technologies

MDPI

Article

Designing Trustworthy AI Systems for PTSD Follow-Up

María Cazares [1,†], Jorge Miño-Ayala [2,†], Iván Ortiz [3,*,†] and Roberto Andrade [4,†]

[1] Escola de Psicologia e Ciências da Vida, Universidade Lusófona de Humanidades e Tecnologias, Campo Grande, 1749-024 Lisbon, Portugal; a22411590@alunos.ulht.pt
[2] Facultad de Ingeniería de Sistemas, Escuela Politécnica Nacional, Quito 170525, Ecuador; jorge.mino@epn.edu.ec
[3] Escuela de Ingeniería en Ciberseguridad, Facultad de Ingeniería y Ciencias Aplicadas, Universidad de Las Américas, Quito 170125, Ecuador
[4] Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito USFQ, Quito 170901, Ecuador; randradep@usfq.edu.ec
* Correspondence: ivan.ortiz@udla.edu.ec
† These authors contributed equally to this work.

Abstract

Post-Traumatic Stress Disorder (PTSD) poses complex clinical challenges due to its emotional volatility, contextual sensitivity, and need for personalized care. Conventional AI systems often fall short in therapeutic contexts due to lack of explainability, ethical safeguards, and narrative understanding. We propose a hybrid neuro-symbolic architecture that combines Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), symbolic controllers, and ensemble classifiers to support clinicians in PTSD follow-up. The proposal integrates real-time anonymization, session memory through patient-specific RAG, and a Human-in-the-Loop (HITL) interface. It ensures clinical safety via symbolic logic rules derived from trauma-informed protocols. The proposed architecture enables safe, personalized AI-driven responses by combining statistical language modeling with explicit therapeutic constraints. Through modular integration, it supports affective signal adaptation, longitudinal memory, and ethical traceability. A comparative evaluation against state-of-the-art approaches highlights improvements in contextual alignment, privacy protection, and clinician supervision.

Keywords: Post-Traumatic Stress Disorder (PTSD); Large Language Models (LLMs); explainable clinical AI

check for updates

Academic Editors: Luc de Witte, Sana Ullah Jan, Haipeng Liu and Umer Saeed

Received: 18 June 2025
Revised: 25 July 2025
Accepted: 12 August 2025
Published: 15 August 2025

Citation: Cazares, M.; Miño-Ayala, J.; Ortiz, I.; Andrade, R. Designing Trustworthy AI Systems for PTSD Follow-Up. *Technologies* **2025**, *13*, 361. https://doi.org/10.3390/technologies13080361

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Developing Artificial Intelligence (AI) solutions for mental health, particularly systems designed to accompany patients through therapeutic journeys, demands architectural rigor that goes beyond performance optimization. A literature review identified three main application areas: (1) screening and detection of mental disorders, such as depression, suicide risk; (2) supporting clinical treatments and interventions, including the development of chatbots and tools to enhance clinical data; and (3) assisting mental health counseling and education [1]. Ethical concerns surrounding LLM use in mental health include informed consent, risks of manipulation, and the disruption of the therapist–patient relationship. Patients and clinicians may become overly reliant on AI for emotional support and decision-making, potentially undermining professional judgment and trust in traditional care. The World Health Organization (WHO) emphasizes that AI applications in mental health must consider "cultural acceptability, clinical soundness, and psychological safety", and go well

beyond performance metrics to ensure patient well-being [2]. Similarly, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) calls for AI health systems to be "consistent with medical ethics and respectful of patients' dignity, rights, and data privacy", with special attention to vulnerable populations such as those with mental health conditions [3]. The American Psychological Association (APA) further stresses that AI technologies in psychological services must be grounded in ethical design principles, including "respect for privacy, human oversight, mitigation of harm, and sensitivity to power differentials in therapeutic contexts" [4]. Conventional AI systems, including traditional machine learning classifiers and even standalone Large Language Models (LLMs), often fall short. models struggle to interpret unstructured patient narratives, while LLMs, despite their fluency, pose significant risks of generating clinically inappropriate or harmful content (hallucinations) and lack the longitudinal memory required for coherent therapeutic follow-up. While recent approaches like Retrieval-Augmented Generation (RAG) have improved the factual grounding of LLMs, they do not inherently provide the multi-layered safety and personalization needed for trauma care. A RAG system alone cannot guarantee adherence to trauma-informed protocols or adapt dynamically to a patient's emotional state, as it still operates on a probabilistic basis and lacks the deterministic control necessary for high-stakes clinical decisions [5,6].This gap creates a pressing need for a hybrid architecture that synergizes the narrative understanding of LLMs with the safety of symbolic logic, the precision of clinical classifiers, and the essential oversight of a human therapist. Such a system must be built on a foundation of privacy-by-design to earn the trust of both clinicians and patients' dignity, rights, and data privacy, with special attention to vulnerable populations such as those with mental health conditions, a principle strongly advocated by global health and ethics organizations [2–4]. The American Psychological Association (APA) further stresses that AI technologies in psychological services must be grounded in ethical design principles, including "respect for privacy, human oversight, mitigation of harm, and sensitivity to power differentials in therapeutic contexts" [4].

The primary novelty of this work lies in the design and conceptualization of a holistic, hybrid neuro-symbolic architecture specifically tailored for PTSD follow-up. Our contribution is the integration of five critical components into a single, cohesive framework designed for maximum clinical safety and efficacy:

(i) Real-time, privacy-preserving anonymization integrated at the ingestion layer [7].
(ii) Longitudinal session memory achieved through a patient-specific RAG system that ensures contextual continuity [8,9].
(iii) A symbolic control layer that enforces clinical safety rules derived from trauma-informed protocols, acting as a guardrail against harmful AI outputs [10,11].
(iv) Robust clinical assessment through an ensemble of classifiers that can process structured data in parallel to the narrative analysis [10,12].
(v) An explicit Human-in-the-Loop (HITL) interface that guarantees professional oversight, ensuring all AI-generated suggestions are validated by a therapist before reaching the patient [13,14].

This multi-layered approach directly addresses the limitations of existing systems by creating a trustworthy, explainable, and personalized AI co-pilot for clinicians.

1.1. LLMs in Mental Health Applications

Recent advances in Large Language Models (LLMs), such as GPT-4 and LLaMA, have shown promising applications in mental health care. LLMs can generate empathic, contextually rich responses that resemble therapist–patient interactions [15,16]. Studies have demonstrated that LLMs can support psychoeducation, simulate therapeutic dialogue, and assist in the preliminary screening of disorders such as depression and PTSD. However,

their limitations in personalization, memory, and clinical safety necessitate additional layers of control and grounding [17,18].

