251
 Start Date: 01/07/2023 – Duration: 66 months
 Coordinator: Iris Sommer, UMCG

D3.2 – eHealth course

WP3 – Creating a trustworthy AI monitoring system based on the speech biomarker

Author (s):	Brita Elvevåg (UiT), Lars Ailo Bongo (UiT), Per Niklas Waaler (UiT), Musarrat Hussain (UiT), Igor Molchanov (UiT)
Reviewer(s):	Iris Sommer (UMCG)
Identifier:	D3.2_UiT_v1
Dissemination level:	Public
Date:	17 th June 2024
Number of pages:	16

Summary	This deliverable report summarizes the objectives of the eHealth course and the 2 versions that have been developed to motivate and empower participants, their families, and caretakers during the TRUSTING clinical trials to ensure consistent app usage over a twelve-month period, allowing for accurate evaluation of a relapse predictor.
----------------	--

Approved by the Coordinator	Date
Iris Sommer	Digitaal ondertekend door Iris Sommer Datum: 2024.07.10 15:21:27 +02'00'

Document history

Date	Change	By whom	Comments
20/6/2024		Brita Elvevåg (UiT), Lars Ailo Bongo (UiT), Per Niklas Waaler (UiT), Musarrat Hussain (UiT), Igor Molchanov (UiT)	First version

Table of Contents

1. Introduction	2
2. Presentation of Results.....	3
2.1 Web Application for eHealth course	3
2.2 Chatbot for delivering the eHealth course.....	9
2.3 Expanding the chatbots abilities with dynamic prompting methods	10
2.4 Controlling the scope of the chatbots responses.....	12
2.5 Ensuring Data Privacy	13
3. Critical Analysis of Results	14
4. Conclusion.....	16

1. Introduction

The purpose of this deliverable is an electronic educational course to educate, motivate and empower participants (as well as their families and caretakers/support systems) during the TRUSTING clinical trials to ensure that they use the app during the complete twelve month period and thus enable us to properly evaluate the accuracy and viability of the relapse predictor. This deliverable is concerned with the development and implementation in an electronic platform of the eHealth course. Further the core materials have been translated into six languages, namely English, Dutch, German, French, Czech and Turkish. However, after delivery of the prototype in June 2024, this course will be further developed in concert with the TRUSTING User Board and adapted for use across all testing sites. This is because although translation of English materials (that account for the majority of the materials that are both available and accessible to us) is straightforward, the resulting materials are not very culturally sensitive and new materials will in fact be developed also that are better suited.

The key goals of this course are to provide essential education about schizophrenia, what relapse is and include an explanation of the potential for false positives and errors in monitoring and diagnosis. Indeed, it is important to educate participants about the eventual interpretation of test results from TRUSTING and the risk of false alarms and their consequences.

The premise of the method of electronic delivery is that ‘adaptive education’ is better suited to those who are ill and who may have reduced cognitive function (e.g., reduced attention spans).

2. Presentation of Results

Prior to commencing work on this deliverable, we reached out to our TRUSTING consortium and to GAMIAN-Europe to establish what publicly available information brochures there were in the 6 TRUSTING languages for patients and the interested public, and also what 1-2 minute videos on these topics existed in these languages. The responses were relatively disappointing and very unequal across the languages, and as such our work on this deliverable is in response to this, namely to create some useful courses that would be accessible in all the 6 TRUSTING languages. Therefore, we have developed a two-part course. The *first* is a web app designed to provide information for families, caretakers, and supportive system. The *second* is a chatbot designed for patients with schizophrenia. We decided to split the course into two parts since we believe both the presentation style and the content of the course needs to be different for the two groups. We are however, currently researching ways of combining the two, either by using the chatbot as a way to learn more about the topic presented in each screen in the web app, or the screens as information to be shown in the chatbot. We therefore believe the best solution requires more research that is outside the scope of this first eHealth course deliverable.

2.1 Web Application for eHealth course

The objective of the eHealth course is to provide essential education about schizophrenia, what relapse is and include an explanation of the potential for false positives and errors in monitoring and diagnosis. The course is structured into three modules, as shown in Figure 1 below. Each module consists of multiple slides, each featuring a picture followed by explanations and a multiple-choice question as shown in Figure 2 (module 1 slides). The purpose of the question on each screen is to ensure users understand and reflect on the concepts presented in each slide. If a user provides wrong answer to the question the system pops-up user with a message saying, "You have selected an incorrect option." Users will not be able to proceed the course without providing a correct answer to the question on each screen. Upon correct answer the "Next" button of the screen gets enabled and allows users to process to the next screen. Additionally, each screen features an "Ask AI Chatbot" button. Currently, the button is disabled; however, we aim to connect each slide of the web application with the complementary chatbot that we also have developed. This will enable users to interact with an AI bot for necessary explanations and other dialogue (see section 2.2-2.4 for details). Since some screens may contain medical terminologies that could be difficult for non-experts to understand, the bot will help explain concepts and answer user questions related to various slides. All the contents in various screens are translated from English into five languages; German, French, Dutch, Czech, and Turkish, using machine translation. In the summer (July 2024) when the GAMIAN-user panel has been established, we aim to verify all the content by obtaining feedback from potential users as well as academic and clinical experts (from the TRUSTING consortium), and to confirm that the content is appropriate in all supported languages. Once it is deployed, this will be made available to the public via the web at: trusting-ehealthcourse.azurewebsites.net.

