informatics

MDPI

Technical Note

GOBLIN: Prototype Architecture and Prospective Validation Roadmap for Clinician-Supervised Management of Auditory Verbal Hallucinations

Zaki Milhem [1,2,†], Rareș-Codrin Lache [1,†], Maria-Luiza Vlădărean [3,*], Ioan-Andrei Bârsan [4], Radu-George Bolborici [5], Alexandru-Florin Ene [5], Ana-Antonia Nicolae [5], Mara-Simina Olinic [6], Alexandru-George Stan [7], Emanuela-Gabriela Hărăguș-Hațegan [6] and Roland Stretea [2,6,†]

1 GOBLIN AI SRL, 500003 Brasov, Romania; milhem.zaki@umfcluj.ro (Z.M.)
2 Department of Neurosciences, Psychiatry and Pediatric Psychiatry, Faculty of Medicine, Iuliu Hațieganu University of Medicine and Pharmacy, 400012 Cluj-Napoca, Romania; roland.miha.stretea@elearn.umfcluj.ro
3 Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, 85748 Munich, Germany
4 Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada; andrei.ioan.barsan@gmail.com
5 Campus Research Institute, Politehnica University Bucharest, 060042 Bucharest, Romania; radu.bolborici@upb.ro (R.-G.B.); alexandru.ene2702@upb.ro (A.-F.E.); ana_antonia.neacsu@upb.ro (A.-A.N.)
6 Clinical Hospital for Infectious Diseases, 400348 Cluj-Napoca, Romania; mara.simi.olinic@elearn.umfcluj.ro (M.-S.O.); emanuela.gabr.hategan@elearn.umfcluj.ro (E.-G.H.-H.)
7 Clinical Emergency Hospital for Children, 400370 Cluj-Napoca, Romania; alexandru.geor.stan@elearn.umfcluj.ro
* Correspondence: maria.vladarean@tum.de
† These authors contributed equally to this work.

Abstract

Auditory verbal hallucinations are one of the most common symptoms in schizophrenia, contributing substantially to distress, functional impairment, and relapse risk. Existing therapeutic approaches, including pharmacological treatment and structured psychotherapies such as cognitive behavioral therapy for psychosis and AVATAR therapy, show promising clinical outcomes. However, scalability remains limited due to resource intensity and reliance on specialized personnel. Against this background, we describe the rationale, prototype architecture, and prospective validation roadmap for GOBLIN—a digital platform prototype combining personalized voice conversion, webcam-based eye-tracking, and large-language-model-assisted clinical documentation and decision support. The platform is intended to support clinician-supervised therapeutic workflows and between-visit monitoring following prospective validation. GOBLIN is intended to investigate whether an integrated digital workflow could address key limitations of current care models in schizophrenia and psychotic disorders, including fragmented follow-up, limited responsiveness to rapid symptom changes, and lack of scalable intervention tools. Our conceptual prototype is the first step toward leveraging the potential of digital, AI-driven platforms to complement existing treatment modalities and improve long-term management of psychosis-related symptoms. The system's modules are intended for clinician-supervised research workflows and remain investigational. This manuscript does not report completed clinical validation, therapeutic efficacy, diagnostic performance, relapse-prediction performance, or validated safety outcomes. Such capabilities remain to be evaluated prospectively

Academic Editors: Ashkan Ebadi, Ricky Leung, Chi Ho Li

Received: 23 July 2026
Revised: 5 October 2026
Accepted: 14 October 2026
Published: 27 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

through staged technical, usability, safety, and clinical studies before clinical implementation can be considered.

Keywords: auditory verbal hallucinations; schizophrenia; AI-based therapy; eye-tracking; AVATAR therapy; longitudinal monitoring; digital therapeutics; voice-based intervention

1. Introduction

Auditory hallucinations (AHs) are perceptual experiences occurring in the absence of a corresponding external auditory stimulus. They encompass a heterogeneous range of phenomena, including non-verbal sounds, music and, most prominently in psychiatric research, auditory verbal hallucinations (AVHs), commonly described as "hearing voices" [1]. Although traditionally associated with psychotic disorders, AVHs are a transdiagnostic phenomenon observed in a variety of psychiatric (mood disorders, borderline personality disorder, trauma- and dissociation-related conditions), neurological and medical conditions (substance-induced states, and neurological disorders such as epilepsy, Parkinson's disease and dementia), as well as in otherwise healthy individuals [1–3].

The clinical significance of AVHs therefore appears to depend not only on their presence, but also on their phenomenological characteristics and on the individual's interpretation of and relationship with the experience [4]. Voices vary substantially in frequency, duration, perceived location, identity, emotional valence, linguistic complexity, controllability and personal meaning. Comparisons between clinical and non-clinical voice-hearers indicate that distress, threatening or negative content, perceived power of the voice, maladaptive appraisals and coping difficulties may be particularly relevant in determining the need for clinical care [3,4].

This issue is notably important in schizophrenia-spectrum disorders, in which auditory hallucinations constitute one of the most frequent positive symptoms. Schizophrenia is a psychotic disorder characterized by delusions, hallucinations, disorganized thinking, and negative symptoms, with substantial heterogeneity in illness course and functional outcome [5]. Auditory verbal hallucinations are experienced by approximately 70% of people with schizophrenia [6], while auditory hallucinations more broadly occur in approximately 60–80% of patients with schizophrenia-spectrum disorders [7]. In this population, voices may be recurrent or persistent and can be associated with substantial distress, impaired functioning and reduced quality of life [8]. Although antipsychotic medication remains the principal pharmacological treatment for schizophrenia and can reduce hallucination severity in many patients, a clinically important subgroup continues to experience persistent or insufficiently responsive hallucinations despite adequate treatment [9,10]. Approximately 25% of people with psychotic disorders have been reported to experience persistent AVHs despite treatment [11].

Beyond incomplete symptom control, schizophrenia care is also constrained by monitoring burden, adverse effects, and fragmented follow-up [12,13]. Psychological interventions such as CBT for psychosis are recommended as adjunctive treatments, although their benefits are generally modest [14,15]. Further limitations include the resource-intensive nature of specialized psychotherapies, which require trained clinicians and sustained patient engagement [15], and the largely episodic, interview-based nature of psychiatric monitoring, which may prevent detection of clinically relevant deterioration between appointments [16].

https://doi.org/10.3390/informatics13110180

Two emerging adjunctive directions appear especially relevant to current gaps in schizophrenia care. The first is the development of more scalable, symptom-focused interventions for persistent AVHs. The second is the use of brief, objective monitoring tools between visits, aimed at detecting early change before clinical deterioration becomes obvious. Both respond to practical constraints of routine care: uneven access to specialized non-pharmacological interventions, and intermittent, encounter-based follow-up that remains largely limited to clinicians' subjective assessment [17].

Regarding the former direction and distressing AVHs in particular, the therapeutic target is not only the presence of the voice itself, but the patient's relationship to it. A growing body of relational therapies addresses this relationship directly: for example, Relating Therapy targets subordinate responding to voices, and Talking with Voices proposes dialogical interventions for distressing voice-hearing experiences [18–20]. Work on command hallucinations has shown that distress and compliance are strongly shaped by beliefs about the voice's power, authority, and ability to harm, which makes these beliefs an important mechanism for intervention [15,21]. AVATAR therapy extends this principle through therapist-guided dialogue with a digital representation of the voice, aiming to reduce its perceived dominance and increase the patient's sense of control [22]. Early randomized studies, including the recent AVATAR2 trial, support its potential to reduce voice-related distress and severity, particularly when delivered in extended formats [23,24], while related virtual reality approaches have shown similar promise for patients with refractory auditory verbal hallucinations [25,26]. Importantly, current evidence does not establish a consistent advantage of full visual embodiment over other active psychological approaches, suggesting that the therapeutic value may lie primarily in the structured dialogue and modification of the patient–voice relationship rather than in visual realism itself [18,19].

The latter direction focuses on non-invasive oculomotor measures as exploratory markers of cognitive and sensorimotor dysfunction in schizophrenia spectrum disorders. Eye-tracking is particularly promising in this context, as diagnosed patients show consistent abnormalities across several oculomotor parameters. These findings are well-documented across multiple independent cohorts and meta-analytic reviews [27,28], suggesting that eye movements may offer a measurable window into processes such as attention, inhibitory control, and sensory integration. Although oculomotor abnormalities have predominantly been characterized as trait-like endophenotypic markers, preliminary evidence suggests that some parameters may reflect clinical state changes, warranting further investigation of their utility for longitudinal monitoring [29]. Their ability to identify clinically meaningful deterioration at the individual patient level is yet to be established.

Taken together, these two lines of research provide the rationale for investigating a combined approach, targeting both distress reduction in persistent AVHs and the improved characterization of within-participant deterioration between visits. To this end, we propose GOBLIN, an integrated digital prototype intended to ease the joint exploration of these research directions. The proposed architecture combines webcam-based oculomotor measurement, AI-facilitated voice-avatar interaction, and clinician oversight in a single platform. The Voice component is informed by relational principles from AVATAR therapy, while the Sight component evaluates oculomotor variables as exploratory candidate measures. GOBLIN is intended to investigate whether clinician-mediated voice interaction may reduce AVH-related distress and whether repeated oculomotor measurements may provide clinically interpretable information about within-participant change between visits. The present manuscript does not establish therapeutic benefit, biomarker validity, psychosis classification, relapse detection or prediction, or clinical effectiveness—these remain subjects for prospective validation.

https://doi.org/10.3390/informatics13110180

The remainder of this technical note details the rationale and technical design of the GOBLIN prototype, and describes the prospective validation methodology required to establish its technical reliability, acceptability, clinical interpretability and therapeutic value. In Section 2 we give a detailed technical description of the system, covering the webcam-based oculomotor measurement module, the AI-facilitated avatar interaction module, and the digital medical assistant module. Section 3 then details the staged methodological roadmap for prospective validation. Section 4 discusses the platform's potential clinical utility, current limitations, and directions for future empirical validation, and Section 5 presents the conclusions.

2. Technical Description of the System

GOBLIN is conceived as an integrated digital platform that brings together three functions within a single clinically oriented system:

- Webcam-based oculomotor measurement;
- Clinician-mediated AVATAR therapy-inspired voice interaction;
- LLM-assisted clinical documentation and clinician-facing information support.

The central design principle of the system is integration under clinical supervision. Its purpose is to combine the three bounded functions within a single workflow: oculomotor measurements, clinician-mediated voice-avatar interaction, and structured clinical documentation support. At present, these functions are detailed as an integrated prototype architecture requiring formal laboratory validation before clinical acceptability or efficacy can be assessed. In translational terms, the concept sits at the level of system integration, where each component is grounded in an existing line of research and their coordinated use for psychosis care remains to be prospectively validated. Throughout this manuscript, references to monitoring, risk flagging, clinical support, and intervention describe intended clinician-supervised functions pending prospective validation. They do not denote validated continuous clinical surveillance, biomarker-based patient classification, relapse detection or prediction, autonomous adaptation, or demonstrated therapeutic scalability.

2.1. A Bird's Eye View of the System Architecture

The platform (Figure 1) integrates multiple modules designed to support the highlighted therapeutic objectives. These modules are accessed through a unified web-based interface supporting both clinician and patient interactions. User access and interaction pathways are determined through backend-managed authentication and role-based access control (RBAC), which define the functionality and workflow available to each user type. System requests initiated through the interface are routed to the appropriate functional module via the backend server layer. The server orchestrates communication between the interface and the relevant processing component, returning the generated output for presentation to the user, based on the specific interaction context.

The following section describes the functionality of the main components providing therapeutic support. As the architecture shown in Figure 1 follows standard client-server design principles, a detailed description of its underlying technologies is omitted. Going forward, we focus on the AI Core components, which represent the main contribution of this work.

https://doi.org/10.3390/informatics13110180

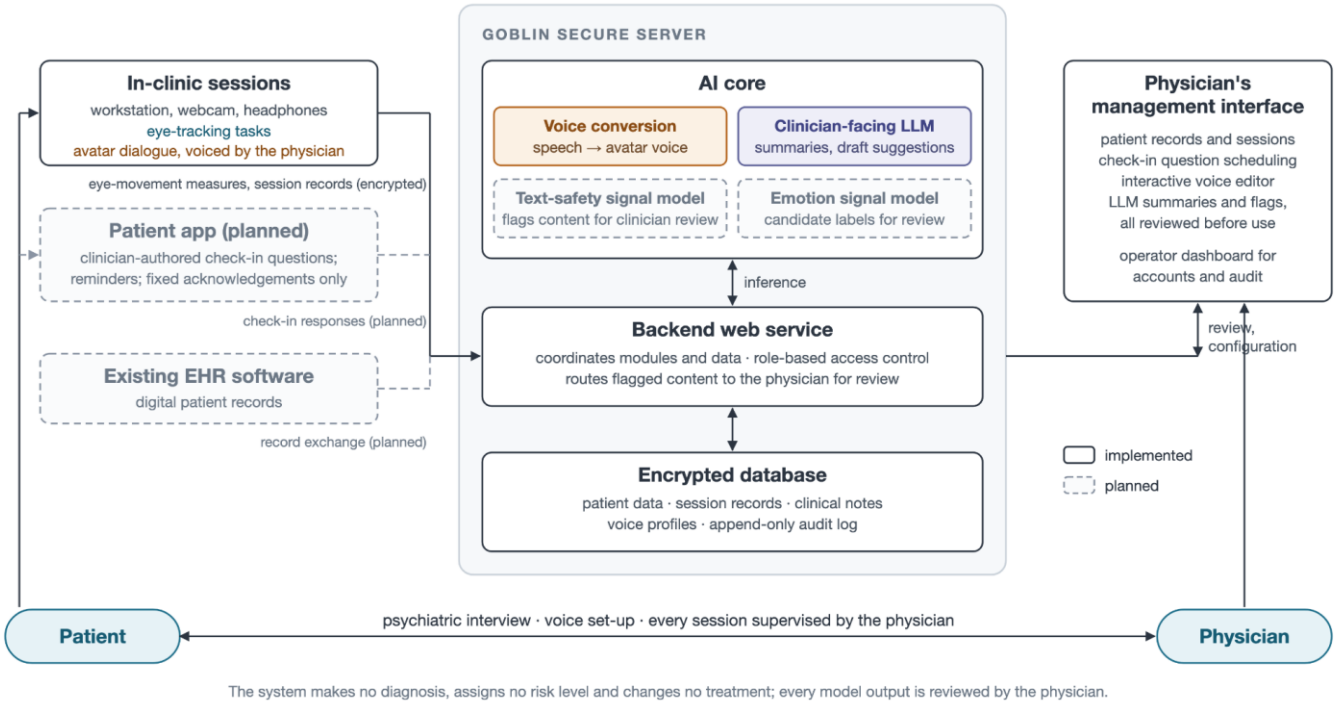


Figure 1. System overview.

2.2. GOBLIN Voice

2.2.1. Clinical Rationale

GOBLIN Voice is an investigational digital component that converts the clinician's speech using a personalized voice profile intended to approximate the patient's distressing hallucinated voice. Within the proposed GOBLIN Therapy workflow, the clinician retains dual communicative agency, alternating between the avatar's voice and their own during structured dialogue to guide, reassure, or reorient the patient and address the voice's perceived power. The current implementation and audio-routing arrangements are described in Section 2.2.2.

AVATAR therapy, as originally proposed by Leff et al. [30,31], uses digital voice synthesis and conversion. Within this framework, the patient selects the voice that best matches their dominant hallucinated voice from digitally morphed variations of the therapist's voice. As a rule of thumb adopted in the literature, a self-reported similarity score of approximately 70% is considered sufficient for this purpose [24]. The therapist's speech is then converted into the selected voice and presented through the avatar in real time, enabling dialogue that externalizes the hallucinated voice into an addressable, interactive entity. The therapist can switch as needed to their natural voice to provide guidance or support.

The phenomenology of AVHs informs requirements for voice personalization. In AVATAR therapy studies, the majority (61.2–75.0%) of patients experience voices with a higher degree of characterization, with identity linked to perceived age, gender, vocal qualities, accent and social background indicators [24,32]. These traits have direct or indirect acoustic correlates [31], motivating voice controls beyond the basic transformations proposed by Garety et al. [24]. Consequently, the practical meaning of this similarity criterion may vary with voice characterization, with highly characterized voices demanding more precise and expressive modeling.

A sense of voice presence during AVATAR sessions has been associated with improvements in hallucination frequency and severity alongside reductions in anxiety

[33]. Perceived environmental "realness" contributes to presence [34], motivating attention to acoustic realism in GOBLIN. In our use case, this perceived authenticity encompasses the collection of characteristics that define the hallucinatory voice, thus highlighting the importance of a "real-enough" synthetic representation. Glenthøj et al. [35] also observed a trend-level association between higher auditory fidelity and frequency of AVHs at follow-up. However, increasing the fidelity of the hallucinated voice must be balanced against the cognitive and emotional burden of the setup process. Ward et al. [22] note that approximately 5% of patients ($n = 75$) from the single-blind trial of Craig et al. [23] withdrew during the preliminary phase, with reasons including distress during avatar creation, reflecting the exposure-like nature of the intervention.

Ward et al. [32] report that 65% of dominant hallucinated voices are perceived as male. Together with the predominantly female psychiatric workforce in our initial target population [36], this suggests that cross-gender voice conversion may be common. Such scenarios involve larger mismatches in fundamental frequency (F0), spectral envelope, and speaker embeddings, which are associated with reduced conversion performance compared to intra-gender settings. These mismatches may limit perceptual realism, motivating targeted conversion methods.

These requirements inform GOBLIN Voice's clinician-operated controls for pitch, vocal tract size, spectral tilt and temporal roughness, supplemented by additional acoustic transformations and speaker-embedding adjustments. Further development aims to improve robustness in cross-gender conversion. The current implementation and planned extensions are described in the following section

2.2.2. Implementation Details

This comprises three distinguishable parts:

1. A graphical module with distinct patient- and clinician-facing interfaces, implementing the vocal parameter controls and the voice mask switch;
2. A voice tuning component for the iterative refinement of a baseline voice toward the target hallucinated voice;
3. A voice conversion (VC) component used by the clinician to morph their voice into the reconstructed hallucinated voice.

Since the first component uses standard client-side rendering paradigms and is not central to our contribution, we do not describe it further. The remainder of this section focuses on the second and third modules. For ease of exposition, we first describe the VC component and assume momentarily that an audio sample of the hallucinated voice is available, and that this voice is human.

Implementing VC. A natural approach to creating a VC model is to learn a mapping from an input recording of the source voice to a temporally aligned recording of the target voice comprising identical utterances. However, creating such a dataset is costly at the scale required to train a well-performing VC model. GOBLIN Voice performs voice conversion by reusing components of a pretrained text-to-speech (TTS) model. Zero-shot voice-cloning TTS models generate speech in a target speaker's voice from a short reference recording. Examples include TorToise [37], Voicebox [38], NaturalSpeech 2 [39] or CosyVoice [40].

For our model, we remove the autoregressive sequence-generating component and instead tokenize the sequence of words produced by the source voice (i.e., the psychiatrist). As before, these extracted token sequences are used, along with additional embeddings encoding the target speaker identity, to condition the mel-spectrogram generation. Ideally, the extracted tokens contain only linguistic (i.e., content and semantics) and para-linguistic (i.e., prosody and emotion) aspects of the source speech, with the target timbre

being modeled by the second-stage flow model. The CosyVoice family of frameworks [40,41] uses an automatic speech recognition objective that biases tokens towards linguistic content. In particular, we choose CosyVoice3 [41], whose pretraining additionally includes language identification and emotion recognition, which motivated its selection for GOBLIN Voice's multilingual and paralinguistic requirements. During development, an informal internal listening comparison assessed preservation of source expressivity, absence of target-speech expressivity leakage, absence of source-timbre leakage, fidelity of the cloned target timbre, and overall speech naturalness and absence of artifacts. Several open-source voice conversion models were compared, including Seed-VC [42], FreeVC [43], Chatterbox [44], MeanVC [45], CosyVoice 2 and CosyVoice 3. The evaluation was conducted as an informal internal assessment rather than a controlled listening study. While the listeners were unanimous in their qualitative judgments, these observations should be interpreted only as development observations, not as quantitative evidence of differences between models and are reported solely as motivation for model selection.

We remark, however, that remnants of the source voice are still perceptible, even if highly diminished compared to alternative models. To further mitigate this effect, we propose tuning the voice conversion mask directly on speech obtained by transforming the psychiatrist's voice into the hallucinated voice, thereby aligning the transformation process with the target acoustic characteristics.