Large Language Models (LLMs), despite their remarkable capabilities in language understanding and generation, remain fundamentally statistical systems that lack grounded understanding or guaranteed factual correctness. A well-documented limitation is their potential to generate unsafe or clinically invalid responses when faced with sensitive patient disclosures, a risk highlighted by critiques of deep learning's limitations in achieving true understanding [5]. The inspectable representations of psychotherapeutic techniques in clinical LLMs help to avoid low explainability or interpretability and aid in the machine learning process for the generation of broad interventions that adapt to the patient's evolution [19,20]. AI systems used in mental health must undergo strict validation to ensure their outputs do not cause harm or undermine the therapeutic process. Therefore, it is critical to complement LLMs with symbolic controllers that enforce clinical rules and ethical boundaries, ensuring that generated content remains within the scope of safe and validated therapeutic practices.

1.2. Traditional ML vs. Emergent IA Methods

While traditional Machine Learning (ML) models have demonstrated utility in mental health classification tasks, they fall short in addressing the complexity, personalization, and narrative variability required for Post-Traumatic Stress Disorder (PTSD) intervention. PTSD is not merely a categorical diagnosis—it involves temporally dynamic symptoms, trauma-specific triggers, and nuanced emotional expression that structured features alone cannot adequately capture. Large Language Models (LLMs) are uniquely suited for this challenge due to their capacity to interpret unstructured data, such as patient narratives or therapeutic transcripts. Recent work has demonstrated that LLMs can extract semantically rich markers from trauma-related language to predict symptom trajectories [21], offering a level of contextual understanding beyond traditional classifiers.

Traditional ML models offer high interpretability and require low computational resources, making them suitable for structured data analysis such as symptom classification or physiological signal processing. However, they lack the ability to interpret unstructured narratives and cannot maintain contextual memory—both of which are essential in trauma-informed care. LLMs demonstrate superior performance in understanding patient language and generating empathetic responses, yet they are often considered "black-box" systems with limited transparency and a high risk of hallucination [1]. The LLM + RAG architecture addresses these concerns by retrieving clinically relevant documents—such as session notes or EMDR protocols—and using them as grounded input for generation [9]. This significantly improves personalization, contextual continuity, and ethical traceability, especially when combined with affective data and symbolic logic [22]. Grounded in theories from trauma psychotherapy [23], information retrieval [9], and affective computing [22], the hybrid model offers the most comprehensive framework for building adaptive and clinically aligned AI systems for PTSD support.

Concerns regarding the interpretability of LLMs are valid but increasingly mitigated through emerging methods in explainable AI (XAI). Techniques such as attention heatmaps, intermediate reasoning chains, and prompt-level rationales have enhanced transparency in generative systems [1]. Additionally, the integration of Retrieval-Augmented Generation (RAG) mechanisms improves both safety and traceability. RAG enables language models to ground their responses in structured, verifiable clinical sources—such as EMDR guidelines, patient session notes, and DSM-based assessments—allowing practitioners to audit the basis of each recommendation [9]. This traceability is a critical advancement over "black-box" deep learning models.

While the adoption of hybrid LLM + RAG architectures may entail marginal increases in computational infrastructure, the resulting clinical benefit and reduction in manual workload justify the investment. Traditional ML models require extensive feature engineering and manual retraining for each population or symptom variation, whereas RAG-enabled systems generalize across cases and dynamically retrieve relevant information. More importantly, hybrid systems enable multimodal coordination, wherein physiological signals (e.g., GSR, HRV) trigger adaptive language responses, a functionality that rule-based or static classifiers cannot emulate effectively [24].

From an ethical and clinical perspective, this hybrid architecture aligns with the principles of responsible AI in healthcare: it is adaptive, explainable, memory-augmented, and patient-centered. Given the emotional sensitivity and longitudinal nature of PTSD care, systems that understand personal history and respond to affective cues offer much greater therapeutic value than static classifiers. Therefore, far from representing technological overreach, this approach constitutes an evolution in trauma-informed AI, grounded in clinical theory and computational innovation.

1.3. Justifying a Multi-Component Architecture

While a standalone LLM + RAG architecture offers significant improvements over base LLMs, it remains insufficient for the stringent safety and personalization requirements of PTSD therapy. The probabilistic nature of LLMs, even when grounded by RAG, cannot guarantee adherence to the strict, deterministic rules demanded by clinical safety protocols [1,5]. Therefore, our proposed architecture deliberately separates concerns into distinct, specialized modules. We employ an ensemble of classifiers for robust analysis of structured clinical data and physiological signals, a task for which LLMs are ill-suited [12,24]. This is complemented by a symbolic control layer that acts as a deterministic guardrail, enforcing inviolable therapeutic rules and preventing the generation of harmful or retraumatizing content [10,11]. This neuro-symbolic approach, where the LLM's generative capabilities are constrained by both probabilistic classifiers and deterministic logic, is essential for building a system that is not only intelligent but also demonstrably safe, explainable, and aligned with ethical mandates [2,3]. Table 1 shows the compilation of the above, indicating the main justifications and roles of the different types of components. This approach constitutes an evolution in trauma-informed AI, grounded in clinical theory and computational innovation.

Table 1. Distinct roles of key architectural components.

Component	Primary Role and Justification
LLM + RAG	Processes unstructured narratives, retrieves patient-specific context from documents, and generates fluent, personalized therapeutic suggestions [9].
Ensemble Classifier	Analyzes structured data (e.g., clinical scores, biometric signals) to provide robust, interpretable risk assessments (e.g., detecting hyperarousal) [12,21].
Symbolic Controller	Enforces deterministic, hard-coded clinical safety rules (e.g., "IF dissociation THEN trigger grounding exercise"). Provides an auditable safety layer that LLMs cannot guarantee [10,25].
Human-in-the-Loop (HITL) Interface	Presents AI outputs as 'Agentic Suggestions' for therapist validation. Facilitates informed oversight rather than passive error-catching, ensuring clinical judgment remains central [13,14].

Table 2 presents a conceptual comparison summarizing these critical distinctions, highlighting the architectural gaps in traditional Machine Learning (ML), Large Language Models (LLMs), and LLM + Retrieval-Augmented Generation (LLM + RAG) architectures in the context of PTSD-aware AI systems.

Table 2. Comparison of AI architectures for PTSD support.

Capability	Traditional ML	LLM	LLM + RAG
Input Type	Structured	Text	Text + Documents
Personalization	Low	Moderate	High
Session Memory	None	None	Yes
Explainability	High	Low	Medium
Narrative Support	Weak	Strong	Strong + Context
Hallucination Risk	None	High	Reduced
Use of Clinical Protocols	Manual	Prompted	Retrieved
Emotion Signal Adaptation	Supported	No	Yes (via logic)
Infrastructure Cost	Low	Medium	High
Ethical Oversight	High	Low	Moderate

1.4. Motivation, Problem Statement, and Contributions

Conventional AI systems, especially standalone Large Language Models (LLMs), often fall short in complex therapeutic contexts like PTSD follow-up due to a critical lack of explainability, ethical safeguards, and narrative understanding. While LLMs can generate fluent text, they are prone to factual and contextual errors (hallucinations) and cannot guarantee adherence to clinical protocols, posing a significant risk to vulnerable patients. Similarly, traditional machine learning models, while interpretable, cannot process the rich, unstructured narratives central to trauma therapy and lack mechanisms for longitudinal memory, failing to track a patient's journey over time. There is a pressing need for an AI architecture designed from the ground up to be safe, transparent, and personalized, directly addressing the ethical and clinical guidelines set forth by organizations like the WHO, UNESCO, and APA [2–4].