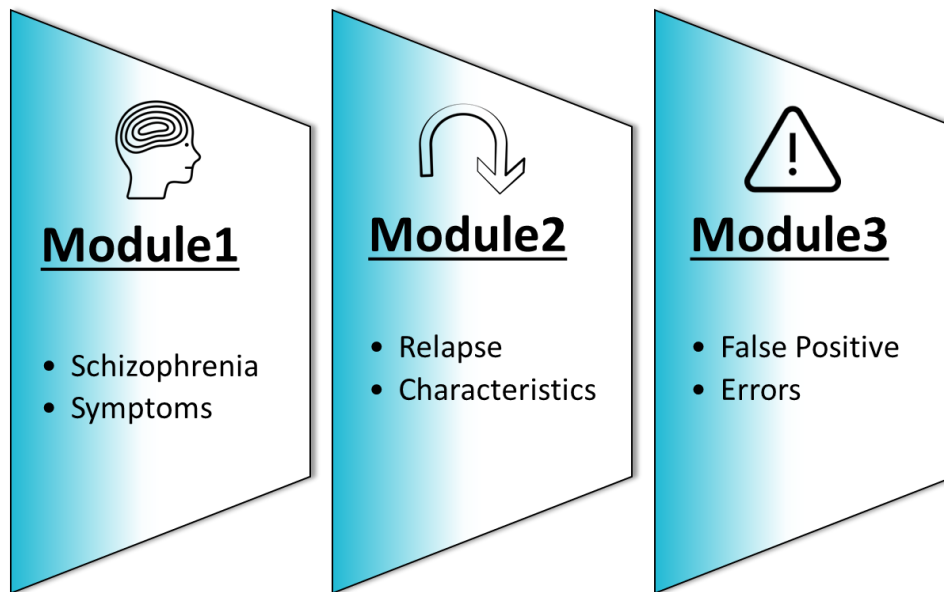


Figure 1: eHealth course web application overview

Module 1 of the course offers a detailed and interactive exploration of schizophrenia, covering its introduction, characteristics, and the various symptoms associated with the disorder. Through a combination of visual aids, descriptions, and interactive questions, users are provided with a comprehensive learning experience that equips them with a deeper understanding of schizophrenia and its impact. Specifically, Module 1 is designed to provide the essential and basic education about schizophrenia explaining that it is a long-term condition that results from changes in the brain and causes a range of different psychological symptoms, and importantly that it is nobody's fault. The module explains that it is a chronic and severe mental disorder that affects how a person thinks, feels, and behaves. Further, people with schizophrenia may seem like they have lost touch with reality, which can be distressing for them and for those around them.

As illustrated in Figure 2 below, Module 1 consists of a total of five screens, each focusing on a different aspect of schizophrenia. These screens include (i) an introduction to schizophrenia, (ii) characteristics of schizophrenia, (iii) positive symptoms, (iv) negative symptoms, and (v) cognitive symptoms associated with schizophrenia. As mentioned earlier, each screen features an image, a description explaining the associated concepts, and presents the user with a question with multiple choices. If the user chooses an incorrect answer they are told this and can choose another response. Upon selecting the correct answer, the user can switch screens to proceed with the course.

The introduction screen provides a comprehensive overview of what schizophrenia is, including its definition, prevalence, and impact on individuals and society. This screen sets the stage for a deeper understanding of the disorder and prepares users for the subsequent detailed discussions.

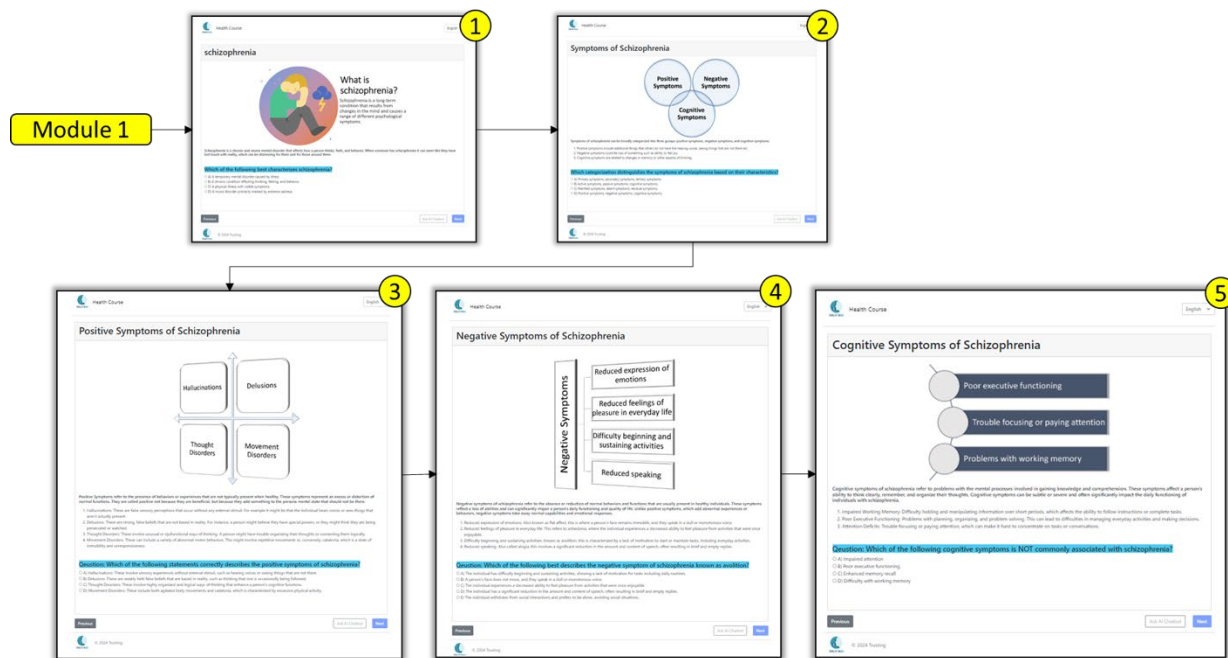


Figure 2: Module 1 screens workflow.

The second screen delves into the characteristics of schizophrenia, outlining the general traits and behaviours exhibited by individuals with the disorder. This includes information on how schizophrenia is diagnosed, common misconceptions, and the importance of early intervention and treatment.