Reconstructing the hallucinated voice**.** Apart from the few cases in which the hallucinated voice belongs to a real person from whom an audio sample can be obtained, the reconstruction must be done ab initio. To this end, we will provide a collection of predefined voices with good coverage of gender, age, and accents from which patients select the subset most similar to their hallucinated voice as a starting point. To refine this initial selection, one or a subset of the following steps may be performed: (i) interpolating voices in the selected subset to obtain a novel voice, possibly achieving greater similarity with the target voice, (ii) employing signal processing techniques to tune the baseline voice according to the vocal dimensions identified in Section 2.2.1, (iii) perturbing the speaker embedding along specific directions, including principal directions of variance. We now elaborate on these approaches.

Following the selection of a baseline target voice, a patient-guided iterative refinement procedure (Figure 2) may further improve perceptual similarity via signal-processing techniques. In practice, perceptually salient transformations associated with speaker characteristics are constrained to a limited set of operations that alter the voice's pitch, spectral envelope, spectral tilt, and temporal roughness [46]. Further fine-grained alterations may be enforced by tuning the speaker embedding by conditioning the reverse process in the flow-matching module of CosyVoice3. Prior studies [47] show that it is possible to use Gibbs sampling [48] with human feedback by asking the user to choose from several variations of the same voice. These variations are generated by perturbing the speaker embedding and regenerating the speech. Prior studies [47] suggest that up to thirty-two refinement steps may be required for achieving a high similarity, which, as discussed in Section 2.2.1, could lead to excessive cognitive and emotional strain on the patient at the outset of therapy. We therefore propose distributing refinement across multiple therapy sessions to allow gradual adjustment to the generated voice, being mindful of the patient's reaction during the process. Our approach uses one slider for each voice operation and enables their adjustment upon listening, unlike other voice design strategies such as VoiceX [49] that expose the user to audios obtained by randomly selected transforms. This interactive control mechanism reduces the time required for exploration and helps avoid inefficient search trajectories or convergence to suboptimal local solutions.

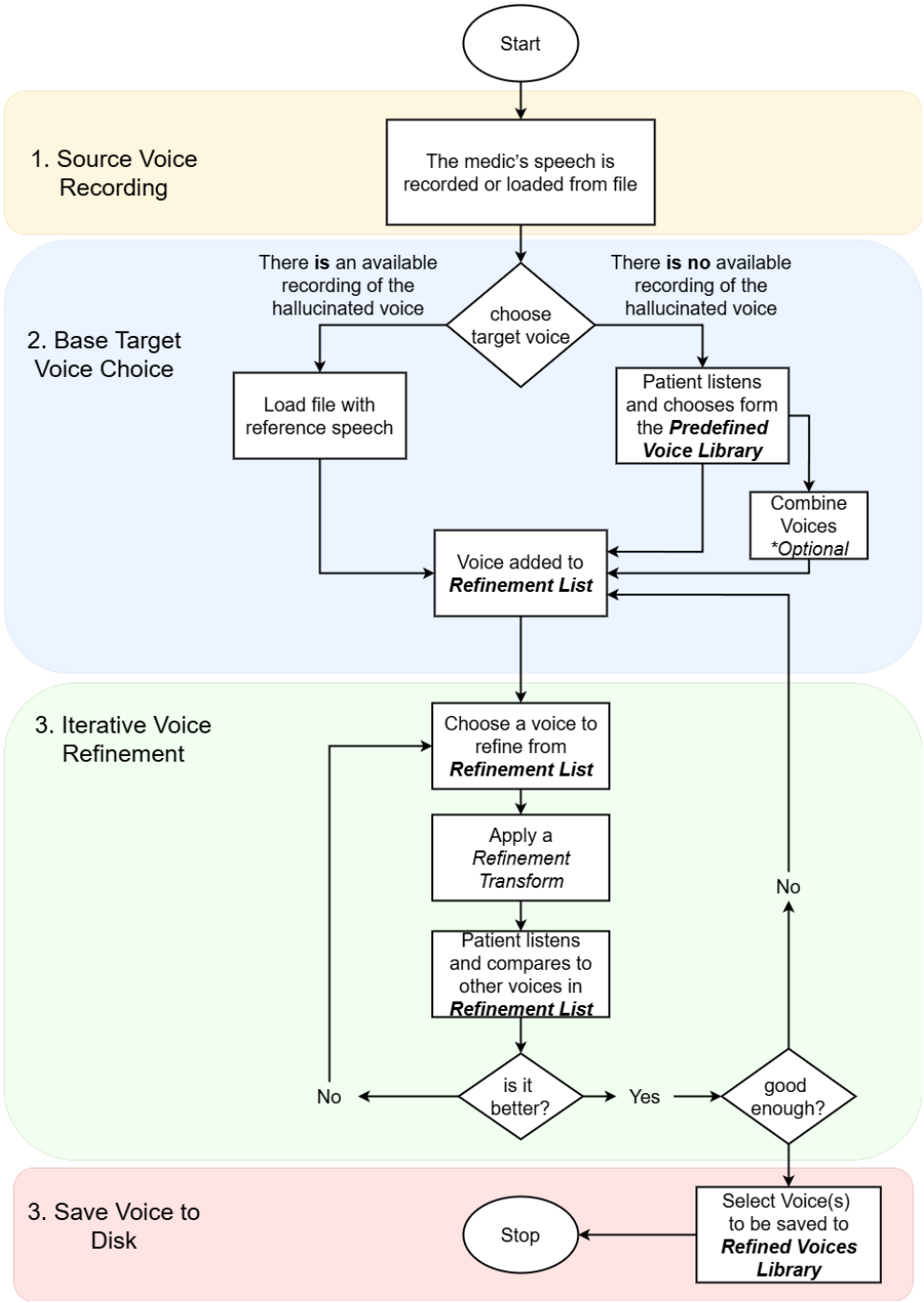


Figure 2. Iterative refinement procedure. *—optional step.

Intended deployment and usage. During therapy, the system continuously records the therapist's speech, which is processed in discrete chunks segmented using a voice activity detection (VAD) algorithm. This enables boundary selection in regions of low or absent speech activity, thereby reducing processing artifacts. The system adopts a pipelined architecture in which recording, voice conversion, signal processing, and playback operate concurrently. This design supports the near-real-time technical audio processing of bounded end-to-end latency, with continuous, albeit slightly delayed, audio output of the converted speech.

Session setup. GOBLIN Voice sessions are audio-only, with no visual avatar or head-mounted display. The clinician and participant sit in separate clinic rooms, or are separated by a wall, linked by a standard audio-conferencing application; the participant

wears headphones, and remote sessions are outside the current validation program. On the clinician's workstation, speech is converted in 2.5 s chunks and routed to a virtual microphone set as the conferencing input, with end-to-end delay on the order of seconds. For natural-voice segments the clinician switches manually to the physical microphone; an in-app toggle is planned but not yet built. The clinician's camera stays off, so the participant sees only a neutral call screen and cannot view the clinician while the avatar voice is active; the participant's camera stays on for one-way viewing, letting the clinician monitor non-verbal distress.

2.3. GOBLIN Sight

2.3.1. Clinical Rationale

GOBLIN Sight is an investigational webcam-based oculomotor module within the GOBLIN platform. Its proposed role is to explore whether eye movement measurements could provide additional context for clinical assessment and longitudinal follow-up in people with schizophrenia spectrum disorders. Task selection is informed by group-level oculomotor findings in these populations involving antisaccade performance, smooth pursuit, visual exploration and blink-related variables [27,50–54]. These findings provide a rationale for investigating the selected measures; their technical reliability and clinical relevance within GOBLIN remain to be established.

The proposed framework includes individual task measurements and the future development of candidate summaries of reference-relative profiles and within-participant change. The task domains and their scientific rationale are outlined below and summarized in Table 1.

Table 1. Standard GOBLIN Sight test battery (excluding calibration).

Task	Approximate Runtime (Minutes)	Candidate Measurements	Scientific Rationale
Antisaccade	4–5	Directional error rate and correction rate.	Examines inhibitory control, informed by group-level findings of elevated antisaccade error rates in psychosis (one of the clearest oculomotor findings, the marker appears to reflect illness liability) [50].
Smooth pursuit	4–5	Target–gaze position error, target–gaze alignment and pursuit gain.	Examines tracking of a moving target, informed by group-level pursuit abnormalities reported in schizophrenia and psychosis research (pursuit quality may be more informative for repeated assessment over time) [27,51].
Free viewing	4–5	Gaze coverage, dwell distribution, scanpath length, fixation count and fixation duration.	Examines spontaneous visual exploration, informed by group-level differences in exploration patterns in schizophrenia (already reproduced in objective discrimination between patients and healthy controls) [53].
Blink metrics	2	Blink rate and blink duration.	Provides contextual measurements whose interpretation requires consideration of medication exposure and testing conditions (useful especially in patients without medication or with more acute symptoms) [54].

The listed measurements describe the proposed research framework and include measures requiring further implementation and technical validation. Their availability and acquisition limitations in the current prototype are described in the sections below.

A more extensive description of the proposed testing methodology is provided in the Supplementary Materials (File S1). The purpose of this test battery is to provide candidate oculomotor outputs that could complement clinician evaluation and follow-up. In planned future composite-development studies, eligible oculomotor measurements may contribute to a personalized scoring framework incorporating reference-relative and within-participant comparisons. This framework would be designed around two exploratory functions:

- Trait-like function—a full test battery in a one-time (ideally initial) session, intended to generate a candidate cross-sectional user measure for investigating the general phenotypic similarity of the user to cohorts described in psychosis research, and a provisional baseline for subsequent measurements;
- State-like function—repeated subsequent sessions with a potentially shorter test battery, interpreted relative to the user's own prior measurements, to evaluate within-participant change over time and its association with independently assessed clinical change.

To support practical testing within clinician-supervised clinical settings, the battery is intended to prioritize tasks that may yield interpretable measurements with webcam-class hardware. Performance across devices and testing conditions remains to be validated.

In a later composite development study, the proposed scoring framework would use standardized proxy scores; each eligible oculomotor measurement would first be direction-corrected so that higher values uniformly indicate greater deviation in the prespecified direction. Each variable would then be standardized against a prospectively derived, demographically matched reference distribution and, for reporting purposes, rescaled to a common 0–100 reference-relative scale [55]. This is proposed as an appropriate first step because the battery includes measures expressed in different units and with opposite prespecified directions; standardization is the conventional way to place such heterogeneous variables on a common scale so they can be compared and considered for aggregation, and standardized values can then be converted into a potentially more clinician-friendly score format without changing their underlying ordering [56]. Demographic matching is also methodologically justified, because normative interpretation in clinical measurement routinely depends on variables such as age, sex and education, and eye-movement measures themselves show systematic age effects and, for some parameters, sex-related variation [57]. For longitudinal follow-up, the same logic would be applied within the patient's medical trajectory; current performance would be compared with the person's own baseline and candidate interval bands would be developed and validated to assess change beyond expected measurement variability, following the rationale of reliable change methods rather than raw score differences alone.

The standardized variables would then besummarized into four proposed theory-led domain scores: antisaccade dysfunction (ASD), pursuit tracking dysfunction (PTD), exploration-environment dysfunction (EED), and blink/arousal deviation (BAD). Grouping variables in this way is intended to preserve the link between the final score and the underlying oculomotor function, while potentially reducing dependence on noise in any single raw marker. Whether this improves reliability would require empirical evaluation. More broadly, some theory-driven composites have been shown to exhibit lower intraindividual variability than single tests and to outperform both raw measures and purely empirical composites on selected measurement properties in longitudinal cognitive aging research [58]. In schizophrenia research specifically, integrated eye movement scoring is already supported by precedent—Morita et al. derived an integrated

score from free-viewing, smooth-pursuit, and fixation variables that distinguished schizophrenia from healthy controls with 82% accuracy [59].

At the composite level, the proposed conceptual prototype includes two exploratory candidate measures with different purposes (candidate specification presented in Supplementary Materials—File S1). The first is GEPS (GOBLIN Eye Psychosis Score), a cross-sectional score intended to summarize the baseline oculomotor profile relative to an appropriate reference distribution, informed by patterns described in psychosis research. The second is GEDS (GOBLIN Eye Destabilization Score), a longitudinal score that would summarize one-sided directional change relative to the patient's own clinically stable technical baseline. The two proposed metrics are intended as candidate composite measurement outputs designed to organize heterogeneous oculomotor measurements into interpretable summaries for prospective validation. The current prototype does not calculate these domain or composite scores. Initial studies will evaluate individual measurements; composite derivation and independent validation are deferred to larger, separately planned studies. Neither measure has been validated as a biomarker, diagnostic classifier, psychosis risk measure, relapse-detection or relapse-prediction tool, or treatment trigger. While GEPS is intended as an exploratory baseline phenotyping score, GEDS is intended as a within-patient change score. Having no established independent clinical meaning at the moment of writing, these scores would be evaluated for their ability to support clinician review, potentially providing an additional dimension of evaluation. If future derivation and independent validation establish clinical interpretability, these candidate scores could subsequently be evaluated as adjunctive clinician-facing summaries. No clinical action thresholds or treatment triggers are proposed at the present stage.

2.3.2. Implementation Details

GOBLIN Sight is a software module that provides a partial implementation of the oculomotor tasks described in Section 2.2.1. Its software architecture is conceptually divided into two tightly coupled modules, as follows:

- An interactive minigame environment that presents the patient with visual cues via a computer display and handles task execution behaviors;
- A webcam-based eye-tracking analysis module that estimates the patient's point of view on the display from sequential webcam frames.

The interactive minigame layer encodes the experimental scenario and presents different visual targets through the user interface while collecting task-specific data. Concomitantly, the eye-tracking analysis module transforms the raw video stream into gaze estimates and descriptive task outputs. The minigame layer is implemented in Python 3.14 using a rendering library.

In the integrated prototype, the task runner (version 0.1.0) operates as a standalone executable launched by the Electron client (software application built with Electron 43.1.1), with one task performed per run. The implemented tasks comprise a randomized mixed pro-/antisaccade block of 24 trials, a 30 s smooth-pursuit task with random direction changes, and a 30 s free-viewing task. These task durations exclude calibration. Blink flags are derived from recordings obtained during these tasks; a dedicated blink-baseline task has not been implemented.

Each run stores gaze estimates, timestamps, task events and calibration data. The client analysis workspace (version 1.1.0) provides descriptive summaries, including valid- and missing-sample percentages, task outcomes, response zone entry times, pursuit position error and free-viewing dwell maps. Domain scores and GEPS/GEDS are not implemented. Testing to date has been limited to internal developer sessions and

automated checks of the analysis engine and results display; formal measurement validation remains pending.

The remainder of this section describes the gaze-estimation pipeline and its current technical limitations.

Equipment and hardware requirements. The pipeline requires a standard monocular RGB webcam (minimum 60 fps acquisition) to acquire images for gaze estimation. Measured values of the frame rate during usage will determine whether the session can be conducted reliably. A chin rest may be used to limit head movement during task performance. The camera is accessed through OpenCV using driver defaults. The decision to work with consumer-grade webcam tracking brings both opportunity and constraint. Recent validation work suggests that modern webcam-based eye-tracking can approach laboratory performance in selected tasks, especially when head position is controlled and the task design is appropriate [60]. However, webcam tracking is more vulnerable to device variability, calibration problems, lighting, viewing distance, and insufficient reporting of hardware criteria than conventional laboratory systems [61]. The initial evaluation will therefore prioritize descriptive task measures collected under standardized conditions, with calibration quality and repeatability assessed empirically.

High-level system description. The gaze position estimation functionality is based on solving a regression problem, in which input features extracted from the video stream are mapped to 2D screen coordinates. The processing pipeline has a three-stage architecture: i) feature extraction from raw video frames, ii) subject and setup-specific calibration and dataset construction, and iii) regression model fitting and inference. The regression model is fitted anew for each task run using the run's calibration data to account for individual differences in facial geometry, gaze behavior, and environmental setup, specifically the relative configuration of the monitor, webcam, and user position.

The first processing stage (Figure 3) extracts a numerical representation of the ocular and facial configuration from each frame of the webcam stream. To this end, we use a facial landmark detection module based on the MediaPipe computer vision framework [62]. In the current implementation, FaceLandmarker provides 478 landmarks, including iris landmarks. The regression input comprises three coordinates from each of 161 selected landmarks and three head-pose angles—yaw, pitch and roll—giving 486 features per frame. The selected landmark coordinates are normalized using an inter-ocular reference frame. Blink flags are computed separately from the eye aspect ratio.

The second stage consists of a calibration procedure performed at the beginning of each task run. During this phase, the user is instructed to fixate on a sequence of 25 targets arranged in a 5 × 5 grid across the display. Each target has a 0.9 s animated cue period followed by a 0.9 s capture period. For each fixation point, feature vectors are recorded during the capture period and paired with the corresponding known 2D display coordinates. This interactive task generates a user- and setup-specific labeled dataset that is stored with the run outputs and used to fit the gaze-regression model for that run.

The third stage of the pipeline consists of fitting a regression model to the calibration dataset to map the extracted landmark and head-pose features to 2D screen coordinates. The current integrated implementation uses ridge regression with standardized input features. The model is fitted separately for each task run using the calibration samples collected for that run, producing a mapping specific to the user and the current camera–display configuration. This mapping is subsequently applied to feature vectors extracted during the task.

Finally, the trained, personalized model is applied during task execution to estimate gaze position in 2D screen coordinates. Webcam frames are acquired through OpenCV 5.0.0.93 using the camera driver's default settings. Acquisition and inference operate sequentially within a processing loop capped at 60 iterations per second, with gaze

estimation performed once per acquired frame when landmark extraction succeeds. Calibration fit alone does not establish the accuracy of this mapping during task execution. This requires independent evaluation using calibration samples or trials that were not used for model fitting.

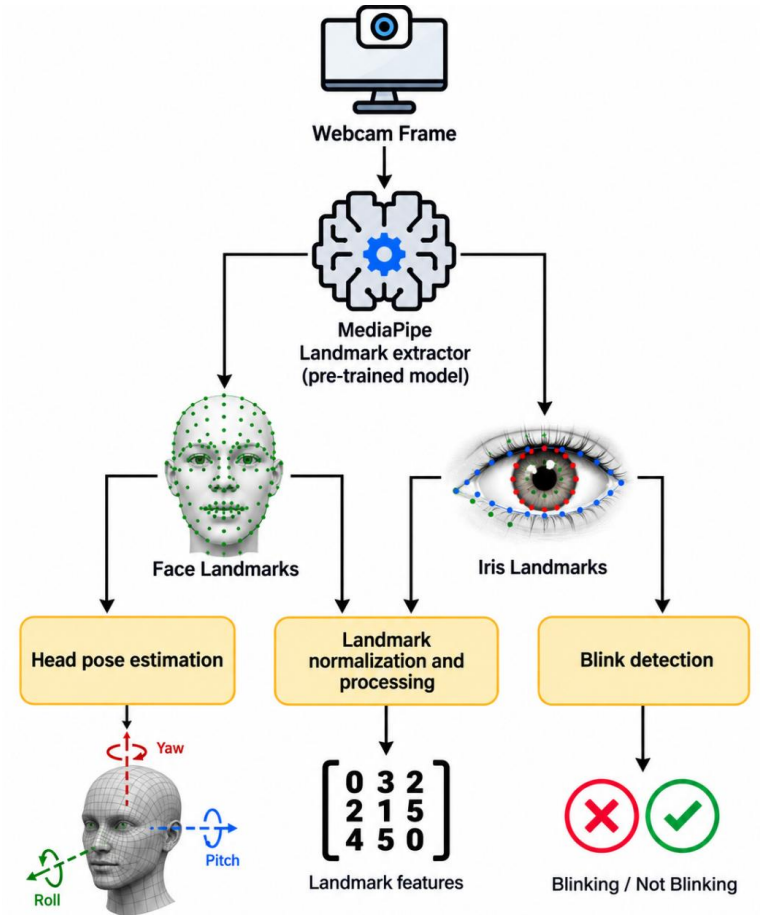


Figure 3. Feature extraction.

Timestamps are assigned after landmark extraction and regression using a high-resolution monotonic counter and are written with 1 µs numerical resolution. They therefore represent processing-completion rather than frame-acquisition time, and the precision of this timing relationship has not been characterized. Consequently, the temporal accuracy required for interpreting saccadic latency and fine event timing has not yet been established. These outputs should not be interpreted until acquisition timing and syncronization with stimulus presentation have been validated.