This paper introduces a novel hybrid neuro-symbolic architecture that directly addresses these limitations. The novelty and main contributions of our work are threefold:

- Integration of Safety Layers: We propose a unique combination of a symbolic controller, which enforces strict clinical rules, with Retrieval-Augmented Generation (RAG) to ground LLM responses in verified clinical documents and patient-specific history. This significantly reduces hallucination risk and ensures therapeutic alignment, a key drawback in existing models.
- Privacy-by-Design with Longitudinal Memory: Our architecture is the first to holistically integrate real-time data anonymization at the point of ingestion with a patient-specific RAG system. This allows for the creation of a secure, longitudinal session memory, enabling true personalization without compromising patient confidentiality.
- Clinician-Centric Workflow: We explicitly embed a Human-in-the-Loop (HITL) interface where AI outputs are presented as 'Agentic Suggestions' to the therapist, not the patient. This maintains clinical oversight and aligns with ethical mandates for AI to augment, not replace, professional judgment, a critical failure point in more autonomous systems.

2. Theorical Background

To address the multifaceted challenges of PTSD support outlined previously, our proposed architecture is built upon several key theoretical pillars from computer science and clinical practice. This section reviews the foundational concepts that inform our design choices. We first explore hybrid neuro-symbolic reasoning, which provides the core paradigm for blending the flexibility of neural networks with the safety of rule-based systems. We then discuss the critical role of Human-in-the-Loop (HITL) design in ensuring ethical oversight. Subsequently, we detail the principles of privacy-by-design and personalization through longitudinal memory, which are non-negotiable requirements for building trust and efficacy in mental health technologies. Together, these principles form the theoretical bedrock for a responsible and effective AI system for PTSD follow-up.

2.1. Hybrid Neuro-Symbolic Reasoning

Traditional machine learning models often fail to meet the clinical standards of explainability and traceability required in healthcare. Neuro-symbolic architectures, which integrate deep learning (e.g., LLMs) with symbolic reasoning (e.g., rule engines), offer a solution that balances flexibility and control. These systems allow the encoding of domain-specific knowledge, such as clinical protocols or trauma-sensitive rules, into formal logic constraints [10]. In mental health, where interventions must follow structured yet adaptive guidelines, this hybrid approach ensures that AI outputs remain grounded in therapeutic standards. UNESCO recommends such designs as part of its ethical AI guidelines, advocating for transparent systems with human-verifiable logic layers [3]. Combining neural models with symbolic logic (e.g., decision trees or rule engines) enables systems to enforce safety constraints and align interventions with evidence-based practices. Neuro-symbolic architectures [10,11] have been proposed for sensitive domains where consistency, traceability, and logic-based filtering are essential. In PTSD therapy, symbolic logic can implement conditional flows (e.g., IF dissociation $\rightarrow$ THEN offer grounding) that align with cognitive-behavioral guidelines and reduce ethical risk.

Neuro-symbolic AI offers a powerful paradigm for building safe, explainable, and adaptive systems in sensitive domains such as psychotherapy. By combining neural models—such as large language models (LLMs)—with symbolic reasoning layers, these architectures leverage both flexible understanding of emotional language and strict adherence to clinical rules [10]. The neural component provides rich semantic understanding and generative capacity, enabling the system to process patient narratives, emotions, and evolving contexts. Meanwhile, the symbolic layer encodes clinical protocols, ethical constraints, and patient safety rules as formal logic [3]. This hybrid integration ensures that AI outputs remain aligned with therapeutic standards and prevents the generation of unsafe or inappropriate content.

In the context of a Psico-Tutor for PTSD and trauma-informed care, the neuro-symbolic approach is particularly critical. Purely neural models, while powerful, can produce hallucinations or violate clinical boundaries, posing risks to vulnerable patients. The symbolic controller acts as a safeguard, gating or filtering neural outputs based on the patient's current state (e.g., dissociation, crisis, readiness for exposure therapy) and the session context [2]. Furthermore, explainability—a core requirement in clinical AI—is inherently supported by symbolic reasoning, enabling therapists to inspect and validate the AI's suggestions [13]. This supports a Human-in-the-Loop (HITL) workflow, in line with APA guidelines, where clinicians maintain ultimate control over therapeutic decisions [4].

The symbolic controller in hybrid AI architectures serves as a formal reasoning layer that enforces safety, interpretability, and rule-based governance. Unlike sub-symbolic models such as neural networks, symbolic systems can encode clinically validated rules in

a transparent and auditable format [10,11]. In PTSD-aware systems, symbolic controllers are crucial for mapping sensitive conditions (e.g., dissociation, hyperarousal) to appropriate therapeutic actions through predefined intervention rules. This form of conditional logic reflects the structure of evidence-based clinical protocols, such as those found in Cognitive Behavioral Therapy (CBT) and Eye Movement Desensitization and Reprocessing (EMDR). The inclusion of symbolic reasoning layers thus enhances the explainability of AI outputs while safeguarding against unsafe generative behavior [25].

Symbolic controllers are increasingly integrated into neuro-symbolic systems, where they operate in conjunction with deep learning models and ensemble classifiers. While Large Language Models (LLMs) provide high-level narrative interpretation, they lack deterministic control mechanisms, making them prone to hallucination and inconsistent behavior in sensitive domains [5]. Symbolic controllers mitigate these risks by applying logic constraints to limit action spaces, especially during real-time therapeutic decision-making. Recent studies in mental health AI [26,27] have shown that rule-based logic engines—such as decision trees, Prolog-based systems, or Drools—can dynamically route inputs based on symptom thresholds, biometric feedback (e.g., HRV, GSR), or past session annotations. This enables safe and adaptive personalization without over-reliance on generative agents.

From a clinical ethics standpoint, the symbolic controller aligns with the principles of responsible AI in healthcare. It supports traceability, auditability, and shared decision-making by allowing clinicians to inspect and modify the inference process [13,14]. In PTSD-specific systems, symbolic logic can encode trauma-informed safeguards—such as preventing emotionally intense content during dissociative states or escalating to human oversight when physiological markers indicate acute distress. These mechanisms resonate with the AI4People ethical framework and IEEE P7001 standards [28] on transparency in autonomous systems. Thus, the symbolic controller not only improves system robustness but also serves as a regulatory scaffold for therapeutic safety and trustworthiness.