The third screen focuses on positive symptoms of schizophrenia, which are those symptoms that add to a person's so-called normal experience, such as hallucinations, delusions, and thought disorders. It explains these symptoms in detail, providing examples and discussing how they affect an individual's perception of reality.

The fourth screen covers negative symptoms, which involve the absence or reduction of normal (or usual) behaviours and functions. These symptoms include lack of motivation, social withdrawal, and diminished emotional expression. This section helps users understand how negative symptoms can severely impact a person's ability to function in daily life.

The final screen addresses cognitive symptoms, which affect memory, attention, and executive function. These symptoms can make it difficult for individuals to process information and make decisions, further complicating their ability to manage everyday tasks.

Module 2 is designed to provide a basis understanding of relapse, from its characteristics and triggers to the warning signs and management strategies. By educating individuals on these aspects, the module aims to equip them with the knowledge and skills necessary to navigate the challenges of recovery and maintain their progress over time. Specifically, Module 2 is designed

to help user in understanding the concept of relapse, its characteristics and management. The course introduces relapse as the recurrence of an illness or disease after a period of improvement. This can happen in various mental health conditions, such as depression, anxiety disorders, bipolar disorder, schizophrenia, and other illnesses.

Similar to Module 1, this module also consists of five screens as shown in Figure 3 below (Module 2 Screen). These screens encompass various important aspects of relapse in order to provide a comprehensive understanding and effective strategies for managing it.

The first screen of module 2 serves as an introduction to the concept of relapse. This section provides a foundational understanding of what relapse is, defining it in the context of recovery from behavioural health challenges. It explains the cyclical nature of relapse and recovery, emphasizing that relapse is a common part of the recovery process. By to some extent normalizing relapse, this screen aims to reduce the stigma often associated with it, thereby encouraging individuals to seek help and support without fear of judgment.

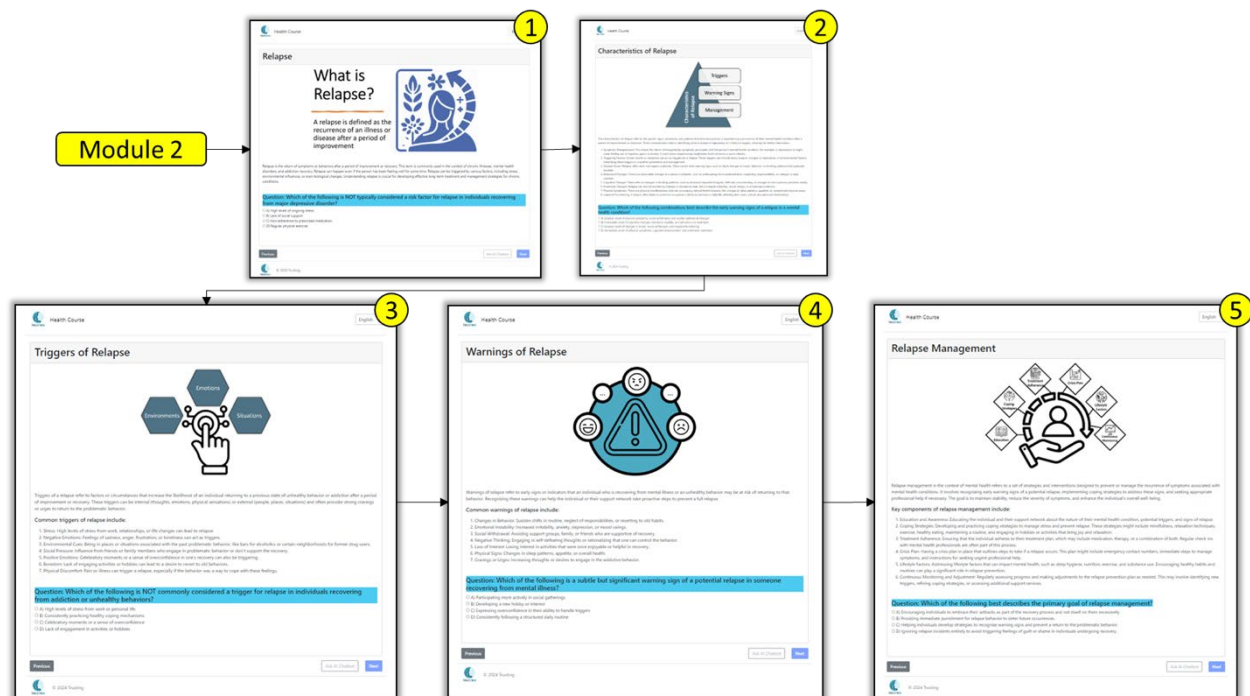


Figure 3: Module 2 screens workflow.

The second screen of module 2 delves into the characteristics of relapse. This part identifies the key signs and symptoms that might indicate an individual is experiencing a relapse. It covers both emotional and behavioural changes that can occur, such as increased stress, withdrawal from social activities, neglect of responsibilities, and a return to previous harmful behaviours.

Understanding these characteristics is crucial for early identification and intervention, enabling individuals and their support networks to take proactive steps to address the relapse.

The third screen of module 2 focuses on the triggers of relapse. It provides an in-depth exploration of various internal and external factors that can precipitate a relapse. Internal triggers might include emotional states like stress, anxiety, or depression, while external triggers could involve environments, people, or situations associated with past substance use or negative behaviours. By identifying these triggers, individuals can develop strategies to avoid or manage them, thereby reducing the likelihood of relapse.

The fourth screen of module 2 outlines the warning signs that precede a relapse. This section is crucial for early detection and prevention. It highlights specific behaviours and thought patterns that often emerge before a full-blown relapse occurs. Examples include rationalizing or justifying the use of substances, thinking about old habits, or planning situations where relapse is likely. Recognizing these warning signs early allows individuals to take preventive actions, such as reaching out for support or using coping mechanisms taught during recovery.