The effective gaze-estimation rate is derived post hoc in the analysis workspace from successive per-sample timestamps. The workspace reports the median, interquartile range and minimum rate, and logs the instantaneous rate for each successive sample. Processing latency is not logged separately. No interpolation is used to increase the gaze-estimation rate. In pursuit analysis, target samples are matched to the nearest earlier gaze estimate.

For each run, the analysis workspace reports the percentages of samples classified as valid, missing, missing during non-blink periods, blink-flagged or edge-clipped. It also records the number of dropouts and the duration of the longest dropout. The current implementation generates warnings when missing samples comprise ≥20% of a run or when a continuous dropout lasts ≥1 s. These are descriptive data-quality warnings and are not enforced task-validity thresholds.

The current implementation does not separately count dropped camera frames: failed captures and frames in which no face is detected are recorded as missing gaze samples. Calibration error is not calculated online, although the raw calibration features and corresponding target coordinates are stored for offline analysis. Drift correction and suppression rules are not implemented. Accordingly, no task-specific quality thresholds are currently enforced, and no gaze samples or derived task outputs are automatically suppressed on the basis of these criteria.

2.4. GOBLIN Assist

2.4.1. Clinical Rationale

The GOBLIN Assist module is intended to address the remaining structural gap identified in current schizophrenia care, namely that clinically relevant information about AVHs is lost between appointments. The LLM module targets the patient's account of their hallucinatory experience, with the aim of testing whether preserving and presenting this account to the clinician can provide clinically relevant information that may contribute to improved quality of care. Whether this indeed improves information continuity, clinician workload, or quality of care remains to be evaluated prospectively.

The module's proposed functions are organized into three contexts:

- Administration of clinician-authored questions between appointments;
- Source-grounded summarization and information support during and after appointments;
- Structured assistance during AVATAR-inspired therapy for the clinician.

Data generated within these contexts would be processed and stored in an access-controlled encrypted environment. Model-generated outputs would be clinician-facing, grounded in and linked to supporting source text, logged, versioned, and editable for clinician review and audit. The model would not diagnose, recommend or alter treatment, assign autonomous clinical risk levels, conduct psychotherapy, or replace clinician judgment. Examples of the intended functionality are provided in the Supplementary Materials (File S2).

Clinical motivation. Three observations in the schizophrenia care literature inform the proposed operating contexts. First, earlier outpatient follow-up after psychiatric discharge is associated with lower suicide risk [63], and structured, measurement-based longitudinal assessment may contribute to improved long-term outcomes [64,65]. Digital assessments have been shown to align with conventional assessments of functioning and positive, negative and depressive symptoms in schizophrenia [66], supporting investigation of clinician-authored questions between appointments as a complement to in-person assessment.

Secondly, autobiographical memory deficits in schizophrenia patients may limit recall of AVHs content during appointments [67,68], and routine assessment of key hallucination dimensions remains inconsistent in practice [69]. Time constraints and workload further complicate comprehensive longitudinal assessment. Computerized decision support in psychosis care has been associated with more care needs being addressed during consultations and more evidence-based treatment decisions [64,70]. These observations motivate the evaluation of a system that preserves timestamped patient-reported content and presents source-linked summaries for clinician verification.

Third, avatar-based interventions require clinicians to voice the avatar, alternate between avatar and therapist roles, and adapt the dialogue to the patient within a structured therapeutic frame. These training-intensive demands may limit broader implementation [24]. This motivates a clinician-facing assistance layer that proposes content from the patient's accumulated hallucination log, while leaving the clinician in

control of what is delivered. Whether this reduces workload or supports broader delivery without compromising safety remains to be evaluated prospectively.

Patient interaction. In the proposed workflow, following the initial appointment, the clinician would use the clinician-facing interface to author and schedule personalized questions about distressing AVHs, medication adherence, possible adverse effects after medication changes or initiations. The patient application would deliver these questions verbatim at specified intervals (e.g., once every 3 days at a set time). Responses would be recorded, timestamped and stored in an encrypted record.

After each response, the application would display a single, fixed, clinician-authored empathic acknowledgement confirming that the response has been saved and directing the patient to discuss it with their treating clinician. The acknowledgement would state that review is asynchronous, the platform is not an emergency service, and urgent help should be sought through local emergency services if the patient feels unable to stay safe. Patient interaction would be restricted to predefined prompts, response capture and this closing statement, with no autonomous dialogue. This restriction addresses concerns about the indiscriminate validation of patient experiences [71].

Under the proposed protocol, submissions would be processed on receipt and reviewed asynchronously by the treating psychiatrist, with review notifications sent to both the treating and designated covering clinician. Consent materials, onboarding and every acknowledgement would explain that the platform is not continuously monitored. Submissions received outside working hours would be reviewed on subsequent working days; consent materials would specify that waits may reach approximately 72 h over ordinary weekends and extend further during holiday periods. Flagged submissions would receive priority, with those received during working hours reviewed the same working day. Items remaining unacknowledged for four working hours would automatically escalate to the clinical lead (MEDICAL_ADMIN). All notification and delivery events would be recorded in an append-only audit log.

Emergency information would be localized for each deployment site and displayed in the patient application and every acknowledgement, including 112 and the relevant suicide-prevention helpline for Romanian deployments. A persistent in-application notice would reiterate that the platform is not an emergency service and is not continuously monitored. Submissions awaiting review would appear in a dashboard queue, with flagged items first and chronological ordering within each group, without score-based ranking. Each item's age would be displayed, and a weekly report would summarize review times.

During the feasibility stage, clinicians would review all submissions in full, allowing missed safety-related text signals to be identified retrospectively. For each submission, the protocol would record time to review, whether any flag was confirmed or judged not relevant, any missed-risk events identified during review, and associated adverse events. The clinical and technical teams would review these data monthly to monitor residual risks and delays.

Safety-related text signals. In the proposed inter-appointment workflow, the system would analyze patient-written responses and automated speech-recognition transcripts for language associated with suicidal ideation. Relevant passages would be highlighted within the full source response and accompanied by a "flagged for review" indicator in the clinician's structured report. This flag would direct attention to potentially relevant content without assigning clinical risk categories, severity tiers, or diagnostic or predictive risk scores. All responses would remain available in full for scheduled clinician review, regardless of the classifier output. Clinical interpretation and subsequent action would remain the treating clinician's responsibility.

Clinician support during appointments. The proposed module would present transcripts, source-linked summaries of between-visit responses, and longitudinal comparisons of patient-reported content. These comparisons would describe reported content without establishing a validated emotional trajectory or clinical change. The clinician would verify, edit, accept, or reject all generated content before using it in clinical interpretation.

AVATAR therapy support. During clinician-supervised AVATAR therapy-inspired sessions, the model is intended to present optional clinician-facing draft suggestions derived from the clinician-approved hallucination record. Each suggestion would be source-grounded, auditable, editable, and dismissible, and would not be delivered directly to the patient or determine the direction of therapy. Candidate outputs would be limited to two types of suggested responses:

- Source-linked excerpts from previously recorded patient-reported hallucination content;
- Proposed supportive responses if the patient struggles to reply;

The clinician has the ability to modify, discard or select the suggestions and the module does not address the patient directly. Additionally, it records and updates any hallucinatory content following the session.

2.4.2. Implementation Details

At present, GOBLIN Assist is not accompanied by quantitative prototype results, and this section serves as a methodological proposal intended to meet the opportunities identified in the psychiatric literature. Empirical benchmarking and clinical validation remain part of the roadmap, as a future contribution.

GOBLIN Assist is intended as an integrated software module providing a digital instantiation of the functionality described in Section 2.4.1. The tentative architecture comprises five distinguishable parts,

1. A graphical module with distinct patient- and clinician-facing interfaces broadly structured as a conversational user interface;
2. A notification and response-management component responsible for scheduling the delivery of clinician-authored questions, collecting patient responses, storing them under the corresponding user identity, and presenting a summary of the responses during subsequent appointments with the clinician;
3. An Automated Speech Recognition component (ASR) used to transcribe patient recordings into text;
4. A neural network-based model intended to generate a candidate probability that patient-provided text contains language associated with self-harm, for clinician review;
5. A neural network-based model intended to generate candidate emotion-related label probabilities from patient-provided textual inputs as well as the treatment session transcripts;
6. A clinician-facing Large Language Model (LLM) intended to produce source-grounded, source-linked, and auditable summaries and optional draft suggestions for clinician review before, during, and after clinician-supervised sessions.

The first and second components would use standard client-side rendering paradigms and established design patterns for notification scheduling, data ingestion, and event-driven processing, respectively. A detailed description is therefore omitted, since these components are not central to our contribution. We focus the remainder of this section on the machine learning models supporting clinician review and information organization.

Given the scarcity of task-specific Romanian clinical datasets, the proposed supervised models would initially use Romanian translations of English-language corpora for development, not validation in the target clinical population. Translation would use an open multilingual neural machine-translation system, followed by automated quality estimation and adequacy and fluency ratings given by two bilingual annotators on stratified random samples. Items below a pre-specified quality threshold would be excluded. All positive-class risk items and a random sample of the remaining data would undergo back-translation to identify clinically relevant semantic drift. A psychiatrist would then review all positive-class items and a random sample of retained negative items to confirm label appropriateness in Romanian; items with altered meaning would be relabeled or excluded.

Performing ASR on session recordings. Our starting point is an ASR model based on OpenAI's open-source multilingual Whisper Large V2 [72], further fine-tuned to SOTA level for the Romanian language by Păiș et al. [73]. The model is 1.5B parameters large and was refined using a variety of Romanian-inclusive ASR datasets such as CommonVoice [74], VoxPopuli [75], RSS [76], RTASC [77] CoRoLa [78], thus being the ideal candidate. To attribute transcribed speech to individual speakers, the pipeline would add speaker diarization, which Whisper does not provide natively. To this end, the proposed diarization pipeline would label speech segments in parallel with Whisper's transcription. The proposed pipeline would use the lightweight WebRTC Voice Activity Detection module [79] to split the dialogue into single-speaker segments. The resulting segments would be passed to a pretrained ECAPA-TDNN neural network [80] to extract speaker embeddings and, in parallel, submit the segment to the Whisper-based model for transcription. The resulting embeddings would then be clustered online using SOTA hierarchical clustering algorithms [81,82] to extract the relevant label. Finally, overlapping segments would be resolved and timestamps aligned in order to produce the labeled and transcribed audio segment.

Detecting emotions and the presence of suicidal ideation in patient transcripts and textual submissions. These outputs are exploratory text-derived probability signals intended to prioritize clinician review, not validated patient-level emotion classifications, risk categories, diagnostic outputs or autonomous escalation triggers. Since text-based classification underlies both tasks, we address them as a conceptual unit. Moreover, they share the structural component responsible for constructing the textual representations on which classification is further performed.

In particular for the emotion recognition task (ERC), the proposed emotion-recognition component would use a multi-label classifier based on the XLM-RoBERTa [83] language representer. XLM-Roberta is a cross-lingual language model that achieves state-of-the-art natural language understanding and allows for cross-lingual transfer. Among other languages, it was trained on a 61.4 GB Romanian-language corpus, making it particularly suited to our purpose. The full description of the dataset behind XLM-Roberta is provided in [83]. A fully connected classification layer would be added to perform the classification of the appropriate dimension on top of XLM-Roberta and the resulting model would be fine-tuned on three Romanian-translated datasets described shortly. The classifier input is a dialogue sequence, while the output is a six-dimensional vector per participant quantifying the presence of each emotion as a percentage. Training would use cross-entropy loss and the AdamW optimizer [84] with dropout for added regularization. The datasets used for training are translations of the GoEmotions [85] and the Emotional Dialogues in OpenSubtitles [86]. These datasets are used for English-based ERC and contain dialogue sequences manually labeled with fine-grained emotion annotations beyond Ekman's six basic emotions [87]. For our purpose, these datasets would be coalesced into a single unit and subject to label uniformization.

The approach to detecting the presence of suicidal ideation in text closely follows the steps of the emotion detection classifier, but in a Multitask Learning context. Multitask learning helps the model generalize better, as it learns from different datasets simultaneously and develops a shared understanding of the tasks. The proposed component would use a separate XLM-RoBERTa instance as a shared representation model, with a binary classification head trained for the review-prioritization task. At each step, the input is a piece of text produced by the patient (either through ASR or their answers to clinician-authored questions), while the output is a value in the interval [0, 1] representing the model's confidence that the text contains language associated with suicidal ideation. For training, the development data would include the CLPsych 2016 ReachOut corpus [88] and the SuicideWatch corpus [89]. ReachOut uses moderator-triage categories (green, amber, red and crisis), rather than clinical risk levels, and these original labels will therefore be retained in the development data. For the intended review-prioritization function, the labels will additionally be mapped to a binary target indicating whether content should be flagged for clinician review. Where tier labels are available, an auxiliary ordinal output will also be retained during model development. The binary output is intended solely to prioritize clinician review and does not constitute an automated clinical risk classification. This simplification is justified by the classifier's purpose: to flag content for clinician review. The optimizer and training parameters remain unchanged relative to the previous setting.

Dataset use will be conditional on confirmation that the applicable licenses or data-use agreements permit translation and derivative model development for research. GoEmotions is distributed under the Apache 2.0 licence. Use of the OpenSubtitles-derived emotional dialogue corpus will depend on the confirmation of both the underlying corpus and annotation-release terms, while the current availability and permitted use of the CLPsych 2016 ReachOut corpus will be verified under its data use agreement with ReachOut Australia; current availability will be confirmed before use. Reddit-derived SuicideWatch collections will be handled in accordance with the applicable platform terms. Any dataset whose terms do not permit the proposed use will be excluded.

To reduce information leakage, development datasets will be split at the level of the originating post, thread or dialogue, and by author where available, rather than at the sentence level. Deterministic grouped splitting with an automated leakage assertion and exact de-duplication are incorporated into the data-preparation pipeline. Exact and near-duplicate removal (MinHash) [90] and an n-gram contamination check between training and evaluation sets will be applied. The grouped split with leakage assertion and exact de-duplication exist in our data-preparation script. The MinHash and contamination checks exist in a separate sanitization stage that will be wired into data preparation before any classifier training.

LLM-based clinician assistance. The proposed LLM component is based on an open-source multilingual Qwen3-family model [91]. Qwen3 was selected as a multilingual model with appropriate licensing for the intended use, documented evaluation in Romanian, and emerging Romanian-specific benchmark evidence, including evaluation of Qwen3-32B on the GMTW-Ro benchmark [92]. The planned development approach would fine-tune this baseline LLM for each of our goals of free-text and dialogue summarization, machine-reading comprehension and question-answering. Prior to any task-specific fine-tuning, the planned pipeline would perform continued pre-training on Romanian and psychiatric domain text corpora. This step's goal is to improve Romanian clinical fluency, psychiatric terminology, colloquial patient language, non-diacritic Romanian handling, and robustness to patient-style short or fragmented responses. The planned continued-pretraining corpus would comprise a curated mixture of Romanian

general and conversational text, and authoritative Romanian language medical and psychiatric material.

Task-specific fine-tuning would use parameter-efficient LoRA adapters [93]. Specifically,

A. The proposed Dialogue Summarization adapter would be fine-tuned using DialogSum [94]. In the intended workflow, this adapter would generate draft summaries of the therapeutic dialogue at the end of a GOBLIN session. This adaptor would also be used to summarize historical session information.
B. Finally, the proposed Medic Assistant adapter would be fine-tuned using the Machine Reading Comprehension dataset [95] and datasets addressing question-answering and instruction-following like Natural Questions [96], TriviaQA [97] and Dolly 15k [98]. The adapter is intended to support the clinician review of patient data during evaluations. A planned alignment stage would use RLHF [99] with the involvement of a specialist medical team instructed to rank two clinician-facing draft suggestions proposed by the model for mock therapeutic situations. These responses would be ranked according to safety, empathy and helpfulness Likert scales.

All planned supervised fine-tuning would use the typical Cross Entropy loss optimized by AdamW with dropout. The planned RLHF stage would follow the methodology proposed by Apple to refine its foundation models [100] by using their proposed soft-label loss function in conjunction with the Bradley–Terry–Luce reward model and Mirror Descent with Leave-OneOut estimation algorithm.

2.5. Proposed Clinical Workflow and Implementation Status

The GOBLIN platform is intended to integrate its individual modules within a clinician-supervised patient pathway; however, the workflow described below represents the proposed integrated use of the system rather than its current clinical implementation. In the proposed workflow, the patient pathway would begin with a clinician-led assessment of diagnosis, current clinical state, capacity to consent, risk profile, medication status, hallucination characteristics and suitability for digital interaction, gathered via the clinician-facing interface during the establishment or revision of individual subject records. If considered appropriate for the platform, the patient would provide informed consent and the clinician would determine which modules should be used.

GOBLIN Sight would then be used to acquire webcam-based oculomotor measurements for clinician review alongside interview findings, symptom scales, medication exposure, sleep, adherence, functioning and risk assessment. The current prototype provides task-level recordings, quality-control information and descriptive outputs; the proposed domain scores and GEPS/GEDS composite measures are not currently implemented and remain subjects for future development and validation. Any future oculomotor outputs would be interpreted only in clinical context and would not independently determine diagnosis, relapse, risk or treatment decisions.

GOBLIN Assist is intended to provide a longitudinal clinical-support layer in which clinicians could define structured questions for delivery between appointments and subsequently review timestamped patient responses. Future implementations may also include automated text signals that prioritize selected responses for clinician review, without assigning autonomous clinical risk levels or making risk-management decisions. The LLM component is intended to remain clinician-facing and restricted to source-grounded information capture, summarization and optional draft support rather than autonomous psychotherapy, clinical advice or treatment modification.

For participants considered suitable for AVATAR-inspired intervention, GOBLIN Voice would support clinician-mediated dialogue using a personalized approximation of

the distressing voice. The clinician would retain control of the therapeutic interaction, including whether to use the converted or natural voice, and whether to pause, redirect or terminate the exchange. Early participant-facing evaluation would be restricted to clinically stable, consenting individuals able to understand and tolerate the therapeutic framework, and would be postponed or excluded in the presence of acute agitation, imminent suicide or violence risk, severe disorganization, intoxication or withdrawal, inability to consent, marked suspiciousness toward the system or substantial distress during voice setup.

Because these functions are at different stages of development, their current technical status should be distinguished from the intended integrated workflow described above. The tables below summarize, for each GOBLIN component, the functionality currently implemented, technical testing performed, data used and results available, known limitations, functionality that remains planned, and the subsequent validation stage. No participant-derived clinical, therapeutic, diagnostic, relapse-prediction, or validated safety outcomes are available at the present stage.

For consistency throughout the remainder of the manuscript, present-tense descriptions refer only to functionality identified as implemented in Table 2, while planned machine learning, patient-facing, scoring, alerting, and clinical workflow functions are described prospectively.

Table 2. Implementation status details.