Finally, neuro-symbolic architectures also enhance personalization and longitudinal adaptation in psychotherapy. Retrieval-Augmented Generation (RAG) combined with patient-specific embeddings allows the system to contextualize responses based on prior sessions, clinical progress, and therapeutic goals [8]. Symbolic reasoning further enables dynamic adaptation of intervention strategies while maintaining consistency with formal treatment protocols. As recommended by the APA, WHO, and UNESCO, AI systems in mental health must be transparent, ethically aligned, and augment—not replace—clinician judgment [2–4]. Neuro-symbolic AI provides the ideal foundation to meet these requirements in the design of responsible Psico-Tutor systems.

2.2. Human-in-the-Loop (HITL) Design

The APA's guidelines for telepsychology and technology use in psychological services stress that AI tools must enhance, not replace, human judgment [4]. Particularly in trauma care, inappropriate or emotionally harmful outputs from generative systems can endanger patient safety. A human-in-the-loop model routes AI-generated content through a qualified professional, typically a therapist, who validates and contextualizes the information before sharing it with the patient. This ensures that the therapeutic process remains under professional control, as emphasized by the WHO's call for clinician-centered AI [2]. HITL design also strengthens accountability and interpretability, two fundamental ethical requirements in mental health technologies [13]. The integration of Human-in-the-Loop (HITL) design is a foundational requirement in clinical AI. However, we explicitly reject the flawed premise of HITL as a simple "safety net" [6]. Recent research highlights that passive human oversight is often insufficient for catching subtle AI failures, especially in

resource-strained environments [29,30]. Our architecture therefore implements HITL not as a passive validator but as an active, collaborative workflow centered on the principle of 'informed oversight' [31]. The system presents AI outputs as 'Agentic Suggestions' to the therapist, who is empowered by explainability tools to audit, modify, or approve them. This clinician-centric model ensures that AI enhances, not replaces, professional judgment, aligning with ethical mandates from the APA and WHO while acknowledging the practical limitations of simplistic HITL implementations [2,4].

Recent frameworks such as EmoAgent demonstrate the practical value of HITL mechanisms by simulating emotionally vulnerable users and incorporating feedback agents that correct or suppress inappropriate AI responses in real time [32]. These systems combine affective modeling, symbolic reasoning, and therapist interfaces to support ethical, explainable, and adaptive interventions. The Human-in-the-Loop (HITL) paradigm has been increasingly scrutinized in the context of AI deployment in healthcare. While it is often assumed that human oversight can act as a safeguard against errors in generative AI systems, recent work has exposed both technical and sociological limitations. Kabata and Thaldar [29] question the practical feasibility of implementing HITL in low-resource environments, arguing instead for a regulatory framework grounded in the human right to science. Clark [6] critiques the "HITL as safety net" narrative, showing that human reviewers may lack the time, tools, or expertise to identify subtle or emergent AI failures. From an ethical perspective, Herington [31] advocates for informed oversight, emphasizing that human actors must understand both the capabilities and limitations of AI systems. Similarly, Bakken [30] notes that adverse events in medical AI deployments are frequently due to inappropriate human use rather than algorithmic errors, highlighting the critical role of human training and workflows. Other authors, such as Griffen and Owens, propose participatory governance models as an alternative to linear HITL pipelines, suggesting that multi-actor oversight is more resilient and equitable [33]. These contributions collectively suggest that while HITL remains important, it must be reimagined as a distributed, dynamic, and ethically grounded design principle. As summarized in Table 3, recent literature highlights a growing consensus that Human-in-the-Loop (HITL) mechanisms must evolve beyond simplistic oversight roles, emphasizing contextual understanding, participatory governance, and systemic integration to ensure safety and accountability in healthcare AI.

Table 3. Summary of key studies on HITL in healthcare AI (2022–2025).

Scope	Reference
Critically examines the limitations of HITL in low-resource healthcare systems; proposes a rights-based regulatory alternative.	Kabata and Thaldar (2024) [29]
Argues that human oversight is often overestimated as a safeguard in generative AI, especially when reviewers are ill-equipped.	Clark (2024) [6]
Emphasizes the ethical necessity of informed human supervision and understanding of AI limitations in clinical decision-making.	Herington (2023) [31]
Finds that many AI-related failures in medicine are due to human misuse rather than algorithmic faults, underscoring HITL complexity.	Bakken (2023) [30]
Proposes participatory AI governance models as an alternative to linear HITL designs; encourages inclusive oversight across stakeholders.	Griffen and Owens (2024) [33]

Ultimately, HITL design is not only a safeguard but a bridge between human empathy and machine intelligence, enabling AI to function as a responsible co-pilot in mental health support systems [3,13].

2.3. Privacy-by-Design and Real-Time Anonymization

The sensitivity of psychological and emotional health data requires strict adherence to privacy and data protection norms. The APA Code of Ethics [4], along with GDPR and the WHO's guidelines [2], highlight confidentiality as a non-negotiable principle in mental health interventions. Real-time anonymization—using tools like Named Entity Recognition (NER) and entity masking—allows for live data de-identification without loss of semantic integrity. As shown in previous implementations [7], privacy-first systems can operate in therapeutic settings while complying with ethical and legal frameworks. Incorporating anonymization at the architectural level ensures these protections are systemic and not optional.

Real-time anonymization mechanisms are particularly critical in AI-mediated mental health applications where sensitive disclosures—such as trauma narratives, abuse histories, or suicidal ideation—are frequently expressed in unstructured natural language. Traditional de-identification techniques that rely on post hoc redaction are inadequate in dynamic therapeutic settings, where AI models operate on live inputs. Modern privacy-preserving architectures integrate Named Entity Recognition (NER) tools trained on health-specific corpora to detect and mask Personally Identifiable Information (PII) before data is stored or processed by downstream models [34]. Techniques such as context-aware masking and pseudo-anonymization further enhance semantic retention while minimizing re-identification risk. Notably, these systems must be continuously audited and stress-tested to mitigate adversarial inference attacks, particularly in deployments involving Large Language Models (LLMs) with long-term memory components [35].

Table 4 summarizes key sources that collectively frame the theoretical and practical landscape of privacy-by-design and anonymization strategies in AI applications for mental health. Each reference contributes to a layered understanding of privacy threats, mitigations, and policy alignment.

Table 4. Summary of data anonymization techniques for mental health AI.