The final screen of module 2 is dedicated to the management of relapse. This screen offers practical strategies and tools for dealing with relapse if it occurs. It emphasizes the importance of having a relapse prevention plan in place, which includes steps like identifying supportive people to contact, using coping strategies learned in therapy, and engaging in self-care activities. The screen also discusses the role of ongoing therapy and support groups in maintaining long-term recovery and preventing future relapses.

Module 3 acknowledges the possibility of errors when monitoring for change (as will be the case in the RCT in TRUSTING). These errors fall into two main categories: The first type, known as a false positive, occurs when the diagnosis system incorrectly detects a disease that the user does not have. The second type, known as a false negative, happens when the system fails to detect a disease that the user actually has. Both types of errors can have various side effects and may impact the user in different ways. This module consists of only one screen as shown in the Figure 4 highlighting the types of errors with a concise description.

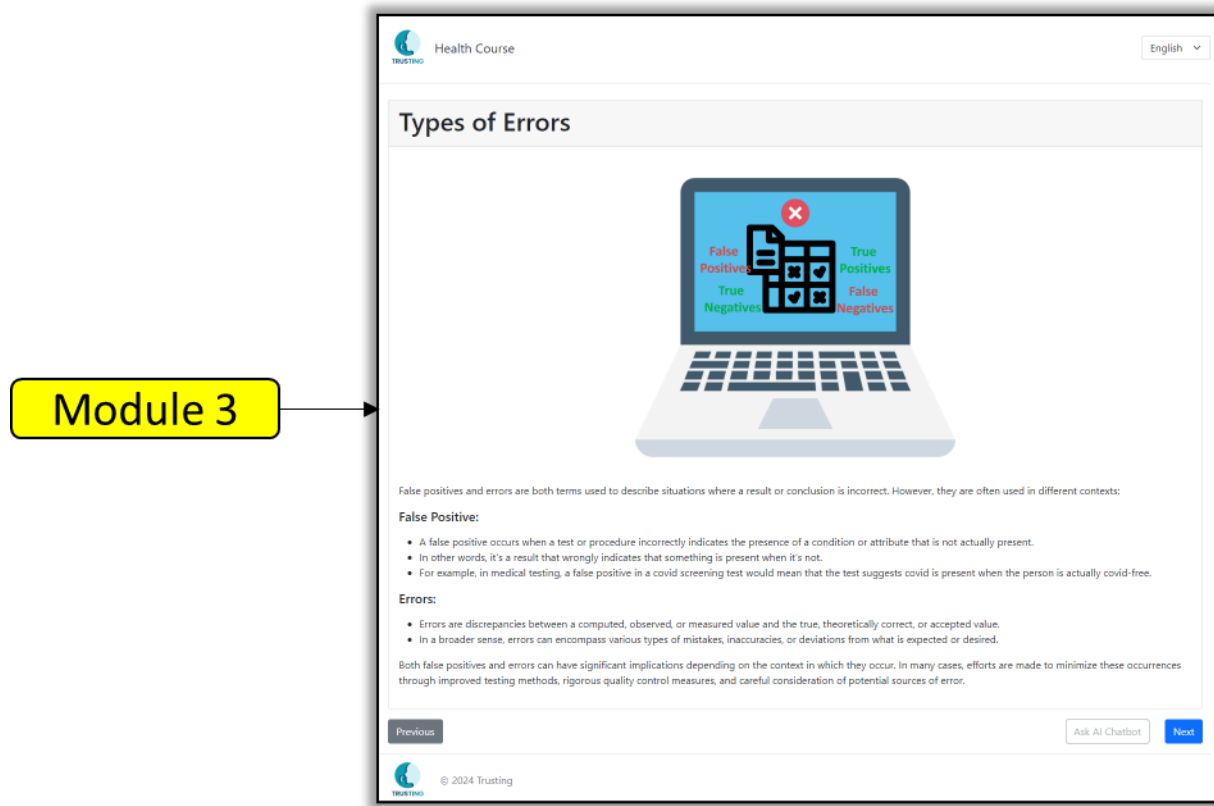


Figure 4: Module 3 screen overview.

To motivate users, we have included an appreciation screen at the end of each module, as shown in Figure 5. Each appreciation screen depicts the user's current status with a motivational message designed to encourage them to complete the course. This feature aims to provide positive reinforcement and a sense of accomplishment, which can be powerful motivators. By visually representing their progress and pairing it with an encouraging message, users are more likely to feel recognized for their efforts and inspired to continue their learning journey. This approach not only enhances user engagement but also helps in maintaining their motivation throughout the course.

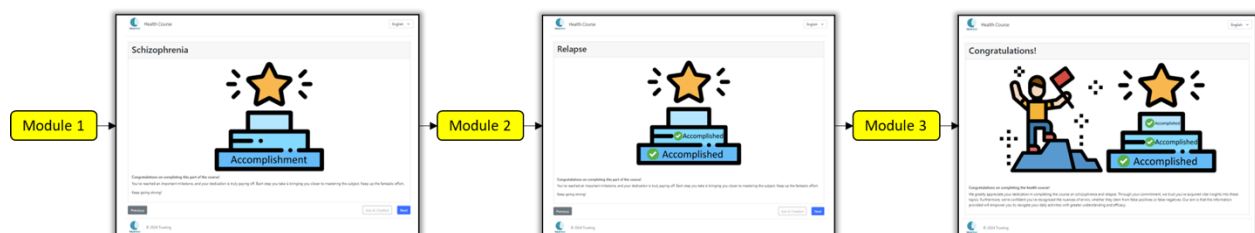


Figure 5: Appreciation screens among modules.