Module (Version)	Implemented	Not Implemented/Planned	Technical Testing	Data Used	Results	Main Limitations	Next Validation Stage	Not Claimable Clinically
Voice (engine 0.1.0; protocol v1)	functional software prototype supporting initial voice personalization; voice conversion for both recorded audio and live communication	integrated switching between converted and natural voice; additional controls for voice realism and personalization; further refinement of voice conversion; support for additional operating systems	automated software unit tests (model mocked); informal developer checks; informal listening comparison of six VC models before implementation; no controlled listening study	publicly available pretrained model (CosyVoice3); developers' own recordings for engineering checks. No patient or participant data; no model fine-tuning performed by the authors	no formal benchmarks; current-version latency, intelligibility and speaker similarity remain unmeasured.	restricted hardware compatibility; multi-second delay; informal persistence of original voice characteristics; limited reference duration; no formal evaluation; audio-call integration not yet hardware-validated	Technical validation (latency, word error rate, speaker-similarity benchmark)	Any effect on distress, voice power or severity; resemblance to the hallucinated voice; suitability of the delay
Sight (runner 0.1.0; contract v1; analysis 1.1.0)	functional webcam-based gaze-estimation prototype with calibration; pro-/antisaccade, smooth-pursuit and free-viewing tasks; blink flagging;	dedicated blink task; camera frame-rate logging; dropped-frame counting; calibration-error estimate; drift correction; quality gating; fixation/saccade classification;	informal developer checks; no formal hardware acceptance	No participant data	None	webcam-limited sampling rate, not yet measured; processing-related timestamp delay; corrective saccades not captured	Technical benchmark against a reference eye tracker; data-quality reporting	Any biomarker, diagnostic, risk, relapse or individual-level property

	secure data transfer and descriptive analysis	pursuit gain; conversion to degrees of visual angle; domain scores; GEPS/GEDS.						
Assist (sandbox rev. v0.6.1; training rev. v0.8.4)	partial implementation with clinician-facing Romanian-language chat prototype with feedback collection and preliminary model-development tools	patient-facing functions, automated analysis and summaries, clinical review workflows and review-prioritization flags; further model refinement and data-sanitization integration	component unit tests and preliminary training runs; sanitization tested separately, not yet integrated	synthetic Romanian examples for pipeline testing and general-domain Romanian web text for language model evaluation; no patient data or clinical training corpus	No formal results	no Romanian clinical corpus or safety-related text detection; data sanitization not yet integrated, only partially implemented	Technical validation of each ML component after implementation	Any detection of risk, emotion or deterioration; any effect on continuity or workload; any patient-facing function
GOBLIN Platform (backend 0.1.0; client 0.1.0; contracts v1)	clinician-facing platform with role-based access, secure authentication, encrypted data storage and transfer, audit logging, Sight and Voice session management, and a dashboard for reviewing recorded outputs	patient-facing application; review notifications, acknowledgement tracking, and escalation workflows; medication and adherence tracking; AI processing; multi-factor authentication; signed software distribution and automatic updates	automated backend, client and interface tests using simulated components; no live-database integration or automated hardware testing	Synthetic seed data	None	internal test build only; data-protection assessment, legal basis for health-data processing, medical device classification, retention policy and signed distribution remain unresolved release requirements	Technical validation (verification, audit completeness)	Any workflow benefit

3. Methodological Roadmap for Prospective Validation

3.1. Rationale

Following system implementation towards sufficient technical maturity, the immediate translational objective is to define how the system will be evaluated in a staged and reproducible manner. Because GOBLIN integrates biometric measurements, clinician-supervised voice avatar interaction and LLM-assisted clinical support, its validation cannot be reduced to a single endpoint. The current guiding principles behind GOBLIN research focus on the following validation program elements: technical validity, acceptability, feasibility, auditability and exploratory clinical anchoring. No clinical efficacy is established in the present work. Any efficacy evaluation will require subsequent controlled studies after predefined technical, usability, safety, and feasibility criteria have been met. This methodological approach is consistent with early-stage AI and digital-health evaluation guidance, which emphasizes transparent reporting of versioning,

intended use, human oversight, failure scenarios and workflow context before more definitive trials [101–103].

Accordingly, the validation strategy presented in the current article is designed as a prospective roadmap, abiding by its guiding principles. The purpose is to define the empirical outputs required before GOBLIN can be evaluated in controlled clinical studies, before any claims can be made. This structure follows the general logic of early-stage evaluation in digital health and AI-assisted healthcare, where systems should first demonstrate technical reliability, workflow integration, usability, safety, and transparent reporting before being tested for clinical effectiveness.

3.2. Overall Validation Architecture

GOBLIN will be evaluated through a staged validation architecture (Table 3). Each stage produces predefined outputs that determine whether the platform should progress to the next level of testing. The first stage verifies whether each module fulfills its technical specifications. The second stage evaluates acceptability and usability for intended users. The third stage evaluates component-level validity against reference systems, clinician annotation, or clinical anchors. The fourth stage evaluates the integrated workflow in a clinician-supervised feasibility pilot. The fifth stage, outside the scope of the present prototype description, would consist of a controlled clinical evaluation of efficacy outcomes. Statistical methods for sample-size planning, missing-data handling, longitudinal modeling and multiple comparisons are specified in File S3.

Table 3. Staged validation architecture for GOBLIN.

Stage	Objective	Indicative Outputs
Technical verification	Verify core technical operation under controlled conditions.	Calibration success, gaze validity, latency, conversion failure, summary generation, alert delivery, audit logs.
Acceptability and usability	Assess whether users can tolerate and complete module workflows.	Completion, dropout, distress/fatigue, SUS/uMARS or equivalent, clinician workload, willingness to repeat.
Exploratory component-level validation	Test whether outputs are reliable, clinically interpretable, and safe.	Sight reliability and interpretability of individual measurements; Assist source-grounded summary accuracy and candidate text-signal performance; Voice technical similarity, dominance modulation, tolerability.
Integrated feasibility	Test the clinician-supervised pathway.	Recruitment, retention, adherence, clinician review completion, review-prioritization flags, overrides, adverse events.
Future efficacy evaluation	Test clinical benefit after feasibility is established.	AVH-related distress, perceived voice power, relapse-related contact, functioning, treatment engagement, clinician workload.

3.3. GOBLIN Sight Validation Methodology

The initial GOBLIN Sight evaluation will determine whether this webcam-based oculomotor module can acquire technically usable gaze data, support the completion and repetition of a standardized task battery, and produce individual measurements with preliminary evidence of reliability and interpretability. Composite-score development is reserved for a subsequent, larger study. Eye-tracking research requires the detailed reporting of hardware, sampling, calibration, stimulus geometry, missingness, data quality and preprocessing decisions for reproducibility [104,105]. The detailed technical specification is provided in the Supplementary Materials (File S1). The expanded

validation methodology is provided in the Supplementary Materials (File S3). We provide below a high-level summary of the methodology.

The initial technical and feasibility study will include approximately 30 healthy controls and 30 clinically stable psychosis-spectrum participants. A retest subgroup of 20–30 participants will complete repeated sessions during clinically stable periods to estimate within-participant measurement variability. A benchmark subgroup of approximately 15–20 participants will complete the same tasks using GOBLIN Sight and a reference eye-tracking system, such as EyeLink, Tobii Pro, Gazepoint, Pupil Labs, or an equivalent validated device. These indicative samples are intended to estimate the technical performance, missingness, participant burden and preliminary reliability of individual measures, with uncertainty intervals. They are not intended to establish normative reference distributions, composite weights, comprehensive confounder-adjustment models or individual-level reliable change thresholds. The final protocol will justify sample sizes against these estimation objectives. Pilot estimates of measurement variability, repeatability and completion will inform subsequent study size calculations, with sensitivity analyses across plausible variance and attrition assumptions. Future efficacy trials will require a prespecified primary clinical outcome, an independently justified clinically meaningful treatment difference, relevant outcome-specific variance estimates, and prespecified power and significance levels. These planning principles are detailed in File S3, Section S3.4.1.

3.3.1. Tests, Metrics and Endpoints

The test battery (Table 4) includes calibration, antisaccade, smooth pursuit, free viewing, and blink baseline tasks. These tasks are retained because they map onto separable oculomotor functions and allow a modular quality control strategy. The initial evaluation will focus on individual oculomotor measurements and task validity metrics. The main technical endpoints are calibration success, calibration error, effective frame rate, valid gaze sample percentage, missing data percentage, task completion, invalid-trial rate, and session validity classification. The main exploratory oculomotor outputs are antisaccade error/correction metrics, pursuit gain/loss and target-gaze correlation, free-viewing scanpath and grid-coverage metrics, and blink rate/duration contextualized by medication and arousal. Eligible measurements may contribute to domain scores and GEPS/GEDS in later derivation and independent validation studies.

Table 4. GOBLIN Sight test battery.

Component	Core Technical Metrics	Endpoint Logic
Calibration	Calibration success, mean error, drift, repeat calibration rate	affected tasks excluded if calibration fails
Antisaccade	Valid trials, invalid/anticipatory trials, omissions	inhibitory control measurement; prioritize directional correctness over latency
Smooth pursuit	Effective FPS, tracking loss, missing samples	pursuit integrity; candidate measure for longitudinal analysis
Free viewing	Image completion, valid samples, fixation detection stability	restricted exploration and visual exploration measurement
Blink metrics	Recording completeness, detection confidence	contextual marker; medication/arousal-sensitive

3.3.2. Technical Benchmarking and Data-Quality Reporting

As webcam-based eye-tracking is sensitive to clinical, behavioral, and technical sources of variation, GOBLIN Sight outputs will be interpreted only after predefined contextual and quality control checks. At each session, medication exposure will be recorded, including antipsychotic type and dose where available, recent dose changes,

benzodiazepines, anticholinergics, hypnotics, and other sedating agents. Medication status will be treated as a potential confounder, particularly for blink rate, pursuit quality, latency, attentional dropout, and task engagement. Where feasible, antipsychotic exposure will be summarized using chlorpromazine-equivalent dose, while recent medication changes will be flagged separately.

Medication adherence and treatment modality will also be recorded as feasibility and external validity variables. Adherence category will be documented for every screened patient, including long-acting injectable treatment, regular oral treatment and partial or non-adherence where applicable. Feasibility outcomes, including eligibility-to-enrolment, refusal, completion, attrition and data-quality yield, will be reported by adherence category. Because the pilot recruits clinically stable participants and marked suspiciousness toward the system may constitute a postponement or exclusion criterion, potential selection bias will be assessed by recording both screened-but-excluded and approached-but-refused patients, including the reason, adherence category and, where routinely available, insight rating, and presenting these in a CONSORT-style flow. Only anonymized screening-log data will be retained for patients who refuse participation. Particular attention will be given to whether patients with poorer insight, persecutory symptoms or lower treatment adherence are less willing or able to complete webcam-based assessments, as this may limit generalisability.

Participant state will also be documented before testing. This includes observable or self-reported sedation, recent sleep duration, subjective sleep quality, daytime sleepiness, fatigue, agitation, intoxication where relevant, and ability to engage with the task. These variables will be used as covariates, stratification factors, exclusion flags, or sensitivity analysis variables depending on the validation stage. If the participant is visibly somnolent, severely fatigued, acutely agitated, intoxicated, or unable to sustain engagement, testing will be postponed, repeated, or marked as clinically uninterpretable.

Cognitive impairment and task comprehension will be handled through brief instruction checks and practice trials before measurement tasks. Failure to understand task rules, repeated non-response, or inability to follow basic instructions will be coded explicitly and will not be interpreted as oculomotor abnormality. Visual and ocular factors will also be recorded, including glasses or contact-lens use, uncorrected visual impairment, major ocular disease, neurological disorders affecting eye movement, screen visibility problems, and lens glare. Participants will be tested with their usual visual correction whenever possible.

Technical quality will be assessed at session and task level. Recorded variables will include webcam resolution, effective frame rate, lighting adequacy, face and eye visibility, head position, viewing distance, screen geometry, calibration error, calibration drift, valid gaze sample percentage, tracking loss, missing data percentage, and task completion. Calibration will be repeated if predefined quality thresholds are not met. If calibration remains inadequate, the session will be excluded from analyses requiring valid oculomotor measurements. In future composite development studies, such sessions would not be eligible for GEPS/GEDS derivation.

Missing or invalid data will be handled according to predefined quality rules. Trials or task segments with excessive tracking loss, blink contamination, unstable head position, off-screen gaze, implausible coordinates or insufficient valid samples will be excluded. Task measurements that fail minimum quality criteria will be reported as missing and will not be imputed for individual interpretation. Missingness will be documented at trial, task and session levels, with reasons including technical or calibration failure, environmental conditions, fatigue, difficulty understanding instructions, non-engagement, refusal, withdrawal or unknown causes. Refusal or exclusion before enrolment will be recorded separately from missing assessments after enrolment.

Missingness will be summarized by task, session and participant group. Available clinical and technical characteristics will be compared between completers and non-completers, with exploratory modeling limited to a small set of prespecified covariates. These comparisons will inform assumptions about missingness but cannot establish whether data are missing at random. Prespecified sensitivity analyses will examine how group-level estimates change under plausible missing-not-at-random assumptions, including delta-adjusted multiple imputation where supported by the available data (File S3, Section S3.4.3). Imputation will not be used to generate individual oculomotor measurements. In future composite development studies, imputed data will not be used for GEPS/GEDS derivation.

Technical benchmarking will compare GOBLIN Sight with a reference eye-tracking system where feasible, such as EyeLink, Tobii Pro, Gazepoint, Pupil Labs, or another validated laboratory/commercial system. The benchmark is a measurement envelope study used to determine which webcam-derived outputs can be interpreted reliably. This approach follows established eye-tracking methods in which webcam or low-cost tracking systems are compared against laboratory systems using spatial accuracy, precision, calibration drift, area-of-interest (AOI) agreement, and task-output agreement [60,104,105]. The main benchmark outputs will be mean/median gaze error, root mean square (RMS) precision or fixation dispersion, calibration drift, AOI classification agreement, antisaccade correct/error agreement, pursuit gain/correlation agreement and free-viewing grid agreement. Indicative technical and measurement performance targets are calibration success ≥ 85–90%, valid gaze samples ≥ 80%, AOI agreement ≥ 80%, pursuit-feature correlation with the reference system $r \geq 0.70$, and test–retest ICC ≥ 0.75 for prespecified individual oculomotor measurements. Reliability estimates will be reported with 95% confidence intervals. These provisional targets will guide further validation; clinical interpretability will require separate evaluation.

Quality gating will determine whether individual task measurements are eligible for analysis. Sessions will be classified as valid, partially valid or invalid using prespecified criteria for calibration, effective sampling rate, valid-gaze percentage and task completion. Partially valid sessions may contribute only measurements from tasks meeting all applicable quality criteria, with exclusions documented. Invalid sessions will be excluded from measurement analyses and will prompt repeat testing or technical review. All attempted study sessions will contribute to feasibility and data quality reporting. The initial study will report individual task measurements and quality indicators without generating domain scores, GEPS or GEDS.

Repeated sessions during clinically stable periods will assess the test–retest reliability of individual oculomotor measurements. Variance components models will estimate between-participant and within-participant variability, with systematic session effects examined separately. Intraclass correlation coefficients, retest differences and retest standard deviations (RetestSD) will be reported with uncertainty intervals, as detailed in File S3, Section S3.4.4. These preliminary estimates will inform subsequent study design; they will not establish validated individual-level reliable change thresholds or demonstrate sensitivity to clinical deterioration.

3.3.3. Future GEPS/GEDS Derivation, Validation and Clinical Anchoring

The initial study will evaluate individual oculomotor measurements; GEPS/GEDS development is reserved for larger, separately planned studies. Variable selection, transformations, weights, adjustment parameters and candidate thresholds will be derived in designated development samples. The complete scoring specification, including preprocessing, missing data and quality rules, will be fixed before evaluation in adequately sized independent validation samples. Participants and their repeated

observations must not overlap between derivation and validation. If an adequate independent sample is unavailable, validation will be deferred and the scores will remain developmental.

Future GEPS reference limits will be estimated from prespecified percentiles of appropriately characterized healthy reference distributions, kept distinct from clinical group distributions. GEDS reliable change limits will be derived from stable-state repeated measurements of the fixed score, accounting for measurement variability and the baseline definition. Both require independent evaluation before individual interpretation (File S3, Section S3.4.2). Diagnostic discrimination, ROC/AUC analyses and diagnostic cutoff selection are outside the initial study.

GEDS clinical anchoring will require a separate longitudinal study with prespecified follow-up and independently assessed clinical change. Repeated measurements will be analyzed using mixed effects models incorporating actual time since baseline and participant-level random effects, with the model form matched to the score distribution. Relevant clinical and technical covariates, and sensitivity analyses concerning missingness and symptom-dependent assessment schedules, will be prespecified (File S3, Section S3.4.4). Change exceeding measurement variability will not, by itself, establish clinical deterioration.

The proposed GEPS/GEDS framework should be understood as a future structured approach to organizing oculomotor information for clinician interpretation. The individual domains included in GOBLIN Sight are already supported by the clinical and experimental literature as relevant to schizophrenia spectrum disorders. However, each marker captures only one aspect of a broader oculomotor phenotype and is vulnerable to task noise, technical variability, medication effects, fatigue, arousal and other state-dependent confounders. A composite approach is therefore methodologically plausible because it preserves domain-level information while reducing dependence on any single variable. Neither score is designed to diagnose psychosis, predict relapse or trigger treatment changes independently; both are intended to add an objective, structured layer of information to clinician interpretation.

3.3.4. Acceptability and Feasibility

Acceptability and feasibility will be evaluated before broader clinical validation, because technical performance alone is insufficient for a digital system intended for use in psychosis-spectrum populations. Even if the eye-tracking pipeline produces valid data under controlled conditions, the module may still be clinically unsuitable if the task battery is too long, cognitively demanding, frustrating, fatiguing, suspiciousness-provoking, or difficult to integrate into routine care. Therefore, GOBLIN Sight will be tested not only as a technical measurement system, but also as a user-facing clinical workflow. Acceptability testing will include three user groups: healthy controls, clinically stable psychosis spectrum participants, and clinicians/operators. Healthy controls will provide an initial estimate of task burden and technical usability in non-clinical users. Clinically stable psychosis spectrum participants will provide the key feasibility signal for the intended clinical population. Marked suspiciousness toward the system, and persecutory content concerning surveillance, represent postponement or exclusion criteria for the webcam module. Clinicians and operators will evaluate whether the outputs are interpretable and whether the test can be realistically integrated into clinical practice. Participants with acute agitation, severe disorganization, imminent risk, or inability to consent will not be included in the early acceptability stage.

The acceptability study will evaluate the complete GOBLIN Sight workflow: onboarding, instruction comprehension, calibration, antisaccade task, smooth pursuit task, free-viewing task, blink-baseline task, post-test questionnaire, and clinician/operator

review. The full battery will first be evaluated as the baseline protocol. If burden is excessive, a shorter follow-up battery may be derived empirically by identifying the tasks with the best balance between technical validity, clinical interpretability, and patient tolerability.

Patient acceptability will be assessed immediately after the task battery using short visual analogue scales and structured questions. The minimum questionnaire will include perceived difficulty, fatigue, distress, frustration, intrusiveness, willingness to repeat, and preference for a full or shortened battery. Pre-test and post-test distress will be collected to determine whether the procedure itself increases discomfort. In addition, participants will be asked an open-ended question about what was confusing, unpleasant, or difficult. This qualitative feedback will be coded into practical categories such as calibration difficulty, eye strain, task boredom, suspiciousness, fatigue, misunderstanding of instructions, or environmental discomfort. Clinician and operator feasibility will be assessed separately. This is necessary because a test may be acceptable to patients but still impractical if it requires excessive staff support, troubleshooting, or interpretation time. Operators will record calibration attempts, repeated instructions, environmental adjustments, technical interruptions, and reasons for incomplete sessions. Clinicians will rate whether the outputs are understandable, whether the quality flags are helpful, whether the presentation of individual oculomotor measurements and quality indicators are clinically interpretable, and whether the workflow could fit into routine outpatient or inpatient follow-up.

3.4. GOBLIN Voice Validation Methodology

GOBLIN Voice prospective validation aims to determine whether the clinician-supervised voice-synthesis and voice-conversion system can reconstruct a perceptually plausible representation of the patient's hallucinated voice, maintain intelligible interactive conversion within acceptable latency, and be operated reliably by clinicians. The detailed technical specification is provided in Section 2, while expanded validation procedures are provided in Supplementary File S3. No validation results are reported in the present manuscript.

3.4.1. Technical Evaluation of the Voice Module

For the planned automated evaluation of the voice tuning process, a simulation of the user walkthrough, similar to the one described in the framework by Tian et al. [106] will be used. An ECAPA-TDNN speaker embedding extractor [80], available in the SpeechBrain toolkit [107] or trained on the VoxSim dataset [108], is used to extract vector representations of the target voice to be reproduced. These embeddings are compared against those of the refined voice to obtain a cosine similarity score at each step. For each iteration, the algorithm explores a number of N values for the current slider and chooses to apply the one with the highest cosine similarity before moving on to the next slider, where the process is repeated. The cosine similarity is monitored over a number of M iterations in order to determine approximately how much time is required to obtain a reasonably good representation for the voice and what similarity score it achieved. This experiment will serve as a quantitative measure to evaluate the method, as it can be automated and does not require human interaction. To establish a reference baseline, we will compute the average similarity between pairs of utterances produced by the same speaker under the same acoustic conditions. The similarity achieved by the proposed system will then be interpreted relative to this baseline, as well as relative to the average starting point before any transformation is applied.