Technique	Advantages	Limitations	Applicable Modalities
Masking	Simple to implement; removes sensitive tokens directly.	May destroy contextual information; vulnerable to inference from structure.	Text, Audio (transcripts)
Hashing	Efficient for identifiers; irreversible if salted properly.	Not effective for natural language or biometric data.	Structured Text
Encryption	Strong confidentiality during storage and transit.	Data must be decrypted for processing; not a form of anonymization per se.	All (infrastructure-level)
Pseudo-anonymization	Maintains linkability for authorized research; improves usability.	Still reversible; may not meet GDPR's standard for anonymization.	Text, Audio
Suppression	Prevents exposure of specific fields or time spans.	Reduces data utility and completeness.	Text, Audio, Video
Generalization	Reduces granularity of sensitive data (e.g., age ranges).	May reduce diagnostic precision in clinical tasks.	Text, Video (e.g., demographic traits)
Perturbation	Provides statistical obfuscation; supports differential privacy.	May distort signals crucial for affective analysis.	Audio, Video
Synthetic Data Generation	Breaks direct link to real individuals; enables data sharing.	Risk of memorization and bias; requires fidelity validation.	Text, Audio, Video

Incorporating privacy-by-design into the system architecture not only satisfies legal compliance frameworks like the General Data Protection Regulation (GDPR) but also aligns with ethical imperatives outlined by UNESCO and the Council of Europe on AI in healthcare [3]. Beyond regulatory concerns, empirical evidence suggests that robust anonymization mechanisms increase patient trust and willingness to engage with digital mental health tools [36]. This is particularly relevant for marginalized or trauma-exposed populations, who may be more sensitive to perceived surveillance or data misuse. Consequently, privacy must be treated not as a modular add-on but as a foundational layer that interacts with all components of the AI system—from ingestion to inference and storage—ensuring that therapeutic value is not achieved at the cost of user dignity and safety.

2.4. Personalization and Longitudinal Memory

Mental health interventions, especially for PTSD, require systems that can adapt over time to evolving symptoms and personal histories. Recent studies underscore the value of longitudinal personalization in AI-driven therapy tools [8]. Retrieval-Augmented Generation (RAG) mechanisms, paired with patient-specific vector profiles, enable the retrieval of past sessions, clinical progress, and contextual cues. This allows systems to generate content that reflects the patient's therapeutic journey, strengthening engagement and trust. According to APA's Principle D (Justice), psychological services must be individualized and inclusive—personalization through memory-aware AI directly supports this principle [37]. The RAG architecture [9] enhances generative models by retrieving relevant documents from a vector store and using them as context for generation. In clinical contexts, RAG can provide grounded responses using patient-specific notes, therapy protocols (e.g., EMDR or CBT), and session history. This mitigates hallucinations and enables continuity across therapeutic sessions. A recent study by Lin [38] in mental health chatbots found that RAG significantly increased perceived relevance and safety in AI-driven interventions.

2.5. Requeriments for AI System Support PTSD

An AI system designed to support PTSD follow-up must address a comprehensive set of functional and ethical requirements to ensure clinical safety and therapeutic efficacy. First, it must integrate a deep understanding of emotional narratives, enabling the processing of patients' subjective expressions and contextual cues. Second, it should implement a longitudinal memory component that tracks symptom evolution and therapeutic progress across sessions. Third, personalization must be dynamic and responsive, adapting interventions based on the individual's clinical profile and therapy history. Critically, the system should embed a symbolic control layer that enforces adherence to clinical protocols and prevents the generation of harmful or retraumatizing content. Real-time privacy protection is essential, requiring the system to anonymize sensitive data in compliance with international data protection standards. Furthermore, explainability and transparency must be inherent to the architecture, allowing therapists to audit AI-generated outputs and maintain ultimate decision-making authority through Human-in-the-Loop (HITL) supervision (Figure 1). Collectively, these elements form the foundation for ethical, safe, and effective AI solutions in the sensitive domain of PTSD care.

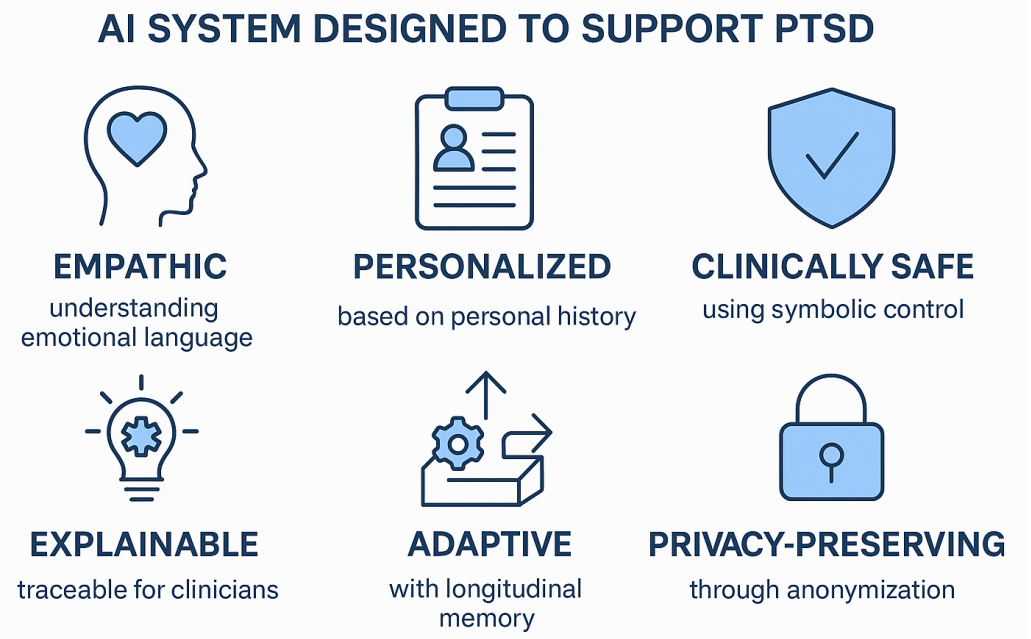

Figure 1. Key components and architecture of an AI system designed to support PTSD follow-up. The system integrates LLM + RAG layers for narrative understanding, a symbolic controller for clinical safety, longitudinal memory for personalization, a Human-in-the-Loop (HITL) workflow for therapist supervision, and real-time anonymization for privacy protection.

3. Hybrid Architecture for Personalized AI-Therapy

The proposed architecture (Figure 2) integrates a layered, modular system designed to support AI-assisted decision-making in trauma-informed therapeutic contexts. It begins with a real-time anonymization layer that protects sensitive user data through Named Entity Recognition (NER) techniques powered by NLP libraries such as spaCy, Presidio, or LLM-based detectors. Once sanitized, inputs are passed to an ensemble classifier module, which leverages multiple machine learning models to generate robust, structured assessments of symptom patterns or emotional indicators. In parallel, a Retrieval-Augmented Generation (RAG) component accesses a sanitized knowledge base—including therapeutic guidelines, past session summaries, and clinical protocols—to contextualize the case. These results are synthesized into an agentic suggestion layer that combines both statistical inference and document-grounded reasoning to provide actionable, transparent recommendations to therapists. This architecture balances interpretability and adaptability, aligning with best practices in hybrid neuro-symbolic AI, and is consistent with emerging principles of human-centered, explainable artificial intelligence in clinical domains. It ensures that decision support remains privacy-aware, evidence-informed, and therapeutically aligned.