The information presented in the web app is based on numerous studies, including:

- Bereda G, Clinical symptoms of schizophrenia. *Psychiatry Ment. Health Res.* 2022; 4(2):1-3.
- Weinberger, Daniel R., and Paul Harrison, eds. *Schizophrenia*. John Wiley & Sons, 2011.
- Carbon, Maren, and Christoph U. Correll. "Thinking and acting beyond the positive: the role of the cognitive and negative symptoms in schizophrenia." *CNS Spectrums* 19.S1 (2014): 35-53.
- Addington, Jean, D. Addington, and Eleanor Maticka-Tyndale. "Cognitive functioning and positive and negative symptoms in schizophrenia." *Schizophrenia Research* 5.2 (1991): 123-134.
- Brandon, Thomas H., Jennifer Irvin Vidrine, and Erika B. Litvin. "Relapse and relapse prevention." *Annu. Rev. Clin. Psychol.* 3 (2007): 257-284.
- Saunders, Bill, and Steven Allsop. "Relapse: a psychological perspective." *British Journal of Addiction* 82.4 (1987): 417-429.
- Allan, Stephanie, et al. "Perspectives of patients, carers and mental health staff on early warning signs of relapse in psychosis: a qualitative investigation." *BJPsych Open* 6.1 (2020): e3.

2.2 Chatbot for delivering the eHealth course

The broad range of cognitive abilities in individuals with schizophrenia makes individual adaptation especially beneficial in an educational setting. With the tremendous advances in large language models (LLMs) over recent years, chatbots are therefore a natural choice as a vehicle for delivering information; they can interpret concepts or words the user struggles with, provide examples and analogies, translate technical language into layman's terms, and adapt the pace of information delivery to the preferences of the individual. Because of these communication advantages, as an alternative to the main eHealth course presented in 2.1, we developed a chatbot that interactively allows the user (patient) to learn more about their condition. To develop this chatbot we used prompt engineering - a technique whereby we instruct the chatbot in natural language on how to behave and what information to convey. The chatbot breaks the information of the course up into bite-sized pieces of information, each of which makes up the content of a single chatbot message. The chatbot pauses the delivery of information at predetermined points to test the users understanding of the material by asking them a few multiple-choice questions which the user must answer correctly to proceed. If the user answers incorrectly, the chatbot will inform them that their answer is incorrect, but also explain why. When they answer correctly, it moves on to the next topic. To ensure that the chatbot always complies with rules such as *"do not move on until the user has answered correctly"* we plan to utilize AI-monitors that monitor the chatbots responses and ensure that the rules are being followed – see 2.4 and Figure 8 for more details. To avoid overwhelming the user with too much information at once, we have also implemented a feedback system that controls the length of the

chatbot messages by either warning the chatbot when they start getting long, or by summarizing them when they exceed a maximum length.

We plan to investigate alternative ways to integrate chatbot support into the web-application course described in 2.1, either as an alternative way of taking the course (e.g. the user may select, in the first screen, between the chatbot version or the regular version of the course) or as a compliment to it (e.g. it answers questions that the user might have about the information on the current slide). The decision on how to integrate the chatbot version of the eHealth course will be based on feedback from members of the TRUSTING consortium. A test-version of the chatbot for delivering the eHealth course is available at <https://python-test-ifi.azurewebsites.net/>.

2.3 Expanding the chatbots abilities with dynamic prompting methods

The chatbot has been instructed to follow specific syntax conventions that allows backend scripts to detect and execute various commands. This setup allows them to interact with backend files and scripts, and, for example, enables it to present images (that we have designed) that illustrates the concepts and themes of the information presented (see Figure 6). It can also request information on a need-to-know basis, which gives it access to a larger body of scripted information than would be possible to convey via non-dynamic prompting techniques – see Figure 7 for a graphical overview of how the chatbot dynamically accesses information during a conversation. Having developed a programming framework that allows the chatbot to interact with backend files makes it easy to implement new features, such as systematically tracking the users current progress in the eHealth course.

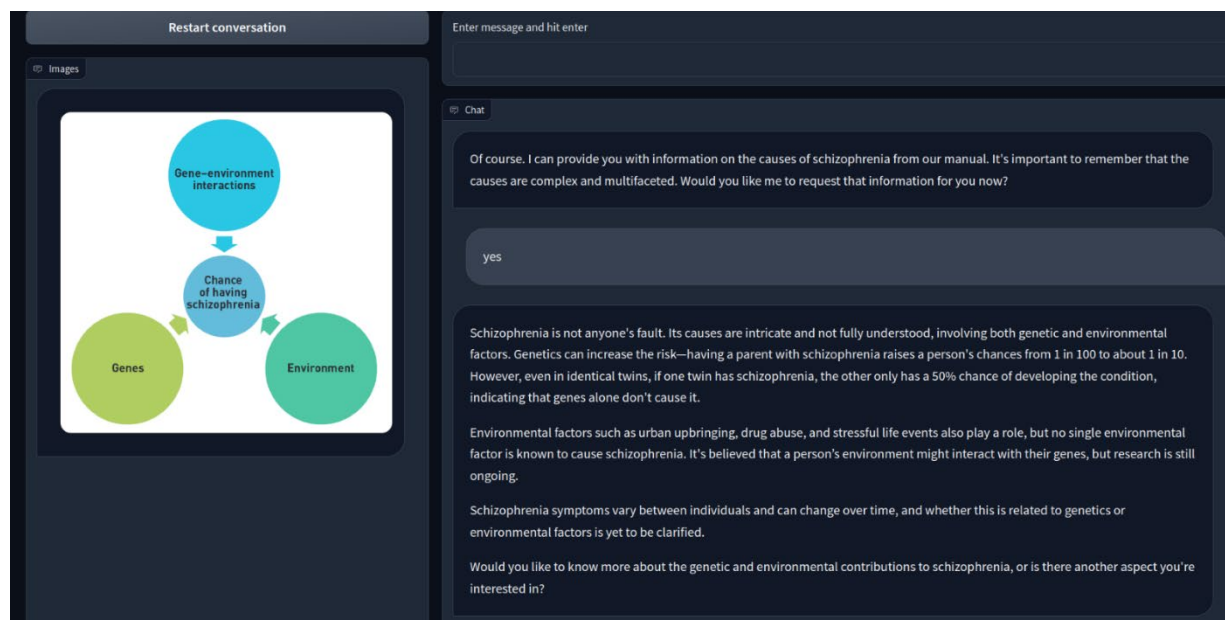


Figure 6 - CHAT EXAMPLE: Example of the chatbot fetching information to answer the users question about schizophrenia and presenting a relevant figure to accompany its response.

