information

MDPI

Systematic Review

AI Assistants Based on Large Language Models for Adolescent Health and Well-Being: A Systematic Review

Andreas Triantafyllidis [1,*], Sofia Segkouli [1], Evdoxia Eirini Lithoxoidou [1], Anastasios Alexiadis [1], Konstantinos Votis [1], Kleio Koutra [2], Vassilis Kilintzis [2], Haridimos Kondylakis [2], Eunate Arana-Arri [3,4], Severin Haug [5], Nikolaos Boumparis [5], Maria Krini [6], Liselot Hudders [7,8] and Dimitrios Tzovaras [1]

1 Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece; sofia@iti.gr (S.S.); elithoxo@iti.gr (E.E.L.); talex@iti.gr (A.A.); kvotis@iti.gr (K.V.); dimitrios.tzovaras@iti.gr (D.T.)
2 Hellenic Mediterranean University, 71410 Heraklion, Greece; kkoutra@hmu.gr (K.K.); billyk@ics.forth.gr (V.K.); kondylak@ics.forth.gr (H.K.)
3 Clinical Epidemiology Unit, Cruces University Hospital, Osakidetza, 48903 Barakaldo, Bizkaia, Spain; eunate.aranaarri@osakidetza.eus
4 Biobizkaia Health Research Institute, 48903 Barakaldo, Bizkaia, Spain
5 Swiss Research Institute for Public Health and Addiction ISGF, Zurich University, 8091 Zurich, Switzerland; severin.haug@isgf.uzh.ch (S.H.); niko.boumparis@isgf.uzh.ch (N.B.)
6 Cyprus Association of Cancer Patients and Friends (PASYKAF), 2000 Nicosia, Cyprus; mariakr@pasykaf.org
7 Center for Persuasive Communication, Department of Communication Sciences, Ghent University, 9000 Ghent, Belgium
8 Tilburg School of Humanities and Digital Sciences, Tilburg University, 5037 AB Tilburg, The Netherlands
* Correspondence: atriand@iti.gr; Tel.: +30-2311-257701

Abstract

Objective: Large Language Models (LLMs) are emerging as a key component for digital health systems based on Artificial Intelligence (AI). The objective of this paper is to systematically review the characteristics, effectiveness, and implementation challenges of LLM-based assistants targeting adolescent health and well-being. **Methods:** A systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. PubMed, Scopus, and Web of Science were searched in November 2025 for studies published from 2022 onwards. Eligible studies examined LLM-based assistants used directly by adolescents in health and well-being contexts and reported quantitative outcomes. Data was extracted independently by multiple reviewers and synthesised narratively. Risk of bias was assessed using the Mixed Methods Appraisal Tool (MMAT). **Results:** Nine studies met the inclusion criteria, involving between 3 and 40 participants and covering mental health support, physical activity promotion, treatment engagement, vaccination awareness, and academic self-efficacy. Most studies were proof-of-concept investigations or pilot studies. LLM-based assistants were associated with reductions in depression, anxiety, stress, and negative affect, improvements in emotional regulation, physical activity, treatment motivation, HPV knowledge, and academic self-efficacy. Risk of bias was generally moderate and the evidence base was limited by small samples, short follow-up periods, and heterogeneous methodologies. **Conclusions:** LLM-based assistants show promise as scalable and accessible tools for supporting adolescent health and well-being, particularly in mental health domains. However, the current evidence remains preliminary. Larger, long-term, and rigorously designed studies are required to establish effectiveness, safety, and implementation feasibility. **Registration:** This review was not pre-registered and no review protocol has been published prior to conducting the review.

Academic Editor: Tania Yankova

Received: 14 July 2026
Revised: 4 August 2026
Accepted: 8 August 2026
Published: 11 August 2026

Keywords: large language models; adolescence; generative AI; digital health; literature review

1. Introduction

Large Language Models (LLMs) represent a major advancement in generative Artificial Intelligence (AI), enabling machines to process, generate, and reason over natural language at unprecedented scale and fluency [1]. Built primarily on the Transformer architecture [2], LLMs are trained on vast corpora of text using self-supervised learning, allowing them to model complex linguistic patterns, contextual dependencies, and semantic relationships. Recent models—such as GPT, PaLM, and LLaMA—demonstrate capabilities extending beyond text generation to include conversational interaction, summarisation, reasoning, and context-aware recommendation, positioning them as a foundation technology for next-generation digital systems [3–5].

The healthcare domain has rapidly emerged as a prominent area of application for LLMs. LLM-empowered conversational assistants have shown potential to support symptom assessment, health education, clinical documentation, patient engagement, and decision support [6–8]. Unlike earlier rule-based chatbots, LLM-based systems can flexibly respond to nuanced user inputs, personalise explanations, and adapt interactions dynamically, thereby producing high-quality outputs that mimic human conversation styles and improving user experience [9]. These characteristics are particularly relevant for preventive care, health promotion, and behavioural interventions, where sustained engagement and clear communication are pivotal.