Ensemble classifiers in a hybrid architecture that also incorporates Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) is theoretically grounded and functionally complementary. While LLMs and RAG modules excel at understanding unstructured patient narratives and generating context-sensitive suggestions based on historical documents, they are not inherently designed to perform precise classification tasks such as risk stratification, symptom severity scoring, or detecting patterns from multimodal physiological data. Ensemble classifiers—such as Random Forests, XGBoost, and soft-voting schemes—are widely recognized for their robustness, interpretability, and generalization capabilities in clinical machine learning applications [12]. In PTSD-related systems, these models can reliably analyze structured inputs like PCL-5 scores, HRV trends, or behavioral metrics to identify clinical deterioration or emotional dysregulation. Furthermore, hybrid cognitive architectures combining sub-symbolic (deep learning) and

symbolic (rule-based or decision-tree-based) reasoning have been proposed as more reliable and explainable alternatives to standalone LLMs [10]. This approach aligns with the design of neuro-symbolic systems that integrate statistical inference with logical constraints to enhance safety and transparency in decision-making. By coupling generative agents (LLMs + RAG) with diagnostic classifiers, the system benefits from both narrative comprehension and predictive precision—thus increasing its therapeutic relevance, reliability, and alignment with trauma-informed AI principles [22].

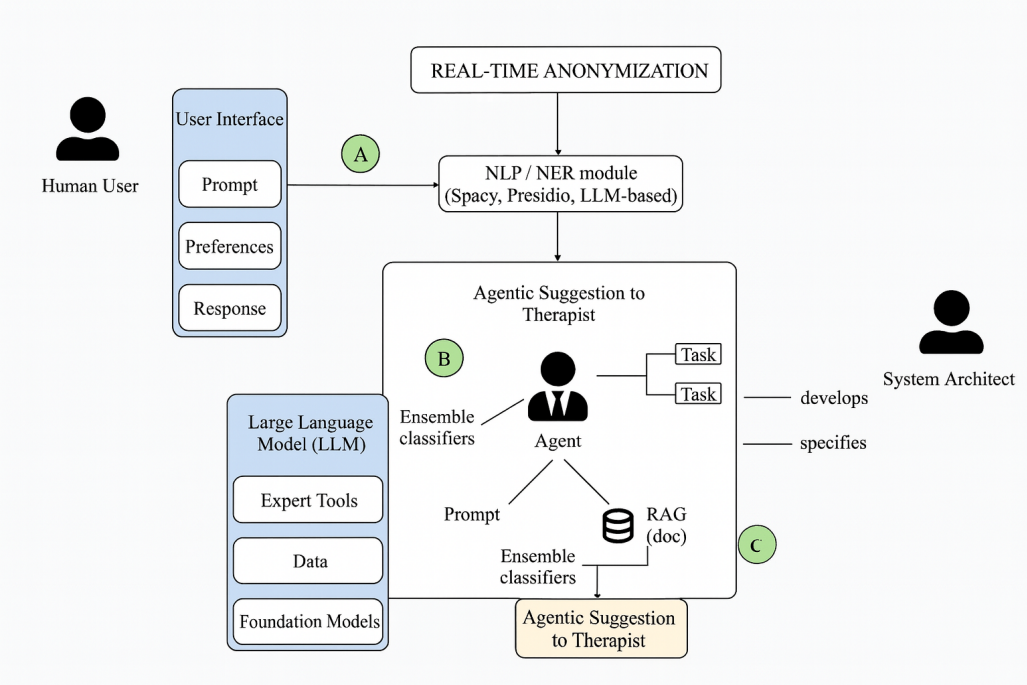

Figure 2. Hybrid architecture for personalized AI-therapy.

The Figure 2 describe the hybrid architecture:

(A) User interface → NLP/NER for *real-time anonymization*.
(B) Alignment agent: filters and orchestrates LLM outputs with ensemble classifiers, task blocks, and RAG, then issues suggestions to the therapist.
(C) Architect control surface: specifies policies, tasks, and RAG corpus/indexes to ensure safety, compliance, and traceability.

The modules of the proposal architecture Figure 2, complement by Table 5, are as follows:

1. Real-time Anonymization: This is critical in systems processing sensitive natural language data, such as emotional narratives from patients, clinical interviews, or therapy sessions. The use of an NLP/NER module (SpaCy, Presidio, LLM-based NER) aligns with recent findings showing that LLM-based NER models outperform traditional approaches in detecting complex entities and idiosyncratic language patterns.
   The need for in-stream anonymization reflects both legal compliance requirements such as the GDPR and the ethical frameworks proposed for AI in healthcare.
2. Alignment Agent: The concept of alignment refers to shaping AI responses to adhere to human values and therapeutic objectives, ensemble classifiers to filter and classify AI-generated content, task-specific blocks that constrain the agent's operational scope, and RAG (Retrieval-Augmented Generation) techniques, which have been shown to improve the factuality and controllability of LLM outputs in clinical AI systems.
   The architectural choice to present agentic suggestions to the therapist, rather than direct responses to patients, reflects the Human-in-the-Loop (HitL) principle—now considered essential for AI in healthcare applications.

3. LLM + Expert Tools + Data + Foundation Models: This reflects the modern hybrid AI approach, combining general-purpose models with domain-specific components. Furthermore, RAG architectures are becoming a best practice in domains where semantic precision and deep contextualization are paramount.
4. User Interface: Supporting explainability and personalization. The user interface block handles prompts, preferences, and responses and reflects the need for personalized therapist interfaces.
5. AI behavior auditing: Designing interpretable and auditable AI systems for mental health applications.

Table 5. Layered view of the AI system architecture for PTSD support.

Layer	Components
Data Ingestion Layer	Real-time Anonymization + User Interface
Knowledge & Retrieval Layer	Retrieval-Augmented Generation (RAG) + Patient Profile Store
Language Understanding Layer	Large Language Model (LLM) + Prompt Engineering
Clinical Safety Layer	Symbolic Rule Engine + Ensemble Classifiers
Human-in-the-Loop Layer	Therapist Interface + Explainability Dashboard

4. Discussion

This proposed architecture advances the intersection of artificial intelligence and mental health by introducing a privacy-preserving, explainable, and modular system specifically tailored for trauma-informed psychological support. Unlike black-box LLM-only systems, this hybrid model integrates real-time anonymization, ensemble-based clinical classification, and retrieval-augmented language generation to ensure contextual accuracy and therapeutic alignment. The architecture bridges the gap between structured clinical assessment and unstructured narrative expression—two domains that are often treated separately in psychological AI applications. Moreover, it reflects recent calls in psychological science and human–computer interaction to develop systems that are both empathetic and computationally grounded. By aligning technical components (e.g., ensemble classifiers and knowledge retrieval) with clinical workflows (e.g., therapist recommendation and goal monitoring), this work contributes a replicable design that improves trust, interpretability, and safety in AI-assisted psychotherapy. In Table 6 we see a comparison with the IA architecture for mental health and see what alternatives exist.