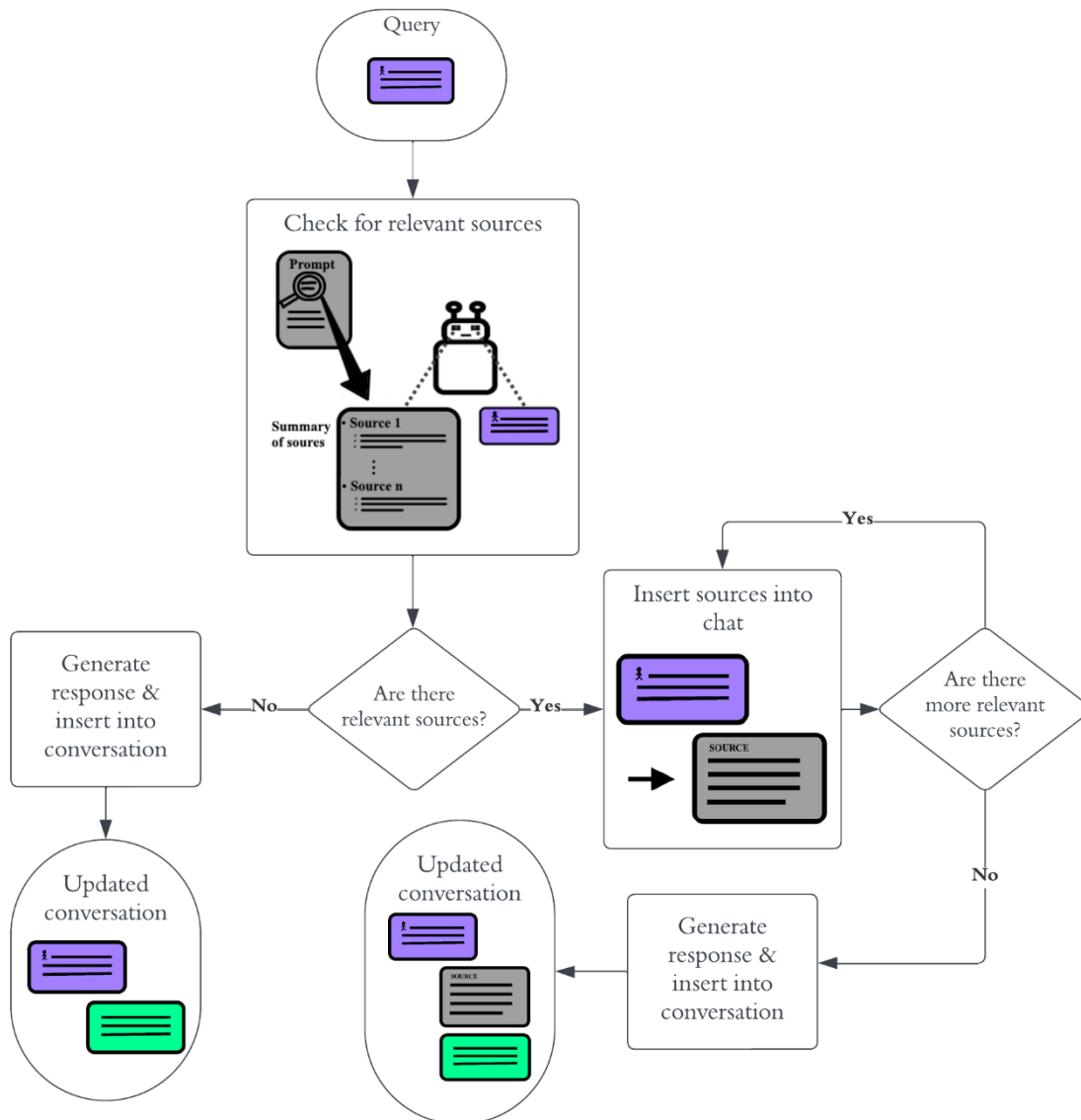


Figure 7 - REQUESTING KNOWLEDGE: Flowchart illustrating how the chatbot retrieves information from its resources and uses it to answer the users' query.

2.4 Controlling the scope of the chatbots responses

Since the chatbot is conveying sensitive information to users who may be in a vulnerable state, it is important to ensure that we can control the scope of the chatbots responses and behaviour. Specifically, we want it to use it as an education tool only and prevent it from taking on other roles, such as a personal therapist, or generating answers based on GPTs innate (un-scripted) knowledge. To ensure that the chatbot adheres to our instructions, is transparent about its limitations, and stays within the scope of its role, we have implemented an AI-filter that monitors the chatbots messages and provides it with corrective feedback if any undesired behaviours are detected (see Figure 8 - AI FILTERS). The AI-filter is composed of several prompt-engineered LLMs, each of which is tasked with detecting a specific type of rule-violation. For example, one AI-monitor is responsible for evaluating the fidelity with which the chatbot conveys scripted resources, and another is responsible for checking that the chatbot does not stray outside the scope of its role (for instance by giving unscripted advice on medication use or taking on the role of a therapist). Distributing these tasks across multiple agents allows them to become more specialised in their function so that each type of undesired behaviour can be reliably detected and (if necessary) corrected for at an early stage. The aim of this approach, in broad terms, is to make it possible to leverage the educational advantages of modern LLMs in an ethically responsible manner by ensuring that their responses are firmly anchored in source-materials that have been validated by relevant experts. Preliminary results suggest that this method is highly effective at anchoring the chatbot its instructions and sources, and substantially increasing the chatbots reliability.