The voice mask validation will distinguish between technical pipeline performance and patient-facing perceptual outcomes. Technical performance will be evaluated

separately in terms of speech intelligibility and latency. Speech intelligibility will be quantified using the word error rate of the transformed output relative to the therapist's input speech, as chunk-based processing may introduce distortions or omissions. As described in Section 2.2.2, this risk is mitigated through voice activity detection. Latency will be measured independently as the interval between the onset of the therapist's recorded speech segment and the onset of playback of the transformed output, including the maximum observed delay under the tested operating conditions.

3.4.2. Acceptability and Feasibility

Following the demonstration of sufficient technical readiness, acceptability, usability and feasibility testing will evaluate whether GOBLIN Voice can be operated reliably by clinicians and tolerated by intended users within a clinician-supervised clinical workflow, without excessive technical burden, emotional distress, cognitive load or disruption of the therapeutic interaction. This stage is distinct from technical benchmarking and is not intended to demonstrate therapeutic efficacy.

Because the target voice may represent an internal auditory experience rather than an externally recorded speaker, participant-facing assessment cannot establish objective acoustic equivalence between the generated and hallucinated voices. Instead, voice resemblance will be evaluated as a subjective perceptual outcome, including participant-rated similarity, realism and fidelity to the perceived characteristics of the distressing voice, together with its tolerability and usability for clinician-supervised avatar dialogue. Technical accuracy will remain evaluated separately under controlled conditions using recorded target voices.

The acceptability and feasibility stage includes two user groups: clinicians/operators and, only after technical readiness has been demonstrated, a small clinician-supervised sample of clinically stable participants experiencing persistent AVHs. Participants with acute agitation, imminent risk, severe disorganization, inability to provide informed consent, or inability to follow task instructions will not be included in early acceptability testing. This stage is intended as a clinician-supervised feasibility and tolerability assessment, not as a definitive therapeutic trial.

Feasibility is evaluated through workflow-completion and implementation indicators. These include eligibility-to-enrolment rate, reasons for refusal, the completion of voice characterization, the completion of voice personalization, time required for customization, number of generated samples or adjustments, completion of the streaming voice mask trial, technical interruptions, troubleshooting time, and resource requirements. Voice characterization completion of at least 80% is considered acceptable in the clinician-supervised feasibility stage. Personalization completion of approximately 70–80% is considered acceptable, with reasons for non-completion coded. Streaming voice mask completion of at least 80% for core test utterances will be considered acceptable. Major technical failures, such as software crashes, synthesis failures, severe artifacts, or role-switch failures, should remain below 5%, with recoverable issues logged and reviewed for workflow improvement.

Usability is assessed separately for clinicians and participants. Clinicians will complete the System Usability Scale [109] and feasibility-oriented measures such as the Feasibility of Intervention Measure [110], alongside workload assessment using the NASA Task Load Index [111]. Clinician-facing usability will also include training time, troubleshooting time, total workflow time, confidence in saved voice configurations and quality-control flags, ease of switching between natural and transformed voice, perceived usefulness for clinician-supervised avatar dialogue, and willingness to use the system again. A SUS score of 68 or above will be treated as acceptable, while scores of 80 or above

will be considered aspirational. Failure to meet predefined usability or workload criteria will trigger workflow revision before further testing.

Participant acceptability focuses on whether the process is understandable, tolerable, and meaningful enough to repeat. Patient-facing items will assess perceived resemblance between the generated voice and the distressing voice, whether the voice is realistic enough for clinician-supervised dialogue, the difficulty of the personalization process, distress elicited by hearing the generated voice, whether the voice feels too realistic or emotionally overwhelming, intrusiveness or privacy-related discomfort, and willingness to repeat the procedure at a future visit. These items will be measured using visual analogue scales, yes/no or unsure responses, and short free-text comments. A mean post-session distress rating of 3/10 or lower, absence of clinically significant distress increase, mean intrusiveness/discomfort ratings of 3/10 or lower, and willingness to repeat in at least 70% of participants will be considered favorable feasibility signals.

Qualitative feedback is collected through brief semi-structured interviews or free-text questions from both participants and clinicians. These data will be used to identify which elements facilitate or hinder engagement, including the clarity of voice characterization instructions, the perceived realism of the generated voice, the emotional burden of the personalization process, clinician confidence, workflow fit, and the balance between sufficient voice presence and participant safety. If the full assessment battery proves too burdensome, a shorter follow-up battery may be derived empirically for later stages.

Safety and tolerability will be monitored throughout voice personalization and clinician-mediated dialogue because the procedure may involve exposure to distressing voice content and interaction with a digital representation of the hallucinated voice. The clinician will observe for signs of clinically significant distress, including requests to discontinue, increasing withdrawal or unusually brief responses compared with the participant's usual communication style, visible anxiety, fear, nervousness or trembling, and verbal indications of discomfort, being overwhelmed or counterproductive engagement with the voice (e.g., "I'll do whatever you say" or "I feel I can't express myself anymore"). Sessions will be paused in the presence of clinically significant distress, agitation, dissociation, suspiciousness, inability to engage, or whenever the clinician considers continuation inappropriate. All interruptions, early terminations, adverse events and safety-related actions will be documented.

If clinically significant distress occurs, the personalized voice will be discontinued and the clinician will return to their natural voice to re-establish a supportive therapeutic environment. The participant's emotional state and appraisal of the voice will then be reassessed, and appropriate anxiety-reducing or supportive measures will be provided. If distress returns to a tolerable level and the participant agrees, the avatar voice may be reintroduced within the same session, thereby avoiding unnecessary termination that could reinforce avoidance. The session will be ended if the participant requests discontinuation, if clinically significant distress does not resolve with these measures, or if the clinician judges further exposure inappropriate; the reason will be documented and the session rescheduled where appropriate.

Potential adverse events will also be monitored during and between sessions and classified according to severity and their plausible relationship to the intervention. These may include acute emotional distress such as anxiety, excessive worry, mood instability or irritability; disturbances in sleep or appetite; concentration difficulties, fatigue or feeling overwhelmed; withdrawal from usual activities; negative changes in the participant's relationship with the hallucinated voice, including increased frequency, perceived dominance or emotional impact; and physical symptoms without an alternative explanation. Participants will be encouraged to report unexpected experiences between

sessions. Reported events will be assessed for their temporal relationship to the intervention while acknowledging that some may reflect the natural course of the underlying disorder, and may prompt the modification of the study procedure where appropriate. For this early stage, the acceptability criterion will be zero serious system-related adverse events, with distress-related stopping events remaining below approximately 10–15%.

To ensure that these procedures are applied consistently across participants, all clinicians delivering the intervention will undergo standardized training prior to study initiation, distinguishing technical operation from therapeutic delivery. The research and development team will introduce the study aims and provide detailed instruction on hardware setup and software operation, including voice recording, voice personalization, and modification procedures. Clinicians will also receive standardized guidance on recruiting eligible participants and obtaining the information required for voice creation. For therapeutic training, we plan to establish a partnership with accredited psychotherapists to support the development and adaptation of AVATAR-inspired techniques and train clinicians in their application. This component would address the rationale behind the voice-based intervention, communication with participants, transitions between therapist and avatar roles, monitoring distress, and responding appropriately to adverse events. The partnership is also intended to provide ongoing clinical supervision, including case discussions and support with difficulties encountered during therapeutic delivery. The training content and supervision arrangements remain to be finalized with the collaborating psychotherapists before participant-facing studies. Where language-model support is implemented and included in the study protocol, technical training will also cover its use alongside the voice conversion component. The research team will remain available throughout the study period to address questions, concerns, or unexpected operational problems, alongside the planned therapeutic supervision.

Together, these safeguards are intended to promote participant safety, ensure consistent implementation of the intervention and facilitate the early identification and appropriate management of any issues arising during the study.

Overall, this validation stage determines whether GOBLIN Voice is acceptable, usable, and feasible enough to progress toward later clinical-context studies. Progression will depend on evidence that clinicians can operate the system with manageable workload, participants can complete the personalization and mask-exposure procedures without excessive distress, the workflow can be delivered with realistic clinical resources, and all technical and safety issues can be adequately logged, reviewed, and corrected.

3.5. GOBLIN Assist Validation Methodology

GOBLIN Assist will be evaluated as a clinician-facing, source-grounded, auditable, and non-autonomous information-support module. The proposed validation will assess technical data capture, source-linked draft summarization, candidate safety-related text-signal generation, clinician verification, and workflow feasibility. It will not evaluate the module as an autonomous monitoring, risk-assessment, treatment-recommendation, or therapeutic system.

3.5.1. GOBLIN Assist Technical Validation

The automated evaluation phase focuses on the performance of the machine learning components independently of therapeutic outcomes.

End-to-end data flow. Using manually-constructed test cases and synthetic patient journeys, pass/fail tests will evaluate whether patient inputs, clinician-authored prompts,

session recordings, transcripts, safety-related review flags, source-linked draft summaries, and clinician-facing draft suggestions are stored under the correct user/session identity.

Scheduled notification and response capture. Using simulated schedules and accounting for edge cases, the scheduled delivery of clinician-authored questions and patient answer collection will be evaluated first in a pass/fail regime (that the events happen as prescribed), then further assessed for appropriately small latency.

ASR module performance. Evaluation will be performed in terms ofseveral dimensions: transcription accuracy, diarization accuracy, robustness to environmental noise, and timestamp quality. Respectively, these will entail the following.

a. Transcription accuracy will be evaluated on the Romanian speech benchmark introduced by [112]. In particular, performance will be assessed based on Word Error Rate (WER) and our model's placement against the benchmarked candidates of the aforementioned paper.
b. Speaker diarization accuracy will be evaluated on SDBench introduced by [113]. In particular, performance will be assessed based on Diarization Error Rate (DER) andspeaker confusion rate (SCR);
c. Transcription robustness will be evaluated on a noisy version of the Romanian speech benchmark. Noise is added at controlled Signal-to-Noise ratios, modeling up to moderate noise levels typical of a medical office. Different types of noise will be tested, such as white, pink and babble noise. Absolute WER as well as relative WER (against clean recordings) will be used for the evaluation and further tuning of the system.

Emotion-related and safety-related text-signal detection performance. The module will be evaluated according to standard metrics in the field [114], macro-F1 and micro-F1 scores, the Hamming loss, per-label precision and recall. Class imbalance will be addressed using stratified sampling and class-weighted loss, with decision thresholds selected on a validation set against a prespecified sensitivity floor for the positive class. Performance will be reported using precision–recall curves, per-class precision and recall, and false negative rates rather than accuracy. Any later change of threshold (for example, because of review flag burden observed in feasibility) is a versioned re-derivation re-validated against the same sensitivity floor and documented before use, never an in-place adjustment during a study. The evaluation of the presence of language associated with suicidal ideation in textual patient responses will be conducted according to the protocol reported by [115]. Validation will further place an emphasis on having low false-negative rates for the presence of suicidal ideation, as well as negative emotions such as sadness, fear, shame, and anger. This approach, combined with the presentation of classifier output to the clinician without automatic binary conclusions, is intended to reduce the risk of clinically relevant symptoms being overlooked. Evaluation will additionally report sensitivity separately for the original ReachOut tiers so that performance on lower-severity positive categories is not obscured by binary aggregation. The false-negative rate for crisis-tier examples will be treated as a primary safety metric.

In addition to aggregate performance metrics, a clinician-authored Romanian behavioral test suite will evaluate language phenomena that may be particularly relevant to safety-related interpretation. The suite will include minimal or closely matched examples addressing negation ("nu vreau să mor" versus "vreau să mor"), hypothetical and quoted statements, temporal framing, ambiguity, sarcasm, idiomatic or metaphorical language, indirect expressions of risk and command-hallucination content. Performance will be reported separately for each capability, and recurrent failure patterns will drive targeted data augmentation. The classifier output remains a review-prioritization signal, and every submission is reviewed in full within the review window, so the consequence of a false negative is bounded by the review interval rather than by the classifier.

https://doi.org/10.3390/informatics13110180

External validation will be performed on a prospectively collected, ethics-approved, de-identified Romanian corpus (patient responses to clinician-authored questions and session transcripts). Samples will be annotated independently by two psychiatrists with adjudication. Inter-rater agreement (Cohen's κ) will be reported. The model will be locked before evaluation, and the corpus will be separated by site or by time from any development data. Performance on this corpus, not on translated public data, is the validation endpoint.

Performance will additionally be reported descriptively, with confidence intervals, across prespecified subgroups including sex, age band, diagnostic subgroup, input modality (typed versus ASR-derived), the presence or absence of diacritics and text length. Although performance estimates and confidence intervals will be reported for each subgroup, the sample size is unlikely to provide sufficient statistical power to detect or confirm differences in performance between subgroups.

LLM assistant for clinician support. We reiterate that the generative LLM component is validated exclusively for clinician-facing use and does not interact directly with patients. Each of the module's different functionalities will undergo a two-stage evaluation. Since several of the proposed functionalities rely on general-purpose conversational language modeling capabilities, evaluation is first conducted using established, generic LLM benchmarks, followed by domain-specific ones for assessing performance in clinically relevant scenarios.

1. Summarization ability is a core functionality supporting clinicians throughout patient evaluation. The model is first evaluated against public generic content and dialogue summarization datasets (Romanian-translated, where appropriate) such as RoLargeSum [116], XL-Sum [117], SAMSum [118], QMSum [119] and DialogSum [94]. Technical validation on medical evidence documents and clinical dialogue is undertaken following the methodology of Tang et al.[120] and Fraille Navarro et al. [121]. In all scenarios, industry-standard scores of ROUGE 1/2/L and BERTScore will be used. Clinical summaries are further subject to human evaluation on a 1-5 Likert scale following the UniEval scheme [122], with additional checks for Romanian fluency, clinical relevance, completeness and presence of hallucinated content, with the latter being marked solely with a binary flag.
2. Instruction-following ability will be evaluated according to the methodology and benchmarks proposed by [123] using an LLM judge.
3. Machine reading comprehension (MRC) will be evaluated using exact match and token-level F1 scores on dedicated benchmarks such as XQuAD-r [124], RoITD [125], RoMed [126]and RoBiologyDataChoiceQ [127].
4. Clinician-facing suggestions for therapeutic interventions will be evaluated on internally-created benchmarks consisting of specialist-created mock therapeutic situations and crisis intervention red team prompts, rated by clinicians on a 1-5 Likert scale for safety, empathy, usefulness, clinical appropriateness and Romanian language appropriateness. The following dimensions are considered: suggestions for routine therapeutic scenarios and scenarios involving escalating safety concerns, suggestions for AVATAR-style dialogue interventions, patient data question answering, and LLM boundary testing (e.g., refusal to respond or follow an unsafe instruction or an out-of-scope medical demand).

Further domain-specific validation will be conducted on the basis of an established list of permitted outputs, restricting the system to source-grounded extraction, summarization, longitudinal comparison, and clinician review flags, while prohibiting new diagnoses, treatment recommendations, and medication advice. First, a small Romanian-language validation set will be created from representative patient responses and therapy session transcripts, with psychiatrist-authored reference summaries and

annotated supporting evidence. The model-generated summaries will be evaluated for factual consistency, clinically relevant omissions, unsupported or contradicted claims, and exact citation accuracy, with all patient quotations verified deterministically against the source text. Second, adversarial testing will assess whether the model produces prohibited diagnostic, treatment, or medication-related recommendations under direct and indirect prompting. Adversarial clinician prompts will include explicitly asking for diagnostic, medication, or treatment protocol suggestions, with the acceptance criterion being set to a strict refusal to respond in each of these cases. Third, prompt-injection resistance will be evaluated using attacks embedded in patient text, clinician speech, medical notes, metadata, and retrieved records, with attack success, policy violations, citation bypass, and system-prompt disclosure reported separately. Finally, the module will undergo multidisciplinary clinical red-teaming involving psychiatrists, ML researchers, security specialists, and privacy personnel. Testing will target subtle diagnostic inference, implied treatment advice, symptom overstatement, temporal errors, speaker misattribution, negation failures, unsupported causality, and excessive confidence under incomplete evidence. All identified failures will be documented according to severity, reproducibility, affected model version, and remediation status.

3.5.2. Acceptability and Feasibility

Acceptability, feasibility and usability testing evaluates whether GOBLIN Assist can be realistically integrated into clinician-supervised schizophrenia care after technical readiness has been demonstrated. This phase evaluates whether the inter-appointment monitoring system, clinician dashboard, summarization tools and avatar therapy assistance functions can be implemented within routine care without excessive burden, disruption of workflow or safety concerns. Two user groups will participate in this stage: clinicians and clinically stable individuals experiencing persistent auditory verbal hallucinations. Participants presenting acute agitation, severe disorganization, inability to provide informed consent or imminent risk requiring immediate intervention will not be included in early feasibility testing.

Feasibility will be assessed through implementation and workflow indicators. These include recruitment and retention rates, completion rates of scheduled inter-appointment assessments, the proportion of successfully delivered clinician-authored prompts, the proportion of patient responses successfully processed and stored, completion of clinician review procedures, technical interruptions and resource requirements. Additional indicators will include time required for clinicians to configure individualized monitoring questions, time required to review generated summaries, number of review prioritization flags generated, proportion of flagged submissions reviewed by clinicians and the frequency of manual correction of generated outputs. Completion of at least 70% of scheduled patient assessments, successful processing of at least 90% of submitted responses and clinician review completion rates of at least 80% will be considered indicative of acceptable feasibility. Major technical failures, including unsuccessful data capture, missing summaries or incorrect alert routing, should remain below 5%.

Usability will be assessed separately for clinicians and participants. Clinicians will complete the System Usability Scale (SUS) [109], the Feasibility of Intervention Measure (FIM) [110] and workload assessments using the NASA Task Load Index (NASA-TLX) [111]. Clinician-specific usability measures will additionally evaluate the system's ease of navigating longitudinal patient records, ease of retrieving previous hallucination content, clarity of generated summaries, interpretability of review prioritization flags, usefulness of avatar therapy suggestions, confidence in system outputs and willingness to use the platform in future practice. Patient-facing usability measures will focus on ease of understanding clinician-authored questions, ease of submitting responses, clarity of

automated acknowledgements, perceived burden of participation and overall usability of the mobile interface. A mean SUS score of 68 or greater will be considered indicative of acceptable usability, while scores above 80 will be considered excellent.

Patient acceptability will focus on the inter-appointment monitoring component. Measures will evaluate ease of understanding the questions, perceived burden of responding, comfort with data collection, perceived helpfulness of symptom logging, willingness to continue using the system and perceived appropriateness of automated responses. Clinician acceptability will evaluate satisfaction with summary quality, perceived usefulness of longitudinal symptom tracking and confidence in system outputs. Acceptability will be assessed using the Acceptability of Intervention Measure (AIM), supplemented by brief Likert-scale ratings and qualitative feedback. Mean AIM scores of at least 3.5/5 and willingness to continue use in at least 70% of participants will be considered favorable indicators of acceptability.

Qualitative feedback from patients and clinicians will identify barriers to and facilitators of implementation. Patient feedback will cover missed check-ins, discontinuation, privacy concerns, question clarity, usefulness of symptom logging, and the appropriateness of the automated acknowledgement. Clinician feedback will address dashboard usability, review-flag burden, manual correction burden, workflow fit, and the usefulness of longitudinal symptom summaries.

Adverse events and tolerability concerns will be actively monitored through participant reports, direct questioning, clinician observation, safety logs, and the review of discontinuation reasons. Mild events, such as transient distress, suspiciousness, frustration, notification-related irritation, disengagement, or increased symptom focus, will be recorded even if they do not meet the criteria for serious adverse events.

Safety monitoring is integrated throughout the validation process because the system processes information related to psychotic symptoms, suicidal ideation and other potentially distressing experiences. The system does not provide diagnoses, treatment recommendations, crisis counseling or autonomous clinical decision-making. All review flags, summaries and generated suggestions remain subject to clinician review before any clinical action is taken. Safety outcomes will include inappropriate system outputs, false-negative and false-positive review flags, retrieval errors, participant complaints, early discontinuation, adverse events and any instances in which generated content is considered potentially misleading or clinically inappropriate. The acceptability threshold for this stage will be zero serious system-related adverse events.

4. Discussion

4.1. Positioning of GOBLIN in Psychosis Care

This paper presents the GOBLIN prototype architecture as a clinician-supervised modular platform intended to integrate three elements: webcam-based eye-tracking for structured oculomotor assessment, an avatar-based module for distressing voices, and a digital assistant layer for documentation and follow-up. Conceptually, GOBLIN draws momentum from several active research lines. In particular, our contribution is the proposal of an integrated architecture along with a structured roadmap for evaluating its technical reliability, feasibility, acceptability, safety, and eventual clinical utility within a clinician-supervised workflow. The intended role of the platform is adjunctive rather than autonomous, positioned in the continuum of psychosis care as an assistant tool. Its intended future role is to support continuity between visits, organize information that is otherwise collected intermittently, and support clinician judgment rather than replace psychiatric assessment or treatment planning. This positioning is consistent with broader literature, where the urgency caused by increasing psychosis burden and limits of current

pharmacological and non-pharmacological management accelerates the discussion and adoption of digital tools as powerful extensions of care pathways [128].