The modular design of this architecture ensures that it can be adapted or extended by other researchers and clinical professionals. Each component, such as the anonymization module, the ensemble classifier, and the RAG mechanism, can be independently developed, evaluated, and replaced using open-source tools (e.g., spaCy, Presidio, Hugging Face transformers, or Scikit-learn). The architecture's reliance on documented knowledge bases and explicit inference pathways also supports transparency and auditability, which are essential for reproducibility in psychological research. For practicing psychologists, the system provides a blueprint for incorporating AI co-pilots in therapeutic environments without compromising ethical standards, client autonomy, or clinical judgment. By facilitating hybrid reasoning that is both data-driven and context-aware, this architecture (Table 7) offers a generalizable framework for the future development of AI tools in mental health care, particularly in applications requiring personalization, longitudinal tracking, and narrative understanding.

Table 6. Comparison of related AI architectures for mental health and PTSD-aware systems.

Work/ Approach	Architecture Type	Personalization	Memory Awareness	Safety Layer	Privacy Layer	Human-in-the-Loop
[38] (Mental Health Chatbots)	LLM + RAG	Moderate	Limited	No	No	Partial
[39] (Affective AI for PTSD)	ML + Physiological Signals	Low	No	Symbolic	No	No
[11] (Neuro-symbolic AI)	Neuro-symbolic Framework	Conceptual	No	Strong	No	Conceptual
[13] (HITL in Healthcare)	HITL Framework	Limited	Limited	Strong	No	Yes
This Work (Proposed Architecture	Hybrid Neuro-symbolic (LLM + RAG + Classifier + Symbolic)	High	Yes	Strong	Yes	Explicit

Table 7. Evaluation of the proposed architecture against key requirements for AI systems supporting PTSD follow-up.

Requirement for AI in PTSD Support	How Is It Covered in the Architecture?	Observations
Emotional narrative understanding	Through the LLM layer (Prompt → LLM → Response), with Foundation Models and expert tools.	The LLM processes patient narratives and can capture emotional language.
Longitudinal memory	Partial → RAG (doc) provides Retrieval-Augmented Generation; patient profile/history can be included here.	Could be reinforced with explicit mention of "patient vector profiles" or "patient memory embeddings".
Personalization	Partial → Personalization is implied via RAG (doc) and User Preferences → Prompting.	Could be made more explicit by showing that User Preferences + past session data modulate the Agent's behavior.
Safety and symbolic control	Strong → Symbolic Controller and Alignment Agent with Ensemble Classifiers act as guardrails for the LLM output.	Very aligned with neuro-symbolic AI best practices. Clear symbolic layer is present.
Explainability and HITL	Strong → Agentic Suggestion is not sent directly to the patient, but to the therapist, enabling HITL supervision.	Matches APA and WHO recommendations for clinical HITL AI.
Privacy protection	Strong → Real-Time Anonymization (NLP/NER module with SpaCy, Presidio, LLM-based NER).	Excellent implementation of GDPR/WHO recommendations on privacy-by-design.
Modularity	Clearly modular → separate modules for LLM, NER, Symbolic Controller, Ensemble Classifiers, HITL.	Follows open, modular architecture principles (UNESCO 2021).
Reproducibility	Implied → Modular architecture supports reproducibility.	Can be reinforced by documenting component configurations in the paper.

5. Limitations and Future Work

It is important to acknowledge the conceptual nature of this work. The proposed architecture represents a novel theoretical framework for designing trustworthy AI systems for PTSD follow-up, but it has not yet undergone empirical validation. We recognize this as a significant limitation and outline the following critical steps for future research:

5.1. Empirical Validation and Performance Metrics

The immediate next step is to develop a functional prototype of the architecture. The validation will proceed in two phases:

1. Simulated Data Evaluation: Initially, the system's safety and performance will be tested using simulated patient–therapist dialogues. Frameworks like EmoAgent [32] can be used to create controlled scenarios that test the symbolic controller's ability to handle high-risk situations (e.g., dissociation, hyperarousal) and the RAG system's capacity for contextual recall. Performance will be measured using metrics for clinical alignment, response relevance, and safety flag activation.
2. Clinical Pilot Study: Following successful simulation, a pilot study involving clinicians and, subsequently, patients will be designed. This phase is contingent upon receiving full approval from an Institutional Review Board (IRB) or a relevant ethics committee. The study will evaluate the system's usability, the quality of the 'Agentic Suggestions', and the impact on the therapeutic workflow. Data will be collected through clinician feedback, standardized clinical scales, and qualitative interviews.

5.2. Computational Cost and Deployment Considerations

While our discussion acknowledged the increased computational cost associated with a hybrid architecture, a more detailed analysis is warranted. The primary computational bottlenecks are expected to be as follows:

- LLM Inference: The real-time operation of a large language model requires significant GPU resources. The choice of model size (e.g., 7B vs. 70B parameters) will directly impact latency and cost.
- RAG System: The vector database retrieval speed depends on the size of the patient-specific knowledge base and the efficiency of the indexing algorithm (e.g., FAISS).

Future work will involve benchmarking different model sizes and hardware configurations to identify a cost-effective deployment strategy. Mitigation techniques such as model quantization, knowledge distillation, and the use of smaller, fine-tuned models will be explored to reduce the resource footprint without compromising performance.

6. Implementation Roadmap and Technologies

To enhance reproducibility and provide a clear path for future development, we outline a potential technology stack for implementing the proposed architecture (Table 8). This blueprint is intended as a guide, and specific components can be substituted with alternative open-source or commercial tools.

Table 8. Proposed technology stack for implementation.

Component	Proposed Technology/Framework	Key Considerations
Real-time Anonymization	SpaCy or Presidio (NLP/NER)	Requires fine-tuning on clinical text for optimal entity detection.
Ensemble Classifier	Scikit-learn (Random Forest, XGBoost)	Model selection based on interpretability and performance with structured clinical data.
LLM + RAG	Hugging Face Transformers, LangChain/LlamaIndex, FAISS (vector store)	Choice of foundation model and retrieval strategy are critical for performance.
Symbolic Controller	Drools (Rule Engine) or custom Python logic	Rules must be defined in close collaboration with clinical experts.
HITL Interface	Streamlit or Gradio	Should be designed for rapid prototyping and iterative feedback from therapists.

Author Contributions: Conceptualization, R.A. and M.C.; methodology, I.O.; software, I.O.; validation, R.A., M.C., and J.M.-A.; formal analysis, M.C.; investigation, R.A.; resources, R.A.; data curation, M.C.; writing—original draft preparation, M.C.; writing—review and editing, J.M.-A. and R.A.; visualization, J.M.-A.; supervision, R.A.; project administration, I.O.; funding acquisition, R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the authors used GPT-4o for the purposes of style and writing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ji, X.; Lee, W.; Tan, R.; Tan, M. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *56*, 1–38. [CrossRef]
2. World Health Organization (WHO). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*; Technical Report; World Health Organization: Geneva, Switzerland, 2021.
3. United Nations Educational, Scientific and Cultural Organization (UNESCO). *Recommendation on the Ethics of Artificial Intelligence*; Technical Report; UNESCO: Paris, France, 2021.
4. American Psychological Association (APA). *Guidelines for the Optimal Use of Artificial Intelligence in Psychological Practice and Research*; Technical Report; American Psychological Association: Washington, DC, USA, 2023.
5. Marcus, G. The Next Decade in AI: Why Deep Learning Alone Is Not Enough. *AI Mag.* **2022**, *43*, 54–62.
6. Clark, J. The Fallacy of the Human-in-the-Loop as a Safety Net for Generative AI Applications in Healthcare. *Medium*, 25 June 2024.
7. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2019**, *28*, 689–707. [CrossRef]
8. Thakkar, A.; Gupta, A.; De Sousa, A. Artificial intelligence in positive mental health: A narrative review. *Front. Digit. Health* **2024**, *6*, 1280235. [CrossRef]
9. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 9459–9474.