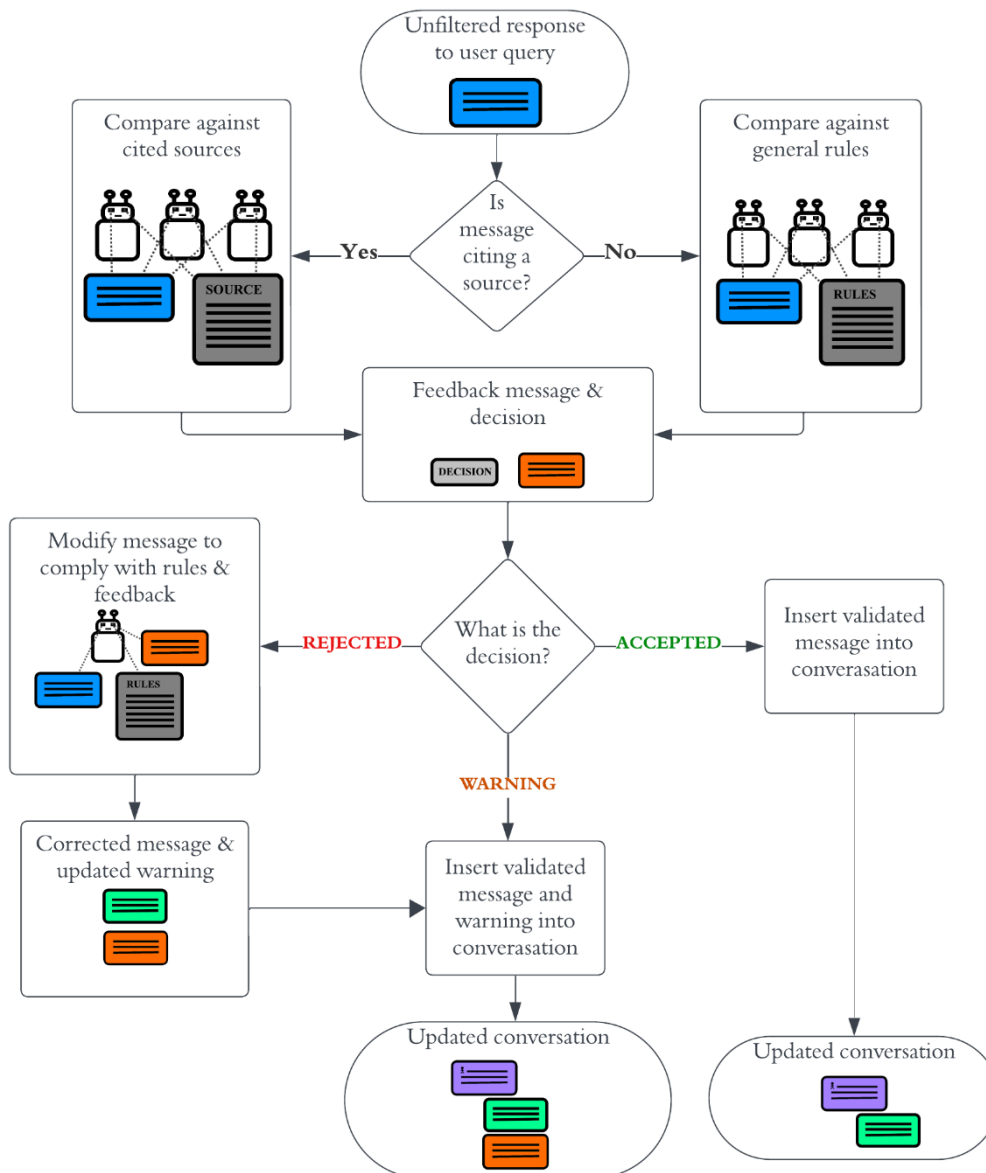


Figure 8 - AI FILTERS: Flowchart showing how AI-agents monitor and influence the conversation in the background to ensure that the chatbot is faithful to its source materials and complies with scripted instructions (detailed in the prompt).

2.5 Ensuring Data Privacy

Given that our chatbot will be used by mental health patients who share sensitive information, ensuring data security and privacy is paramount. Currently, we do not plan to store any data from

the users. To ensure that no data is being shared with third parties, we utilize OpenAI models through Microsoft Azure, leveraging its advanced security features such as encryption, access controls, and compliance with industry security standards. Thus, the data flows only through controlled channels, providing the highest level of data protection and privacy. Although we have designed the chatbot for sensitive data, we note that the current version does not yet have all data protection mechanisms enabled and so will not and should not be used for sensitive data. During development we may log conversations for debugging and evaluation purposes and note that Microsoft logs chatbot input and output for threat protection reasons. We do however have full control over the dataflow in the chatbot and so in the eventual deployed version that will be used by TRUSTING users, we plan to disable the conversation logging and we will request Microsoft to also disable their logging. The chatbot will therefore use the same mechanism and policies as chat.uit.no, which is approved for sensitive data by UiT. In addition, the above described protection mechanisms should ensure that the chatbot stays on topic and is not used for anything else than for the intended purpose of providing educational information.

3. Critical Analysis of Results

We have successfully delivered the eHealth course. However, there remains much scope for improvement including areas for future research. Below we note some key challenges ahead that we plan to address and solve:

(i) The current version of the eHealth courses will be rigorously evaluated by TRUSTING consortium members and clinicians and then - once it is established - by GAMIAN user panel who are fluent in all 6 of the TRUSTING languages. The chatbots developed so far are prompted in English. Despite of this, they seem to translate very well to other languages, and they change language appropriately (despite no instructions to do so) if the user starts the conversation in a different language. Going forwards, we plan to test the chatbot translations more systematically by getting feedback from users in each language. We will also systematically test if conversing with the chatbot in non-English languages influences various aspects of its reliability and performance. Is the performance of the AI filters affected? Does it reliably request images and sources that are relevant to the context?