Conceptually, GOBLIN is well-positioned relative to existing digital psychosis tools that have already been developed for psychosis care, including AVATAR-based interventions for AVHs, VR-assisted therapies, digitally supported CBT, smartphone-based relapse monitoring and moderated online psychosocial platforms. AVATAR therapy systems (presented previously) provide the closest therapeutic precedent, using a digital representation of the distressing voice to modify the patient–voice relationship and reduce perceived voice dominance. VR systems such as gameChange target avoidance, anxiety and paranoia through immersive exposure-based scenarios [129], while tools such as SlowMo and Actissist provide structured digital or blended CBT support for paranoia, reasoning biases and coping [130,131]. Other platforms, including ClinTouch, CrossCheck and EMPOWER, focus mainly on symptom monitoring, ecological momentary assessment or relapse early-warning detection [132–134], whereas Horyzons and PRIME emphasize psychosocial recovery, motivation, peer/community engagement and moderated digital support [135,136]. In this context, the proposed contribution of GOBLIN is architectural: the integration of exploratory oculomotor measurement, AVATAR-inspired voice interaction and LLM-supported clinical summarization within a single clinician-supervised workflow.

4.2. Clinical Use Cases and Novelty

The main contribution of GOBLIN is that it is designed to be evaluated against two gaps in psychosis care at the same time: persistent, distressing voices and poor visibility of change between appointments. AVATAR-based work has already shown that voice-focused dialogue can reduce distress [24], while digital monitoring studies suggest that repeated patient-generated data can help capture clinically meaningful change outside the consultation room [137]. GOBLIN's novelty lies in integrating these functions in the same clinician-supervised pathway instead of treating them as separate tools. The potential clinical significance is practical rather than theoretical. In routine care, worsening is often recognized late, especially when insight is poor, reporting is inconsistent, or follow-up is fragmented [128]. If technically and clinically validated, such an integrated platform could capture structured behavioral and patient-reported information, preserve longitudinal context, and present source-grounded summaries to the treating psychiatrist. Whether this improves the timing or focus of clinical review requires prospective evaluation.

4.3. Technical Novelty

A core component of the technical novelty brought by GOBLIN to AVATAR therapy-informed approaches is its hybrid voice conversion architecture, which is specifically designed to improve fidelity and address the challenges of cross-gender voice conversion discussed previously. While prior AVATAR-based systems rely on parametric voice transformation techniques such as adjusting pitch, vocal tract size, spectral tilt and temporal roughness, the proposed system extends this paradigm by combining the state-of-the-art machine learning-based voice cloning technology CosyVoice 3 with voice embedding engineering and classical signal processing operations. This hybrid approach is intended to allow finer control over the generated voice characteristics, potentially improving perceptual similarity between the converted voice used during the therapy session and the hallucinated voice described by the user. Additionally, it may improve speech naturalness and reduce perceptible artifacts.

A methodological feature of the proposed GOBLIN Sight framework is that its reference distributions and scoring parameters could be recalibrated as additional validation data become available. Any such recalibration would require dataset

documentation, version control, locked-model testing, and renewed technical and clinical validation. This flexibility should not be interpreted as established scalability, improved accuracy, validated subgroup performance, or an adaptive clinical capability.

4.4. Limitations

The limitations of the GOBLIN system span three interrelated domains: constraints arising from individual technical components, therapeutic and population-level barriers, and regulatory and governance requirements. The therapeutic component carries both eligibility and contextual limitations. The dialogue-based intervention presupposes a capacity for structured, sustained engagement that may not be present in acutely disorganized, agitated, or severely inattentive patients; those experiencing florid psychosis or marked thought disorder are unlikely to benefit and should not be considered candidates [138]. The technique additionally requires the patient to identify and characterize a dominant distressing voice, which presents practical difficulties for individuals whose voices are multiple, amorphous, or experienced primarily as a non-verbal presence [32]. A further technical constraint pertains to the temporal coherence of streaming voice transformation. The intrinsic end-to-end latency introduced by voice-morphing pipelines can undermine the rhythmic and interpersonal synchrony of therapeutic dialogue, rendering the conversational flow perceptually mechanical and diminishing the avatar's credibility as a responsive interlocutor. This is especially salient given that the sense of presence—the patient's subjective perception of the digital entity as a lifelike embodiment—is increasingly recognized as an integral mechanism of action in facilitating therapeutic outcomes [33]. An additional area requiring empirical study is whether the lack of a visual avatar impacts therapeutic effectiveness, a relationship that has not been systematically investigated in the current literature. At the population level, cultural adaptation remains in its early stages, with no evidence yet available from low- and middle-income country settings [139]. Practical barriers, including limited technology literacy, concerns about digital surveillance, restricted internet access, device availability, language diversity, and low literacy, may exclude a substantial proportion of eligible patients. The preferences of patients who favor non-digital alternatives must be respected, and treatment choice should be preserved throughout [140].

The primary limitations of the GOBLIN Sight oculomotor measurement module concern the interpretation of its outputs at the individual patient level. Eye-tracking data alone cannot function as a stand-alone classifier capable of reliably distinguishing between acute and remitted psychosis. It is more appropriately conceptualized as a candidate adjunctive phenotyping and longitudinal monitoring framework requiring prospective validation, rather than an autonomous indicator of clinical decompensation. Translating group-level oculomotor findings into actionable individual-level clinical decisions is inherently difficult, as oculomotor indices exhibit significant inter-individual heterogeneity and do not consistently correlate linearly with symptom severity. Single-patient interpretation is further complicated by confounding variables including antipsychotic medication effects, sedation, fluctuating task engagement, motion artifacts, and the intrinsic heterogeneity of psychotic illness. A further limitation is that the proposed oculomotor domain and composite scores, including GEPS and GEDS, are not implemented in the current prototype and have not been clinically validated. Although eye-tracking abnormalities in schizophrenia are well documented at a group level, this does not establish that webcam-based eye-tracking can diagnose psychosis, predict relapse, detect treatment non-adherence, or provide an individual-level biomarker in routine care. GEPS and GEDS should therefore be understood as proposed exploratory candidate scores for future derivation and validation. Their reliability, clinical interpretability, sensitivity to medication and fatigue, robustness across devices, and

longitudinal association with clinician-rated deterioration remain to be established prospectively. The scores' derivation is therefore deferred to later studies, after initial validation of individual oculomotor measurements. The webcam-based nature of the module also introduces technical and contextual limitations. Eye-tracking quality may be affected by lighting, camera resolution, frame rate, screen size, viewing distance, visual correction, glasses glare, head position, calibration quality, participant engagement, cognitive impairment, sedation, sleep deprivation and missing data. These factors may influence the derived oculomotor variables independently of psychosis-related mechanisms.

Limitations of the LLM-module concern reliability, safety and governance. Because model outputs depend on patient-reported hallucination data and clinician-authored information, underreporting, ambiguous descriptions and patients' impaired insight may impact the quality of the information utilized in the construction of the clinician summaries. In addition, LLM-based systems may generate inaccurate, overgeneralized, or poorly calibrated outputs, and safety-related text-signal models remain vulnerable to both false-negative and false-positive review flags. For this reason, GOBLIN Assist cannot function as an autonomous evaluator of risk, a diagnostic system, or an independent treatment recommender. All outputs must remain clinician-facing, source-grounded, source-linked, auditable, editable, and subject to clinician verification before any clinical use.

4.5. Regulatory Framework, Risk Management and Governance

We further discuss safety and governance constraints. Within the regulatory framework of the EU AI Act, of which several articles are yet to come into effect, it is explicitly recognized that AI systems operating within the health sector can pose substantive risks to user safety, necessitating rigorous precautionary measures to mitigate potential harm [141]. Accordingly, GOBLIN must be designed and operated as a clinician-in-the-loop system at all times. Neither the voice therapy literature nor the biomarker literature provides a basis for fully autonomous operation independent of clinical oversight, and the platform must not be deployed as such [142]. A comprehensive and continuously updated systemic safety architecture is required throughout the system's lifecycle, incorporating clearly defined protocols for session interruption, escalation pathways to a supervising clinician, and immediate takeover procedures for situations involving acute distress or self-harm ideation. Residual risks must be documented and mitigation measures validated prior to deployment [142,143]. The platform must maintain sufficient accuracy, sensorimotor robustness, and optimal response speeds to preserve the temporal coherence and therapeutic integrity of the clinical dialogue. This includes technical measures to manage errors or inconsistencies that may emerge from the human-system interaction, ensuring that the interpersonal frame of the intervention remains stable [144]. In light of the high sensitivity of psychosis-related health data, strict privacy-by-design principles are applied throughout the system. These include rigorous data minimization, the application of pseudonymization where feasible, robust access controls, and secure storage, while maintaining full auditability of the session logic to ensure clinical accountability [145]. Future development will be undertaken in accordance with the General Data Protection Regulation (GDPR) with specific consideration given to lawful processing, purpose limitation, data minimization, storage limitation, technical and organizational security measures, and the protection of data subject rights [146].

Another important consideration is system-level risk management. Even with a clinician-in-the-loop design, residual risks remain. These include invalid eye-tracking data being misinterpreted, calibration failure, false reassurance from apparently stable oculomotor outputs or future composite scores, excessive distress during voice

personalization, the reinforcement of delusional content, LLM summarization errors, missed safety-related text signals and false-positive review flags, clinician overreliance on automated outputs and data-governance failures. For this reason, GOBLIN requires ongoing residual-risk review rather than one-time validation. Operationally, the proposed clinical-stage system would be designed to maintain audit logs for technical failures, invalid Sight sessions, calibration attempts, missing data, any future implemented score outputs, safety-related review flags, clinician acknowledgements, clinician overrides, LLM outputs, source references, edited summaries, Voice interruptions, patient distress ratings and adverse events. These logs are to be reviewed periodically by the clinical and technical team to identify recurrent failure modes, update quality-control thresholds, refine escalation rules and document version changes. Any modification to scoring logic, safety-related text-signal logic, prompt structure, voice-generation parameters or safety procedures should be versioned and traceable. This governance model preserves the intended role of GOBLIN as an adjunctive clinical-support system. The platform may organize information, generate candidate signals and support structured interaction, but clinical meaning remains dependent on human interpretation, context and longitudinal psychiatric assessment. The completeness, reliability, and usability of these audit mechanisms remain to be verified prospectively.

The present Technical Note reports no human-participant study, participant-derived technical or clinical results, clinical deployment, patient intervention, therapeutic outcomes, diagnostic performance, or validated safety outcomes. Its contribution is limited to the description of a prototype software architecture and a prospective validation roadmap. Any future evaluation involving patients, healthy volunteers, clinicians, identifiable voice recordings, eye-tracking data, clinical interviews, therapeutic sessions, or clinical outcomes will require prospective institutional ethics approval or a formally documented determination, as applicable, before recruitment or data collection. Such studies will use study-specific informed-consent procedures, appropriate safeguards for vulnerable participants, purpose-limited data processing, controlled access, predefined retention periods, and explicit provisions governing data sharing and future reuse.

4.6. Future Directions

To the best of our knowledge, the specific combination of exploratory oculomotor measurement, clinician-supervised AVATAR-inspired voice interaction, and source-grounded LLM-assisted clinical documentation has not previously been described within a single integrated platform architecture for psychosis care. The principal novelty of GOBLIN therefore lies in the proposed integration of these complementary components within a clinician-controlled and auditable workflow, rather than in any currently established clinical capability. Future work will determine whether the integrated prototype can achieve adequate technical reliability, usability, acceptability, safety, source-grounding, and clinical interpretability. Any therapeutic benefit, biomarker validity, relapse-related utility, or broader clinical effectiveness would require subsequent prospective validation.

5. Conclusions

Auditory verbal hallucinations are a clinically significant phenomenon that can occur across several psychiatric and non-psychiatric conditions, including schizophrenia. Among individuals with schizophrenia who experience AVHs, a subset develop persistent hallucinations [147] that can remain clinically challenging and insufficiently addressed by existing treatment paradigms. This technical note presents GOBLIN as an integrated prototype architecture combining voice-based interaction, artificial intelligence, and biometric measurements, together with a staged roadmap for prospective validation. The proposed final system design and functionality are intended to address structural limitations in current care models, including limited accessibility, intermittent assessment, and fragmented follow-up. The present manuscript does not provide evidence of clinical effectiveness, therapeutic efficacy, diagnostic performance, relapse-prediction performance, or validated safety outcomes. These capabilities require sequential technical, usability, safety, and clinical validation before clinical implementation can be considered, and remain as future work. At its current stage, GOBLIN should therefore be understood as an investigational, clinician-supervised platform with potential to complement existing therapeutic modalities and support more personalized, longitudinal, and data-informed clinician-supervised psychosis care. The extent to which this potential can be realized in clinical practice, however, remains to be established.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/informatics13110180/s1; File S1: GOBLIN Sight technical specification; File S2: Functional examples of GOBLIN Assist; File S3: GOBLIN Prospective validation methodology. References [148–163] are only cited in the Supplementary Materials.

Author Contributions: Conceptualization, Z.M., R.-C.L., M.-L.V., I.-A.B. and R.S.; methodology, M.-L.V., I.-A.B., M.-S.O. and R.S.; software, R.-G.B., A.-F.E. and A.-A.N.; writing—original draft preparation, M.-L.V., R.-G.B., A.-F.E., A.-A.N., M.-S.O., A.-G.S., E.-G.H.-H. and R.S.; writing—review and editing, M.-L.V. and R.S.; visualization, M.-L.V. and R.S.; supervision, Z.M. and R.-C.L.; project administration, Z.M. and R.-C.L.; funding acquisition, Z.M. and R.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research and project were financed by the Romanian Government through the POCIDIF program 2021–2027 from the European Regional Development Fund of the European Union, under grant(SMIS) number 336094, GOBLIN AI SRL Beneficiary.

Institutional Review Board Statement: Not applicable. The manuscript is limited to the description of a prototype software architecture and a prospective validation roadmap and reports no human-participant study, participant-derived analysis, clinical intervention, or clinical outcomes.

Informed Consent Statement: Not applicable. No human-participant data, voice recordings, or participant-derived results are included or analyzed in the revised manuscript.

Data Availability Statement: No original human-participant dataset is reported or analyzed in this manuscript, and no participant-derived data support the claims presented in this Technical Note. Public data, pretrained models, and external resources referenced in the technical description are identified and cited in the manuscript.

Acknowledgments: Generative-AI tools were used exclusively for post-production language editing and graphic layout. They were not used to generate scientific content, technical data, results, interpretations or conclusions.

Conflicts of Interest: The funders (Romanian Government) had no role in the conceptualization and technical implementation of the system, in the writing of the manuscript or in the decision to submit the manuscript for publication. As part of the grant agreement, GOBLIN AI SRL received funding

for industrial research and experimental development. No company profit was involved. Dr. Milhem and Mr. Lache are associated with GOBLIN AI SRL and were involved in initial conceptualization, project administration and supervision. They did not participate in the technical implementation or in the formulation of the manuscript's conclusions or prospective research directions. The manuscript presents a prototype and conceptual framework only, and no claims of established clinical performance or commercial superiority are made. The other authors declare no conflicts of interest.

References

1. Auditory Hallucinations. In *Handbook of Clinical Neurology*; Elsevier: Amsterdam, The Netherlands. 2015; Volume 129, pp. 433–455, ISBN 978-0-444-62630-1.
2. Maijer, K.; Begemann, M.J.H.; Palmen, S.J.M.C.; Leucht, S.; Sommer, I.E.C. Auditory Hallucinations across the Lifespan: A Systematic Review and Meta-Analysis. *Psychol. Med.* **2018**, *48*, 879–888. https://doi.org/10.1017/S0033291717002367.
3. Johns, L.C.; Kompus, K.; Connell, M.; Humpston, C.; Lincoln, T.M.; Longden, E.; Preti, A.; Alderson-Day, B.; Badcock, J.C.; Cella, M.; et al. Auditory Verbal Hallucinations in Persons With and Without a Need for Care. *Schizophr. Bull.* **2014**, *40*, S255–S264. https://doi.org/10.1093/schbul/sbu005.
4. Upthegrove, R.; Broome, M.R.; Caldwell, K.; Ives, J.; Oyebode, F.; Wood, S.J. Understanding Auditory Verbal Hallucinations: A Systematic Review of Current Evidence. *Acta Psychiatr Scand* **2016**, *133*, 352–367. https://doi.org/10.1111/acps.12531.
5. Jauhar, S.; Johnstone, M.; McKenna, P.J. Schizophrenia. *Lancet* **2022**, *399*, 473–486. https://doi.org/10.1016/S0140-6736(21)01730-X.
6. Cobandag, M.; Sigala, N. The Effectiveness of Non-Pharmacological Treatments for Auditory Verbal Hallucinations in Schizophrenia Spectrum Disorders: A Systematic Review and Meta-Analysis. *Eur. Psychiatr.* **2025**, *68*, e155. https://doi.org/10.1192/j.eurpsy.2025.10115.
7. Lim, A.; Hoek, H.W.; Deen, M.L.; Blom, J.D.; Bruggeman, R.; Cahn, W.; De Haan, L.; Kahn, R.S.; Meijer, C.J.; Myin-Germeys, I.; et al. Prevalence and Classification of Hallucinations in Multiple Sensory Modalities in Schizophrenia Spectrum Disorders. *Schizophr. Res.* **2016**, *176*, 493–499. https://doi.org/10.1016/j.schres.2016.06.010.
8. Dyck, M.S.; Mathiak, K.A.; Bergert, S.; Sarkheil, P.; Koush, Y.; Alawi, E.M.; Zvyagintsev, M.; Gaebler, A.J.; Shergill, S.S.; Mathiak, K. Targeting Treatment-Resistant Auditory Verbal Hallucinations in Schizophrenia with fMRI-Based Neurofeedback—Exploring Different Cases of Schizophrenia. *Front. Psychiatry* **2016**, *7*, 37. https://doi.org/10.3389/fpsyt.2016.00037.
9. Sommer, I.E.C.; Slotema, C.W.; Daskalakis, Z.J.; Derks, E.M.; Blom, J.D.; Van Der Gaag, M. The Treatment of Hallucinations in Schizophrenia Spectrum Disorders. *Schizophrenia Bulletin* **2012**, *38*, 704–714. https://doi.org/10.1093/schbul/sbs034.
10. McCutcheon, R.A.; Pillinger, T.; Varvari, I.; Halstead, S.; Ayinde, O.O.; Crossley, N.A.; Correll, C.U.; Hahn, M.; Howes, O.D.; Kane, J.M.; et al. INTEGRATE: International Guidelines for the Algorithmic Treatment of Schizophrenia. *Lancet Psychiatry* **2025**, *12*, 384–394. https://doi.org/10.1016/S2215-0366(25)00031-8. Correction in *Lancet Psychiatry* **2025**, *12*, e9. https://doi.org/10.1016/S2215-0366(25)00131-2.
11. Shergill, S.S.; Murray, R.M.; McGuire, P.K. Auditory Hallucinations: A Review of Psychological Treatments. *Schizophr. Res.* **1998**, *32*, 137–150. https://doi.org/10.1016/S0920-9964(98)00052-8.
12. Ceraso, A.; Lin, J.J.; Schneider-Thoma, J.; Siafis, S.; Tardy, M.; Komossa, K.; Heres, S.; Kissling, W.; Davis, J.M.; Leucht, S. Maintenance Treatment with Antipsychotic Drugs for Schizophrenia. *Cochrane Database Syst. Rev.* **2020**, *8*, CD008016. https://doi.org/10.1002/14651858.CD008016.pub3.
13. Keepers, G.A.; Fochtmann, L.J.; Anzia, J.M.; Benjamin, S.; Lyness, J.M.; Mojtabai, R.; Servis, M.; Walaszek, A.; Buckley, P.; Lenzenweger, M.F.; et al. The American Psychiatric Association Practice Guideline for the Treatment of Patients With Schizophrenia. *Am. J. Psychiatry* **2020**, *177*, 868–872.
14. Bighelli, I.; Salanti, G.; Huhn, M.; Schneider-Thoma, J.; Krause, M.; Reitmeir, C.; Wallis, S.; Schwermann, F.; Pitschel-Walz, G.; Barbui, C.; et al. Psychological Interventions to Reduce Positive Symptoms in Schizophrenia: Systematic Review and Network Meta-analysis. *World Psychiatry* **2018**, *17*, 316–329. https://doi.org/10.1002/wps.20577.
15. Berendsen, S.; Berendse, S.; van der Torren, J.; Vermeulen, J.; de Haan, L. Cognitive Behavioural Therapy for the Treatment of Schizophrenia Spectrum Disorders: An Umbrella Review of Meta-Analyses of Randomised Controlled Trials. *EClinicalMedicine* **2024**, *67*, 102392. https://doi.org/10.1016/j.eclinm.2023.102392.