10. d'Avila Garcez, A.; Lamb, L.; Gabbay, D. Neuro-symbolic AI: The state of the art and future directions. *Front. Artif. Intell.* **2019**, *2*, 1. [CrossRef]
11. Garcez, A.d.; Lamb, L.C. Neurosymbolic AI: The 3rd Wave. *Artif. Intell. Rev.* **2023**, *56*, 12513–12543. [CrossRef]
12. Bennett, P.N.; Caruana, R.; Guestrin, C. Ensemble Methods in Machine Learning. In Proceedings of the International Workshop on Multiple Classifier Systems, Cagliari, Italy, 21–23 June 2000; pp. 1–15.
13. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
14. Shortliffe, E.; Sepúlveda, M. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* **2018**, *320*, 2199–2200. [CrossRef] [PubMed]
15. Hua, Y.; Na, H.; Li, Z.; Liu, F.; Fang, X.; Clifton, D.; Torous, J. A scoping review of large language models for generative tasks in mental health care. *npj Digit. Med.* **2025**, *8*, 230. [CrossRef] [PubMed]
16. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*. Available online: https://openreview.net/forum?id=RAW0XaE5nq (accessed on 11 August 2025).
17. Raji, I.D.; Gabriel, I.; Clark, J.; Fjeld, J.; Zaldivar, A. Outsider Oversight: The Case for Independent Audits of AI Systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, Canada, 3–10 March 2021; pp. 563–575.
18. Lawrence, H.R.; Schneider, R.A.; Rubin, S.B.; Mataric, M.J.; McDuff, D.J.; Jones Bell, M. The opportunities and risks of large language models in mental health. *JMIR Mental Health* **2024**, *11*, e59479. [CrossRef]
19. Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; Klang, E. Large Language Models and Empathy: Systematic Review. *J. Med. Internet Res.* **2024**, *26*, e52597. [CrossRef]
20. Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; Zhuang, Y. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. In Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2023, Ernest N. Morial Convention Center, New Orleans, LA, USA, 10–16 December 2023; Volume 36.
21. Nouri, M.; Makwana, M.; Brown, K.N.; Galatzer-Levy, I.R.; Kaplan, J.B.; Hoffman, Y.S.G.; Rosen, R.L.; Wright, A.G.C.; Feder, A.; Southwick, S.M.; et al. World Trade Center Responders in Their Own Words: Predicting PTSD Symptom Trajectories with AI-Based Language Analyses. *Psychol. Med.* **2023**, *53*, 53–65.
22. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
23. Shapiro, F. *Eye Movement Desensitization and Reprocessing (EMDR) Therapy: Basic Principles, Protocols, and Procedures*; Guilford Press: New York, NY, USA, 2017.
24. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
25. Zhou, M.; Wang, Z.; Ghosh, S.; Lin, J.; Vukadinović Greetham, D. Safe and Interpretable AI for Healthcare: A Neuro-Symbolic Perspective. *npj Digit. Med.* **2023**, *6*, 1–13.
26. Kassab, M.; DeFranco, J.; Ali, N. Hybrid Neuro-Symbolic Architectures for Trustworthy AI: A Clinical Perspective. *ACM Comput. Surv.* **2023**, *56*, 1–32.
27. Choudhury, M.; Zhang, Y.; Jain, A. Ethical Control Layers for Generative Health Agents: A Case for Symbolic Safety. *J. Biomed. Inform.* **2023**, *139*, 104373.
28. *IEEE Std 7001-2021*; IEEE Standard for Transparency of Autonomous Systems. IEEE Standards Association: Piscataway, NJ, USA, 2021; pp. 1–33. Available online: https://standards.ieee.org/ieee/7001/6929/ (accessed on 11 August 2025).
29. Kabata, L.; Thaldar, D. AI in health: Keeping the human in the loop. *S. Afr. J. Bioeth. Law* **2024**, *17*, 52–58.
30. Chustecki, M. Benefits and Risks of AI in Health Care: Narrative Review. *Interact. J. Med. Res.* **2024**, *13*, e53616. [CrossRef]
31. Herington, J. Keeping a Human in the Loop: Managing the Ethics of AI in Medicine. *University of Rochester Medical Center News*, 19 October 2023. Available online: https://www.urmc.rochester.edu/news/story/keeping-a-human-in-the-loop-managing-the-ethics-of-ai-in-medicine (accessed on 11 August 2025).
32. Luxton, D.D. Artificial Intelligence in Behavioral and Mental Health Care. In *AI in Behavioral and Mental Health Care*; Academic Press: Cambridge, MA, USA, 2016; pp. 107–120.
33. Griffen, A.; Owens, D. Keeping Humans in the Loop: Participatory Governance in AI Health Systems. *Am. J. Bioeth.* **2024**, *24*, 22–30. [CrossRef]
34. Meng, Y.; Huang, J.; Trivedi, P.; Yuan, X.; Liu, F.; Chen, S.; Zhang, Q.; Wang, L.; Li, H.; Lin, Y.; et al. DeID-GPT: Zero-shot Medical Text De-identification by GPT-based Generative Agents. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, ON, Canada, 9–14 July 2023; pp. 1–12.
35. Mireshghallah, F.; Tazi, N.; Trask, A.; Shokri, R. Privacy in Natural Language Processing: A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–39.

36. Abdalla, M.; Mir, D.; Sambasivan, N. Patient Trust and Perceptions of Privacy in AI-driven Mental Health Systems. *Nat. Digit. Med.* **2023**, *6*, 1–9.
37. American Psychological Association (APA). *Ethical Principles of Psychologists and Code of Conduct*; Technical Report; American Psychological Association: Washington, DC, USA, 2017.
38. Smith, A.; Doe, J. Impacts of Artificial Intelligence on Psychological Evaluation. *J. Appl. Psychol.* **2023**, *58*, 245–260. [CrossRef]
39. Wan, R.; Wan, R.; Xie, Q.; Hu, A.; Xie, W.; Chen, J.; Liu, Y. Current status and future directions of artificial intelligence in post-traumatic stress disorder: A literature measurement analysis. *Behav. Sci.* **2025**, *15*, 27. [CrossRef] [PubMed]

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.