(ii) Educational materials presented in the form of an interactive web app or chatbot transcends the mere digitalization of glossy materials and potentially increases engagement for those with smaller attention spans than the healthy population. As such, the dynamic forms of delivering information that we employ may be considered an important improvement over and above the current *status quo*. However, the flexibility of chatbots also comes with greater risks due to the vast range of responses that the chatbots can potentially generate. We aim to harness the benefits of modern LLMs while minimizing the risks of harmful, misleading, or inaccurate responses by anchoring their responses in scripted materials and enforcing strict compliance with instructions and source-materials using AI-filters. Within the next 12 months, we plan to test

further these safety-mechanisms by gathering feedback from human testers. In addition to human testers, we will use AI (prompted LLM agents) to simulate conversations from a multitude of scenarios. The benefit of using AI testers in addition to human testers is that we can automatically generate large numbers of simulated conversations in multiple scenarios. Repeated sampling increases the chance of catching rare errors or behaviours we have not accounted for. In addition, by running multiple independent simulations of a particular situation (for example, a user who trying to push the chatbot “out-of-character”), we can leverage the stochastic nature of LLMs to get more samples and more variety in the input of the chatbot. Results from these experiments can be statistically analysed so that we can discern between genuine improvement and chance improvement; a distinction which is particularly important when working with prompt engineering. We have already used this method to identify situations in which the chatbots sometimes behaves in an undesirable manner, and fine-tuned the AI-filters and prompts to improve chatbot behaviours in those situations.

(iii) We plan to design web-applications on Azure specifically for the purpose of collecting feedback so that we can easily share the deployments (chatbots and web app at trusting-ehealthcourse.azurewebsites.net) with testers. To streamline the process of collecting feedback and test-results (such as conversations with the chatbot), we plan to incorporate features such as a “Submit feedback” button and an associated textbox where the testers can enter and submit any comments they might have at any point during the session. To facilitate constructive feedback that we can act upon, we will include a “Feedback wiki” page which presents the tester with information that details what we want feedback on. When the testers click the “Submit” button after writing a comment, their comment (e.g., “This *information is inaccurate*”) will automatically be saved to a feedback-database along with relevant “snapshot” data (e.g., the most recently generated chatbot message) that gives context to the comment. With this approach, feedback can be collected and organized automatically based on language and which deployment is being tested (which chatbot/version/web-applications). Upon starting each new test session, a new directory will be automatically created which stores results and feedback. Using this approach, the process of generating feedback is contained within the web-applications. This simplifies collection and analysis of feedback and makes it a more scalable process.

Some of the aspects of the chatbot we will include in the “Feedback wiki” are

- Does the chatbot produce responses that are relevant, useful, and specific to the context and questions posed (i.e., not generic)?
- Does its answer seem complete, or is there information missing which should be presented in this context?
- Are its responses aligned with the current gold-standards for schizophrenia information and clinical care?
- Does it respect the boundaries of its role?
- Does it ask appropriate questions to figure out what information is relevant before fetching it, or does it start presenting “walls of text” prematurely?
- Does it hallucinate or produce factually incorrect or inappropriate information?

- Any strange behaviours (e.g. referring to an image it has not shown)?

Once this kind of feedback has been collected, we can fine-tune those aspects of the chatbot to ensure it reliably complies with the desired behavioural rules, such as avoiding lengthy generic responses or staying within the bounds of its source-materials and instructions.

(iv) Where do we actually want this course to be in terms of the experience of the TRUSTING app? This issue will be discussed with the TRUSTING consortium. Put differently, should this be on the TRUSTING website as a standalone course or as something users are to do before / during the RCT of the TRUSTING app? Currently, the eHealth course and chatbots are a standalone web-apps. However, our plan is to integrate the eHealth web app into the TRUSTING app, specifically as a mandatory part of the introduction to the TRUSTING app (after setup but before arriving at the home screen). Similarly, we plan for the chatbot to be integrated into the eHealth course as an alternative that the user can select on the first screen of the eHealth course. Considering the study participants, it is advantageous to integrate all features into one digital platform to produce a minimally distracting experience, and, to the degree it is possible, avoid linking to external resources.

4. Conclusion

We have developed two versions of the eHealth course: one as a web application, which presents information on various screens, covering one small topic per screen, and another conducted in a chatbot environment. The eHealth web app consists of three modules: Module 1 focuses on schizophrenia, aiming to provide a basic understanding of the condition: what it is, its symptoms, and its causes (emphasizing that it is no one's fault). Module 2 educates users on the concept of relapse, including its characteristics, early signs, triggers (especially stress), and how to manage a relapse. Finally, Module 3 explains basic probabilistic and statistical concepts, primarily false positives and prediction errors, to help users understand that there might be errors in the diagnosis and management of the illness in real-world scenarios. In the chatbot version of the course, the user is guided through the Modules in a similar slide-by-slide manner, with each chatbot response typically consisting of text plus image or a multiple-choice question. We have developed a framework that allows the chatbot to interact with backend functions and files which expands its abilities (e.g., presenting images or requesting information). This version of the course is particularly suitable for users with cognitive impairments, since it is highly flexible by virtue of being powered by GPT-4 and can adapt its style of information delivery to user preference. Both versions of the course have been developed to be understandable by users with a wide range of cognitive abilities and with the potential cognitive impairments of schizophrenia patients in mind; they deliver information in small digestible pieces and offer a more interactive and engaging experience compared to brochures by frequently asking questions.