16. Gleeson, J.F.; McGuckian, T.B.; Fernandez, D.K.; Fraser, M.I.; Pepe, A.; Taskis, R.; Alvarez-Jimenez, M.; Farhall, J.F.; Gumley, A. Systematic Review of Early Warning Signs of Relapse and Behavioural Antecedents of Symptom Worsening in People Living with Schizophrenia Spectrum Disorders. *Clin. Psychol. Rev.* **2024**, *107*, 102357. https://doi.org/10.1016/j.cpr.2023.102357.
17. Ellory, I.; Malasi, J.; Wood, L.; McGuire, P.; Middleton, A.; Persaud, K. *Digital Health Technologies to Help Manage Symptoms of Psychosis and Prevent Relapse in Adults and Young People: Early Value Assessment*; National Institute for Health and Care Excellence: Manchester, UK, 2024.
18. Hayward, M.; Jones, A.-M.; Bogen-Johnston, L.; Thomas, N.; Strauss, C. Relating Therapy for Distressing Auditory Hallucinations: A Pilot Randomized Controlled Trial. *Schizophr. Res.* **2017**, *183*, 137–142. https://doi.org/10.1016/j.schres.2016.11.019.
19. Longden, E.; Corstens, D.; Bowe, S.; Pyle, M.; Emsley, R.; Peters, S.; Branitsky, A.; Chauhan, N.; Dehmahdi, N.; Jones, W.; et al. A Psychological Intervention for Engaging Dialogically with Auditory Hallucinations (Talking With Voices): A Single-Site, Randomised Controlled Feasibility Trial. *Schizophr. Res.* **2022**, *250*, 172–179. https://doi.org/10.1016/j.schres.2022.11.007.
20. Thomas, N.; Ward, T.; Longden, E.; Bell, I.H.; Brand, R.M.; Christensen, M.J.; Glenthøj, L.B.; Rus-Calafell, M.; Hayward, M. Relational Therapies for People Who Hear Voices: Operationalisation and Current Status of an Emergent Group of Psychological Therapies. *Schizophr. Bull.* **2026**, *52*, sbaf142. https://doi.org/10.1093/schbul/sbaf142.
21. Trower, P.; Birchwood, M.; Meaden, A.; Byrne, S.; Nelson, A.; Ross, K. Cognitive Therapy for Command Hallucinations: Randomised Controlled Trial. *Br. J. Psychiatry* **2004**, *184*, 312–320. https://doi.org/10.1192/bjp.184.4.312.
22. Ward, T.; Rus-Calafell, M.; Ramadhan, Z.; Soumelidou, O.; Fornells-Ambrojo, M.; Garety, P.; Craig, T.K.J. AVATAR Therapy for Distressing Voices: A Comprehensive Account of Therapeutic Targets. *Schizophr. Bull.* **2020**, *46*, 1038–1044. https://doi.org/10.1093/schbul/sbaa061.
23. Craig, T.K.; Rus-Calafell, M.; Ward, T.; Leff, J.P.; Huckvale, M.; Howarth, E.; Emsley, R.; Garety, P.A. AVATAR Therapy for Auditory Verbal Hallucinations in People with Psychosis: A Single-Blind, Randomised Controlled Trial. *Lancet Psychiatry* **2018**, *5*, 31–40. https://doi.org/10.1016/S2215-0366(17)30427-3. Erratum in *Lancet Psychiatry* **2018**, *5*, e1. https://doi.org/10.1016/S2215-0366(17)30473-X.
24. Garety, P.A.; Edwards, C.J.; Jafari, H.; Emsley, R.; Huckvale, M.; Rus-Calafell, M.; Fornells-Ambrojo, M.; Gumley, A.; Haddock, G.; Bucci, S.; et al. Digital AVATAR Therapy for Distressing Voices in Psychosis: The Phase 2/3 AVATAR2 Trial. *Nat. Med.* **2024**, *30*, 3658–3668. https://doi.org/10.1038/s41591-024-03252-8. Correction in *Nat. Med.* **2026**, *32*, 2689. https://doi.org/10.1038/s41591-026-04540-1.
25. Smith, L.C.; Vernal, D.L.; Mariegaard, L.S.; Christensen, A.G.; Jansen, J.E.; Schytte, G.; Stokbro, L.A.; Albert, N.; Christensen, M.J.; Thomas, N.; et al. Immersive Virtual Reality-Assisted Therapy Targeting Persistent Auditory Verbal Hallucinations in Patients Diagnosed with Schizophrenia Spectrum Disorders in Denmark: The Challenge Assessor-Masked, Randomised Clinical Trial. *Lancet Psychiatry* **2025**, *12*, 557–567. https://doi.org/10.1016/S2215-0366(25)00161-0.
26. Dellazizzo, L.; Potvin, S.; Phraxayavong, K.; Dumais, A. One-Year Randomized Trial Comparing Virtual Reality-Assisted Therapy to Cognitive-Behavioral Therapy for Patients with Treatment-Resistant Schizophrenia. *npj Schizophr.* **2021**, *7*, 9. https://doi.org/10.1038/s41537-021-00139-2.
27. O'Driscoll, G.A.; Callahan, B.L. Smooth Pursuit in Schizophrenia: A Meta-Analytic Review of Research since 1993. *Brain Cogn.* **2008**, *68*, 359–370. https://doi.org/10.1016/j.bandc.2008.08.023.
28. Wolf, A.; Ueda, K.; Hirano, Y. Recent Updates of Eye Movement Abnormalities in Patients with Schizophrenia: A Scoping Review. *Psychiatry Clin. Neurosci.* **2021**, *75*, 82–100. https://doi.org/10.1111/pcn.13188.
29. Kallimani, D.; Theleritis, C.; Evdokimidis, I.; Stefanis, N.C.; Chatzimanolis, I.; Smyrnis, N. The Effect of Change in Clinical State on Eye Movement Dysfunction in Schizophrenia. *Eur. Psychiatry* **2009**, *24*, 17–26. https://doi.org/10.1016/j.eurpsy.2008.08.003.
30. Leff, J.; Williams, G.; Huckvale, M.; Arbuthnot, M.; Leff, A.P. Avatar Therapy for Persecutory Auditory Hallucinations: What Is It and How Does It Work? *Psychosis* **2014**, *6*, 166–176. https://doi.org/10.1080/17522439.2013.773457.
31. Leff, J.; Williams, G.; Huckvale, M.A.; Arbuthnot, M.; Leff, A.P. Computer-Assisted Therapy for Medication-Resistant Auditory Hallucinations: Proof-of-Concept Study. *Br. J. Psychiatry* **2013**, *202*, 428–433. https://doi.org/10.1192/bjp.bp.112.124883.
32. Ward, T.; Lister, R.; Fornells-Ambrojo, M.; Rus-Calafell, M.; Edwards, C.J.; O'Brien, C.; Craig, T.K.; Garety, P. The Role of Characterisation in Everyday Voice Engagement and AVATAR Therapy Dialogue. *Psychol. Med. 52*, 3846–3853. https://doi.org/10.1017/S0033291721000659.
33. Rus-Calafell, M.; Ward, T.; Zhang, X.C.; Edwards, C.J.; Garety, P.; Craig, T. The Role of Sense of Voice Presence and Anxiety Reduction in AVATAR Therapy. *J. Clin. Med.* **2020**, *9*, 2748. https://doi.org/10.3390/jcm9092748.

34. Schubert, T.; Friedmann, F.; Regenbrecht, H. The Experience of Presence: Factor Analytic Insights. *Presence Teleoperators Virtual Environ.* **2001**, *10*, 266–281. https://doi.org/10.1162/105474601300343603.
35. Glenthøj, L.B.; Mariegaard, L.S.; Smith, L.C.; Christensen, M.J.; Jensen, S.N.; Rus-Calafell, M.; Thomas, N.; Nordentoft, M.; Vernal, D.L. Impact of Avatar Features and Presence on Treatment Outcomes in Virtual Reality-Assisted Therapy for Auditory Hallucinations. *Schizophr. Res.* **2026**, *287*, 1–2. https://doi.org/10.1016/j.schres.2025.11.004.
36. Kilic, O.; Riecher-Rössler, A.; Galderisi, S.; Gorwood, P.; Frangou, S.; Pinto da Costa, M. The Role of Gender as a Barrier to the Professional Development of Psychiatrists. *Eur. Psychiatry* **2023**, *66*, e89. https://doi.org/10.1192/j.eurpsy.2023.2462.
37. Betker, J. Better Speech Synthesis through Scaling. *arXiv* **2023**, arXiv:2305.07243.
38. Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Adi, V.M.Y.; Hsu, J.M.W.-N. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale.
39. Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; Bian, J. NaturalSpeech 2: Latent Diffusion Models Are Natural and Zero-Shot Speech and Singing Synthesizers 2023. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
40. Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; et al. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. *arXiv* **2024**, arXiv:2412.10117.
41. Du, Z.; Gao, C.; Wang, Y.; Yu, F.; Zhao, T.; Wang, H.; Lv, X.; Wang, H.; Ni, C.; Shi, X.; et al. CosyVoice 3: Towards In-the-Wild Speech Generation via Scaling-up and Post-Training. *arXiv* **2025**, arXiv:2505.17589.
42. Liu, S. Zero-Shot Voice Conversion with Diffusion Transformers. *arXiv* **2024**, arXiv:2411.0994.
43. Li, J.; Tu, W.; Xiao, L. Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
44. Chatterbox: Open Source Text-to-Speech|Resemble AI. Available online: https://www.resemble.ai/learn/models/chatterbox (accessed on 13 October 2026).
45. Ma, G.; Yao, J.; Ning, Z.; Jiang, Y.; Xiong, L.; Xie, L.; Zhu, P. MeanVC: Lightweight and Streaming Zero-Shot Voice Conversion via Mean Flows 2025. In Proceedings of the 2026 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 3–8 May 2026.
46. Stylianou, Y. Voice Transformation: A Survey. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 3585–3588.
47. Tian, Y.; Liu, J.; Lee, T. User-Driven Voice Generation and Editing through Latent Space Navigation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024.
48. Harrison, P.; Marjieh, R.; Adolfi, F.; van Rijn, P.; Anglada-Tort, M.; Tchernichovski, O.; Larrouy-Maestri, P.; Jacoby, N. Gibbs Sampling with People. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 10659–10671.
49. Mertes, S.; Don, D.W.; Grothe, O.; Kuch, J.; Schlagowski, R.; André, E. VoiceX: A Text-To-Speech Framework for Custom Voices 2024.
50. Reilly, J.L.; Frankovich, K.; Hill, S.; Gershon, E.S.; Keefe, R.S.E.; Keshavan, M.S.; Pearlson, G.D.; Tamminga, C.A.; Sweeney, J.A. Elevated Antisaccade Error Rate as an Intermediate Phenotype for Psychosis Across Diagnostic Categories. *Schizophr. Bull.* **2014**, *40*, 1011–1021. https://doi.org/10.1093/schbul/sbt132.
51. Meyhoefer, I.; Sprenger, A.; Derad, D.; Grotegerd, D.; Leenings, R.; Leehr, E.J.; Breuer, F.; Surmann, M.; Rolfes, K.; Arolt, V.; et al. Evidence from Comprehensive Independent Validation Studies for Smooth Pursuit Dysfunction as a Sensorimotor Biomarker for Psychosis. *Sci. Rep.* **2024**, *14*, 13859. https://doi.org/10.1038/s41598-024-64487-6.
52. Zhang, L.; Zhang, X.; Fang, X.; Zhou, C.; Wen, L.; Pan, X.; Zhang, F.; Chen, J. Eye Movement Characteristics in Male Patients with Deficit and Non-Deficit Schizophrenia and Their Relationships with Psychiatric Symptoms and Cognitive Function. *BMC Neurosci.* **2021**, *22*, 70. https://doi.org/10.1186/s12868-021-00673-w.
53. Lakhlifi, M.; Laprevote, V.; Schwan, R.; Schwitzer, T. Free Viewing Exploration in Schizophrenia: Review of Evidence from Laboratory Settings to Natural Environment. *Encephale* **2020**, *46*, 115–122. https://doi.org/10.1016/j.encep.2019.11.012.
54. Nielsen, J.R.W.; Dietz, M.; Jefsen, O.H. Blink Rates in Patients with Schizophrenia Compared to Healthy Controls: A Meta-Analysis. *Schizophr. Res.* **2025**, *279*, 87–93. https://doi.org/10.1016/j.schres.2025.03.019.
55. Andrade, C. Z Scores, Standard Scores, and Composite Test Scores Explained. *Indian J. Psychol. Med.* **2021**, *43*, 555–557. https://doi.org/10.1177/02537176211046525.

56. 56. delCacho-Tena, A.; Christ, B.R.; Arango-Lasprilla, J.C.; Perrin, P.B.; Rivera, D.; Olabarrieta-Landa, L. Normative Data Estimation in Neuropsychological Tests: A Systematic Review. *Arch. Clin. Neuropsychol.* **2023**, *39*, 383–398. https://doi.org/10.1093/arclin/acad084.
57. Takahashi, J.; Miura, K.; Morita, K.; Fujimoto, M.; Miyata, S.; Okazaki, K.; Matsumoto, J.; Hasegawa, N.; Hirano, Y.; Yamamori, H.; et al. Effects of Age and Sex on Eye Movement Characteristics. *Neuropsychopharmacol. Rep.* **2021**, *41*, 152–158. https://doi.org/10.1002/npr2.12163.
58. Jonaitis, E.M.; Koscik, R.L.; Clark, L.R.; Ma, Y.; Betthauser, T.J.; Berman, S.E.; Allison, S.L.; Mueller, K.D.; Hermann, B.P.; Van Hulle, C.A.; et al. Measuring Longitudinal Cognition: Individual Tests versus Composites. *Alzheimer's Dement.* **2019**, *11*, 74–84. https://doi.org/10.1016/j.dadm.2018.11.006.
59. Morita, K.; Miura, K.; Fujimoto, M.; Yamamori, H.; Yasuda, Y.; Iwase, M.; Kasai, K.; Hashimoto, R. Eye Movement as a Biomarker of Schizophrenia: Using an Integrated Eye Movement Score. *Psychiatry Clin. Neurosci.* **2017**, *71*, 104–114. https://doi.org/10.1111/pcn.12460.
60. Kaduk, T.; Goeke, C.; Finger, H.; König, P. Webcam Eye Tracking Close to Laboratory Standards: Comparing a New Webcam-Based System and the EyeLink 1000. *Behav. Res. Methods* **2024**, *56*, 5002–5022. https://doi.org/10.3758/s13428-023-02237-8.
61. Patterson, A.S.; Nicklin, C.; Vitta, J.P. Methodological Recommendations for Webcam-Based Eye Tracking: A Scoping Review. *Res. Methods Appl. Linguist.* **2025**, *4*, 100244. https://doi.org/10.1016/j.rmal.2025.100244.
62. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines 2019. In Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019, Long Beach, CA, USA, 17–20 June 2019.
63. Che, S.E.; Gwon, Y.G.; Kim, K.-H. Follow-Up Timing After Discharge and Suicide Risk Among Patients Hospitalized With Psychiatric Illness. *JAMA Netw. Open* **2023**, *6*, e2336767. https://doi.org/10.1001/jamanetworkopen.2023.36767.
64. Dinakaran, D.; Sreeraj, V.S.; Venkatasubramanian, G. Measurement Based Care in Schizophrenia-Feasibility in Routine Clinical Practice. *Asian J. Psychiatr.* **2020**, *49*, 101954. https://doi.org/10.1016/j.ajp.2020.101954.
65. Helldin, L.; Mohn, C.; Olsson-Tall, M.; van Dijk-Härd, I.; Olsson, A.-K. Extended Lifetime Expectancy in Schizophrenia, Role of Measurement Based Care. *Arch. Psychiatr. Nurs.* **2024**, *53*, 151–157. https://doi.org/10.1016/j.apnu.2024.10.012.
66. Wright, A.C.; Browne, J.; Skiest, H.; Bhiku, K.; Baker, J.T.; Cather, C. The Relationship between Conventional Clinical Assessments and Momentary Assessments of Symptoms and Functioning in Schizophrenia Spectrum Disorders: A Systematic Review. *Schizophr. Res.* **2021**, *232*, 11–27. https://doi.org/10.1016/j.schres.2021.04.010.
67. Herold, C.J.; Lässer, M.M.; Schröder, J. Autobiographical Memory Impairment in Chronic Schizophrenia: Significance and Clinical Correlates. *J. Neuropsychol.* **2023**, *17*, 89–107. https://doi.org/10.1111/jnp.12288.
68. Ricarte, J.J.; Ros, L.; Latorre, J.M.; Watkins, E. Mapping Autobiographical Memory in Schizophrenia: Clinical Implications. *Clin Psychol. Rev.* **2017**, *51*, 96–108. https://doi.org/10.1016/j.cpr.2016.11.004.
69. Isaacson, M.; Hazell, C.M.; Cape, J.; Hickson, E.; Islam, F.; Gill, A.; Simon, K.; Patel, R.; Souray, J.; Raune, D. The Cognitive-Phenomenological Assessment of Delusions and Hallucinations at the Early Intervention in Psychosis Service Stage: The Results of a Quality Improvement Project. *Early Interv. Psychiatry* **2022**, *16*, 1345–1352. https://doi.org/10.1111/eip.13283.
70. Roebroek, L.O.; Bruins, J.; Boonstra, A.; Veling, W.; Jörg, F.; Sportel, B.E.; Delespaul, P.A.; Castelein, S. The Effects of a Computerized Clinical Decision Aid on Clinical Decision-Making in Psychosis Care. *J. Psychiatr. Res.* **2022**, *156*, 532–537. https://doi.org/10.1016/j.jpsychires.2022.10.053.
71. Yeung, J.A.; Dalmasso, J.; Foschini, L.; Dobson, R.J.; Kraljevic, Z. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. *arXiv* **2025**, arXiv:2509.10970.
72. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision 2022. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
73. Păis, V.; Mititelu, V.B.; Ion, R.; Irimia, E. Evaluating a Fine-Tuned Whisper Model on Underrepresented Romanian Speech. In Proceedings of the 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 25–27 October 2023; pp. 141–145.
74. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus 2020. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, 11–16 May 2020.

75. Wang, C.; Rivière, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; Dupoux, E. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation 2021. In Proceedings of the ACL 2021—59th Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 1–6 August 2021.
76. Stan, A.; Yamagishi, J.; King, S.; Aylett, M. The Romanian Speech Synthesis (RSS) Corpus: Building a High Quality HMM-Based Speech Synthesis System Using a High Sampling Rate. *Speech Commun.* **2011**, *53*, 442–450. https://doi.org/10.1016/j.specom.2010.12.002.
77. Păiş, V.; Ion, R.; Avram, A.-M.; Irimia, E.; Mititelu, V.B.; Mitrofan, M. Human-Machine Interaction Speech Corpus from the ROBIN Project. In Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 13–15 October 2021; pp. 91–96.
78. Tufiş, D.; Mititelu, V.B.; Irimia, E.; Păiş, V.; Ion, R.; Diewald, N.; Mitrofan, M.; Onofrei, M. Little strokes fell great oaks. Creating corola, the reference corpus of contemporary romanian. *Rev. Roum. Linguist.* **2019**, *64*, 225–226.
79. Google WebRTC Voice Activity Detection (VAD) Module. Available online: https://www.mathworks.com/matlabcentral/fileexchange/78895-google-webrtc-voice-activity-detection-vad-module (accessed on 31 March 2026).
80. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3830–3834.
81. Menon, A.K.; Rajagopalan, A.; Sumengen, B.; Citovsky, G.; Cao, Q.; Kumar, S. Online Hierarchical Clustering Approximations 2019. *arXiv* **2019**, arXiv:1909.09667.
82. Monath, N.; Zaheer, M.; McCallum, A. Online Level-Wise Hierarchical Clustering. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery, New York, NY, USA, 4 August 2023; pp. 1733–1745.
83. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale 2020. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020.
84. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization 2019. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.
85. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions 2020. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020.
86. Welivita, A.; Xie, Y.; Pu, P. A Large-Scale Dataset for Empathetic Response Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Moens, M.-F., Huang, X., Specia, L., Yih, S.W., Eds.; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 1251–1264.
87. Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200. https://doi.org/10.1080/02699939208411068.
88. Milne, D.N.; Pink, G.; Hachey, B.; Calvo, R.A. CLPsych 2016 Shared Task: Triaging Content in Online Peer-Support Forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*; Hollingshead, K., Ungar, L., Eds.; Association for Computational Linguistics: San Diego, CA, USA, June 2016; pp. 118–127.
89. Ji, S.; Yu, C.P.; Fung, S.; Pan, S.; Long, G. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity* **2018**, *2018*, 6157249. https://doi.org/10.1155/2018/6157249.
90. Broder, A.Z. On the Resemblance and Containment of Documents. In *Proceedings of the Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*; IEEE Computer Society: Salerno, Italy, 1998; pp. 21–29.
91. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report 2025.
92. Bulzan, A.-Ş.; Morariu, B.; Joldea, A.-R.; Cernăzanu-Glăvan, D.; Cernăzanu-Glăvan, C. GMTW-Ro: A Deterministic Benchmark for Evaluating Large Language Models on Grounded Romanian Tasks. *Front. Artif. Intell.* **2026**, *9*, 1831918. https://doi.org/10.3389/frai.2026.1831918.
93. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models 2021. In Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022), Online, 25–29 April 2022.
94. Chen, Y.; Liu, Y.; Chen, L.; Zhang, Y. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 5062–5074.

95. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text 2016. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, Texas, 1–4 November 2016.
96. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural Questions: A Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 452–466. https://doi.org/10.1162/tacl_a_00276.
97. Joshi, M.; Choi, E.; Weld, D.S.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension 2017. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017.
98. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. Available online: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm (accessed on 31 March 2026).
99. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback 2022. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022.
100. Gunter, T.; Wang, Z.; Wang, C.; Pang, R.; Narayanan, A.; Zhang, A.; Zhang, B.; Chen, C.; Chiu, C.-C.; Qiu, D.; et al. Apple Intelligence Foundation Language Models 2024. In Proceedings of the 39th Annual Conference on Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025.
101. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting Guideline for the Early Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *BMJ* **2022**, *377*, e070904. https://doi.org/10.1136/bmj-2022-070904.
102. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group. Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension. *Nat. Med.* **2020**, *26*, 1364–1374. https://doi.org/10.1038/s41591-020-1034-x.
103. Cruz Rivera, S.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J.; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension. *Nat. Med.* **2020**, *26*, 1351–1363. https://doi.org/10.1038/s41591-020-1037-7.
104. Dunn, M.J.; Alexander, R.G.; Amiebenomo, O.M.; Arblaster, G.; Atan, D.; Erichsen, J.T.; Ettinger, U.; Giardini, M.E.; Gilchrist, I.D.; Hamilton, R.; et al. Minimal Reporting Guideline for Research Involving Eye Tracking (2023 Edition). *Behav. Res. Methods* **2024**, *56*, 4351–4357. https://doi.org/10.3758/s13428-023-02187-1. Correction in *Behav. Res. Methods* **2024**, *56*, 4218–4219. https://doi.org/10.3758/s13428-024-02438-9
105. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Halszka, J.; van de Weijer, J. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; Oxford University Press: Oxford, UK, 2011; ISBN 978-0-19-969708-3.
106. Tian, Y.; Liu, J.; Lee, T. Personalized Voice Synthesis through Human-in-the-Loop Coordinate Descent. *arXiv* **2024**, arXiv:2408.17068.
107. Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; et al. SpeechBrain: A General-Purpose Speech Toolkit. *arXiv* **2021**, arXiv:2106.04624.
108. Ahn, J.; Kim, Y.; Choi, Y.; Kwak, D.; Kim, J.-H.; Mun, S.; Chung, J.S. VoxSim: A Perceptual Voice Similarity Dataset 2024. In Proceedings of the Annual Conference of the International Speech Communication Association, Kos, Greece, 1–5 September 2024.
109. Brooke, J. SUS: A Quick and Dirty Usability Scale. *Usability Eval. Ind.* **1995**, *189*, 4–7.
110. Weiner, B.J.; Lewis, C.C.; Stanick, C.; Powell, B.J.; Dorsey, C.N.; Clary, A.S.; Boynton, M.H.; Halko, H. Psychometric Assessment of Three Newly Developed Implementation Outcome Measures. *Implement. Sci.* **2017**, *12*, 108. https://doi.org/10.1186/s13012-017-0635-3.
111. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183, ISBN 978-0-444-70388-0.
112. Diaconu, A.; Vînaga, M.; Alexe, B. RO-N3WS: Enhancing Generalization in Low-Resource ASR with Diverse Romanian Speech Benchmarks 2026. *arXiv* **2026,** arXiv.2603.02368. https://doi.org/10.48550/arXiv.2603.02368.
113. Pacheco, E.; Orhon, A.; Durmus, B.; Munyampirwa, B.; Leonov, A. SDBench: A Comprehensive Benchmark Suite for Speaker Diarization 2025. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025.

114. Harutyunyan, A.; Kumar, S. Fine-Grained Emotion Detection on GoEmotions: Experimental Comparison of Classical Machine Learning, BiLSTM, and Transformer Models. *arXiv* **2026**, arXiv:2601.18162.
115. Chim, J.; Tsakalidis, A.; Gkoumas, D.; Atzil-Slonim, D.; Ophir, Y.; Zirikly, A.; Resnik, P.; Liakata, M. Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*; Yates, A., Desmet, B., Prud'hommeaux, E., Zirikly, A., Bedrick, S., MacAvaney, S., Bar, K., Ireland, M., Ophir, Y., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 177–190.
116. Avram, A.-M.; Timpuriu, M.; Iuga, A.; Matei, V.-C.; Taiatu, I.-M.; Găină, T.; Cercel, D.-C.; Cercel, M.-C.; Pop, F. RoLargeSum: A Large Dialect-Aware Romanian News Dataset for Summary, Headline, and Keyword Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 2049–2066.
117. Hasan, T.; Bhattacharjee, A.; Islam, M.S.; Mubasshir, K.; Li, Y.-F.; Kang, Y.-B.; Rahman, M.S.; Shahriyar, R. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Kerrville, TX, USA 2021; pp. 4693–4703.
118. Gliwa, B.; Mochol, I.; Biesek, M.; Wawer, A. SAMSum Corpus: A Human-Annotated Dialogue Dataset for Abstractive Summarization. In Proceedings of the 2nd Workshop on New Frontiers in Summarization, Hong Kong, China, 4 November 2019; pp. 70–79.
119. Zhong, M.; Yin, D.; Yu, T.; Zaidi, A.; Mutuma, M.; Jha, R.; Awadallah, A.H.; Celikyilmaz, A.; Liu, Y.; Qiu, X.; et al. QMSum: A New Benchmark for Query-Based Multi-Domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 5905–5921.
120. Tang, L.; Sun, Z.; Idnay, B.; Nestor, J.G.; Soroush, A.; Elias, P.A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J.F.; et al. Evaluating Large Language Models on Medical Evidence Summarization. *npj Digit. Med.* **2023**, *6*, 158. https://doi.org/10.1038/s41746-023-00896-7.
121. Fraile Navarro, D.; Coiera, E.; Hambly, T.W.; Triplett, Z.; Asif, N.; Susanto, A.; Chowdhury, A.; Azcoaga Lorenzo, A.; Dras, M.; Berkovsky, S. Expert Evaluation of Large Language Models for Clinical Dialogue Summarization. *Sci. Rep.* **2025**, *15*, 1195. https://doi.org/10.1038/s41598-024-84850-x.
122. Zhong, M.; Liu, Y.; Yin, D.; Mao, Y.; Jiao, Y.; Liu, P.; Zhu, C.; Ji, H.; Han, J. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2022; pp. 2023–2038.
123. Masala, M.; Ilie-Ablachim, D.C.; Dima, A.; Corlatescu, D.; Zavelca, M.; Olaru, O.; Terian, S.; Terian, A.; Leordeanu, M.; Velicu, H.; et al. "Vorbeşti Românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions 2024. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024.
124. Juletxara/Xquad_xtreme · Datasets at Hugging Face. Available online: https://huggingface.co/datasets/juletxara/xquad_xtreme (accessed on 5 June 2026).
125. Dragosnicolae555/RoITD · Datasets at Hugging Face. Available online: https://huggingface.co/datasets/dragosnicolae555/RoITD (accessed on 13 October 2026).
126. Rogoz, A.-C.; Ionescu, R.T.; Anghel, A.-V.; Antone-Iordache, I.-L.; Coniac, S.; Ionescu, A.I. A Large-Scale Benchmark for Evaluating Large Language Models on Medical Question Answering in Romanian. *npj Digit. Med.* **2026**, *9*, 268.
127. Ghinea, D.-D.; Corbeanu, A.-N.; Dumitran, M.-A. RoBiologyDataChoiceQA: A Romanian Dataset for Improving Biology Understanding of Large Language Models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*; Adelani, D.I., Arnett, C., Ataman, D., Chang, T.A., Gonen, H., Raja, R., Schmidt, F., Stap, D., Wang, J., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 551–567.
128. Smith, K.A.; Hardy, A.; Vinnikova, A.; Blease, C.; Milligan, L.; Hidalgo-Mazzei, D.; Lambe, S.; Marzano, L.; Uhlhaas, P.J.; Ostinelli, E.G.; et al. Digital Mental Health for Schizophrenia and Other Severe Mental Illnesses: An International Consensus on Current Challenges and Potential Solutions. *JMIR Mental Health* **2024**, *11*, e57155. https://doi.org/10.2196/57155.
129. Freeman, D.; Lambe, S.; Kabir, T.; Petit, A.; Rosebrock, L.; Yu, L.-M.; Dudley, R.; Chapman, K.; Morrison, A.; O'Regan, E.; et al. Automated Virtual Reality Therapy to Treat Agoraphobic Avoidance and Distress in Patients with Psychosis (gameChange): A

Multicentre, Parallel-Group, Single-Blind, Randomised, Controlled Trial in England with Mediation and Moderation Analyses. *Lancet Psychiatry* **2022**, *9*, 375–388. https://doi.org/10.1016/S2215-0366(22)00060-8.

130. Garety, P.; Ward, T.; Emsley, R.; Greenwood, K.; Freeman, D.; Fowler, D.; Kuipers, E.; Bebbington, P.; Dunn, G.; Hardy, A. *Digitally Supported CBT to Reduce Paranoia and Improve Reasoning for People with Schizophrenia-Spectrum Psychosis: The SlowMo RCT*; Efficacy and Mechanism Evaluation; NIHR Journals Library: Southampton, UK, 2021.
131. Bucci, S.; Berry, N.; Ainsworth, J.; Berry, K.; Edge, D.; Eisner, E.; Emsley, R.; Forbes, G.; Hassan, L.; Lewis, S.; et al. Effects of Actissist, a Digital Health Intervention for Early Psychosis: A Randomized Clinical Trial. *Psychiatry Res.* **2024**, *339*, 116025. https://doi.org/10.1016/j.psychres.2024.116025.
132. Palmier-Claus, J.E.; Ainsworth, J.; Machin, M.; Barrowclough, C.; Dunn, G.; Barkus, E.; Rogers, A.; Wykes, T.; Kapur, S.; Buchan, I.; et al. The Feasibility and Validity of Ambulatory Self-Report of Psychotic Symptoms Using a Smartphone Software Application. *BMC Psychiatry* **2012**, *12*, 172. https://doi.org/10.1186/1471-244X-12-172.
133. Ben-Zeev, D.; Brian, R.; Wang, R.; Wang, W.; Campbell, A.T.; Aung, M.S.H.; Merrill, M.; Tseng, V.W.S.; Choudhury, T.; Hauser, M.; et al. CrossCheck: Integrating Self-Report, Behavioral Sensing, and Smartphone Use to Identify Digital Indicators of Psychotic Relapse. *Psychiatr. Rehabil. J.* **2017**, *40*, 266–275. https://doi.org/10.1037/prj0000243.
134. Gumley, A.I.; Bradstreet, S.; Ainsworth, J.; Allan, S.; Alvarez-Jimenez, M.; Birchwood, M.; Briggs, A.; Bucci, S.; Cotton, S.; Engel, L.; et al. Digital Smartphone Intervention to Recognise and Manage Early Warning Signs in Schizophrenia to Prevent Relapse: The EMPOWER Feasibility Cluster RCT. *Health Technol. Assess* **2022**, *26*, 1–174. https://doi.org/10.3310/HLZE0479.
135. Schlosser, D.A.; Campellone, T.R.; Truong, B.; Etter, K.; Vergani, S.; Komaiko, K.; Vinogradov, S. Efficacy of PRIME, a Mobile App Intervention Designed to Improve Motivation in Young People With Schizophrenia. *Schizophr. Bull.* **2018**, *44*, 1010–1020. https://doi.org/10.1093/schbul/sby078.
136. Alvarez-Jimenez, M.; Koval, P.; Schmaal, L.; Bendall, S.; O'Sullivan, S.; Cagliarini, D.; D'Alfonso, S.; Rice, S.; Valentine, L.; Penn, D.L.; et al. The Horyzons Project: A Randomized Controlled Trial of a Novel Online Social Therapy to Maintain Treatment Effects from Specialist First-Episode Psychosis Services. *World Psychiatry* **2021**, *20*, 233–243. https://doi.org/10.1002/wps.20858.
137. Vecchio, I.; Mifsud, L.; Castro e Almeida, S.; Passecker, J. Diagnostic Digital Phenotyping in Schizophrenia-Spectrum Disorders: A Systematic Review. *npj Digit. Med.* **2025**, *9*, 16. https://doi.org/10.1038/s41746-025-02194-w.
138. Aali, G.; Kariotis, T.; Shokraneh, F. Avatar Therapy for People with Schizophrenia or Related Disorders. *Cochrane Database Syst. Rev.* **2020**, *2020*, CD011898. https://doi.org/10.1002/14651858.CD011898.pub2.
139. Creative Facilitation Testimonial—AVATAR3. Available online: https://medhumsplatform.org/creative-facilitation-unit/creative-facilitation-testimonial-avatar3/(accessed on 31 March 2026).
140. Dennard, S. How Do People with Psychosis Experience and Interact with Digital Mental Health Interventions? [Internet]. AMETHYST; 2024 Aug 23. Available Online: https://www.voicestherapy.com/post/how-do-people-with-psychosis-experience-and-interact-with-digital-mental-health-interventions (accessed on 31 March 2026).
141. Finck, M. Recital 47. In *EU Artificial Intelligence Act*; Oxford University Press: Oxford, UK, 2026.
142. Moratinos, G.L. Human Oversight (Article 14). In *EU Artificial Intelligence Act*; Wolters Kluwer/CEDAM: Milan, Italy, 2025.
143. Andrés, M.B. Risk Management System (Article 9). In *EU Artificial Intelligence Act*; La Ley, Soluciones Legales: Las Rozas, Madrid, 2024.
144. J. E. Constantino Torres. Accuracy, Robustness and Cybersecurity (Article 15). In *EU Artificial Intelligence Act*; Wolters Kluwer: Alphen aan den Rijn, The Netherlands, 2024.
145. Esposito, M.S. Data and Data Governance (Article 10). In *EU Artificial Intelligence Act*; Wolters Kluwer/CEDAM: Milan, Italy, 2025.
146. Regulation-2016/679-EN-Gdpr-EUR-Lex. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed on 28 August 2025).
147. Köhler-Forsberg, O.; Madsen, T.; Behrendt-Møller, I.; Nordentoft, M. The 10-Year Trajectories of Auditory Hallucinations among 496 Patients with a First Schizophrenia-Spectrum Disorder: Findings from the OPUS Cohort. *Schizophr. Res.* **2022**, *243*, 385–391. https://doi.org/10.1016/j.schres.2021.06.033.
148. Goldsack, J.C.; Coravos, A.; Bakker, J.P.; Bent, B.; Dowling, A.V.; Fitzer-Attas, C.; Godfrey, A.; Godino, J.G.; Gujar, N.; Izmailova, E.; et al. Verification, Analytical Validation, and Clinical Validation (V3): The Foundation of Determining Fit-for-Purpose for Biometric Monitoring Technologies (BioMeTs). *npj Digit. Med.* **2020**, *3*, 55. https://doi.org/10.1038/s41746-020-0260-4.
149. Bowen, D.J.; Kreuter, M.; Spring, B.; Cofta-Woerpel, L.; Linnan, L.; Weiner, D.; Bakken, S.; Kaplan, C.P.; Squiers, L.; Fabrizio, C.; et al. How We Design Feasibility Studies. *Am. J. Prev. Med.* **2009**, *36*, 452–457. https://doi.org/10.1016/j.amepre.2009.02.002.

150. Eldridge, S.M.; Chan, C.L.; Campbell, M.J.; Bond, C.M.; Hopewell, S.; Thabane, L.; Lancaster, G.A. CONSORT 2010 Statement: Extension to Randomised Pilot and Feasibility Trials. *BMJ* **2016**, *355*, i5239. https://doi.org/10.1136/bmj.i5239.
151. Sekhon, M.; Cartwright, M.; Francis, J.J. Acceptability of Healthcare Interventions: An Overview of Reviews and Development of a Theoretical Framework. *BMC Health Serv. Res.* **2017**, *17*, 88. https://doi.org/10.1186/s12913-017-2031-8.
152. Sekhon, M.; Cartwright, M.; Francis, J.J. Development of a Theory-Informed Questionnaire to Assess the Acceptability of Healthcare Interventions. *BMC Health Serv. Res.* **2022**, *22*, 279. https://doi.org/10.1186/s12913-022-07577-3.
153. Jacobson, N.S.; Truax, P. Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *J. Consult Clin. Psychol.* **1991**, *59*, 12–19. https://doi.org/10.1037//0022-006x.59.1.12.
154. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* **1986**, *1*, 307–310.
155. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012. Erratum in *J. Chiropr. Med.* **2017**, *16*, 346. https://doi.org/10.1016/j.jcm.2017.10.001.
156. Stoyanov, S.R.; Hides, L.; Kavanagh, D.J.; Wilson, H. Development and Validation of the User Version of the Mobile Application Rating Scale (uMARS). *JMIR mHealth uHealth* **2016**, *4*, e72. https://doi.org/10.2196/mhealth.5849.
157. Eisner, E.; Richardson, C.; Thomas, N.; Rus-Calafell, M.; Syrett, S.; Firth, J.; Gumley, A.; Hardy, A.; Allan, S.; Kabir, T.; et al. Measurement of Adverse Events in Studies of Digital Health Interventions for Psychosis: Guidance and Recommendations Based on a Literature Search and Framework Analysis of Standard Operating Procedures. *Schizophr. Bull.* **2024**, *50*, 1456–1470. https://doi.org/10.1093/schbul/sbae048.
158. Garety, P.; Edwards, C.J.; Ward, T.; Emsley, R.; Huckvale, M.; McCrone, P.; Rus-Calafell, M.; Fornells-Ambrojo, M.; Gumley, A.; Haddock, G.; et al. Optimising AVATAR Therapy for People Who Hear Distressing Voices: Study Protocol for the AVATAR2 Multi-Centre Randomised Controlled Trial. *Trials* **2021**, *22*, 366. https://doi.org/10.1186/s13063-021-05301-w. Correction in *Trials* **2024**, *25*, 803. https://doi.org/10.1186/s13063-024-08549-0.
159. Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science **1986**, *1*, 54–75.
160. Rubin, D.B. Multiple Imputation for Nonresponse in Surveys. In *Wiley Series in Probability and Statistics*; 1st ed.; Wiley: Hoboken, NJ, USA, 1987; ISBN 978-0-471-08705-2.
161. Haynes, W. Benjamini–Hochberg Method. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013; p. 78, ISBN 978-1-4419-9862-0.
162. Benjamini, Y.; Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Statist.* **2001**, *29*, 1165–1188. https://doi.org/10.1214/aos/1013699998.
163. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.