1.1. Adolescence as a Critical and Distinct Health Context

Adolescence is a formative developmental phase characterised by rapid biological maturation, cognitive development, identity formation, and evolving social relationships [10]. It is also a period during which many mental health conditions first emerge, health behaviours are established, and vulnerabilities related to stress, behavioural risk-taking, and social pressure intensify [11]. Interventions delivered during this life stage can yield substantial long-term health benefits, underscoring the importance of accessible, scalable, and youth-centred health support solutions.

Digital technologies are deeply embedded in adolescents' everyday lives, with high rates of smartphone use and online engagement for communication, information seeking, and social interaction [12]. Digital health interventions—particularly those delivered via conversational AI assistants—offer unique opportunities to meet adolescents "where they are," providing on-demand, non-judgmental access to health information and support [13]. LLM-powered assistants, by virtue of their conversational fluency and adaptability, may be especially well suited to addressing adolescents' diverse health needs, including mental well-being, physical activity, preventive care, and academic functioning [14,15].

1.2. Opportunities and Risks of LLM-Based Assistants for Adolescents

LLM assistants have several theoretical advantages in adolescent health contexts. They can deliver personalised psychoeducation, encourage positive behaviours, support self-reflection, and reduce barriers associated with stigma or limited access to professional services [16]. Studies suggest that conversational AI tools can enhance engagement, improve mental health literacy, and support behavioural change among users [17,18].

However, the deployment of LLMs in adolescent health also raises significant ethical, safety, and methodological concerns. These include risks of emotional dependency, misinformation or hallucinated content, inappropriate advice, data privacy violations, al-

gorithmic bias, and the absence of age-appropriate safeguards [19–21]. Adolescents may be particularly susceptible to over-reliance on AI systems or misinterpretation of health guidance without adequate contextualisation or professional oversight. Regulatory frameworks and clinical guidelines for the use of generative AI in pediatric and adolescent populations remain underdeveloped, highlighting the need for empirical evidence to inform safe and responsible implementation [22].

Although this review focuses on adolescents, LLM-based assistants have also been increasingly investigated in adult and older adult populations, where they have shown potential for supporting health education, chronic disease management, mental health, and patient engagement [7,8,23]. However, concerns regarding accuracy, safety, bias, privacy, and overreliance remain significant [20,24]. These findings provide useful context but may not directly translate to adolescent populations due to important developmental and psychosocial differences.

1.3. Need for a Systematic Review

Despite rapidly growing interest and progress in generative AI and LLMs for healthcare, evidence regarding the effectiveness, design characteristics, and limitations of LLM-based interventions for adolescent health remains fragmented [25]. Existing reviews have primarily examined conversational agents prior to the emergence and rapid growth of LLMs [26,27], focused mostly on adult populations [28,29], investigated a single technology, e.g., ChatGPT [24,30], or explored a single condition [31]. To address this gap, we present a systematic review of LLM-empowered assistants for adolescent health and well-being. To the best of our knowledge, this is the first review to focus specifically on LLM-based interventions targeting adolescents across multiple health domains. Rather than adopting a condition-specific lens, this review applies a cross-domain perspective to capture how LLMs are currently being designed, evaluated, and applied in diverse adolescent health contexts.

1.4. Objectives of This Review

The objective of this systematic review is to examine the characteristics, performance, benefits, and limitations of LLM-based assistants used to support adolescent health and well-being. Specifically, we synthesize evidence on study design, participant characteristics, intervention features, underlying LLM technologies, data sources, and reported outcomes. By evaluating effectiveness and identifying methodological and ethical challenges, this review aims to inform researchers, clinicians, developers, and policymakers seeking to design, evaluate, and implement safe and effective LLM-driven health solutions for adolescents. Ultimately, this work contributes to advancing evidence-based integration of generative AI into adolescent health promotion and care.

2. Methodology

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement [32]. The study selection process, data extraction, synthesis, and reporting were guided by the PRISMA 2020 checklist (provided as Supplementary Material S1) and flow diagram (Figure 1). This review was not prospectively registered and no review protocol was published prior to conducting the review.

We conducted a search of the Scopus, PubMed, and Web of Science databases to identify studies examining LLM-based assistants for adolescents in the context of health and well-being. The search was limited to manuscripts published from 2022 onward, marking the period when research interest in LLMs surged due to the landmark introduction of

ChatGPT. The inclusion criteria were: (a) the study should focus on LLM-based assistants used by adolescents; we defined adolescents as individuals aged 10 to 19 years, and included studies if their samples fell predominantly within this range—studies involving participants adjacent to the adolescent age range were also eligible if they addressed developmental stages overlapping with adolescence and adolescent-relevant health outcomes; (b) the study should target the health and well-being domain; this was defined broadly to include mental health, emotional regulation, physical health and activity, treatment engagement, health education, and academically related well-being (e.g., stress, self-efficacy); (c) the study should clearly highlight the utilization of one or more LLMs; (d) quantitative outcomes of the study should be presented; (e) the paper should be written in English.

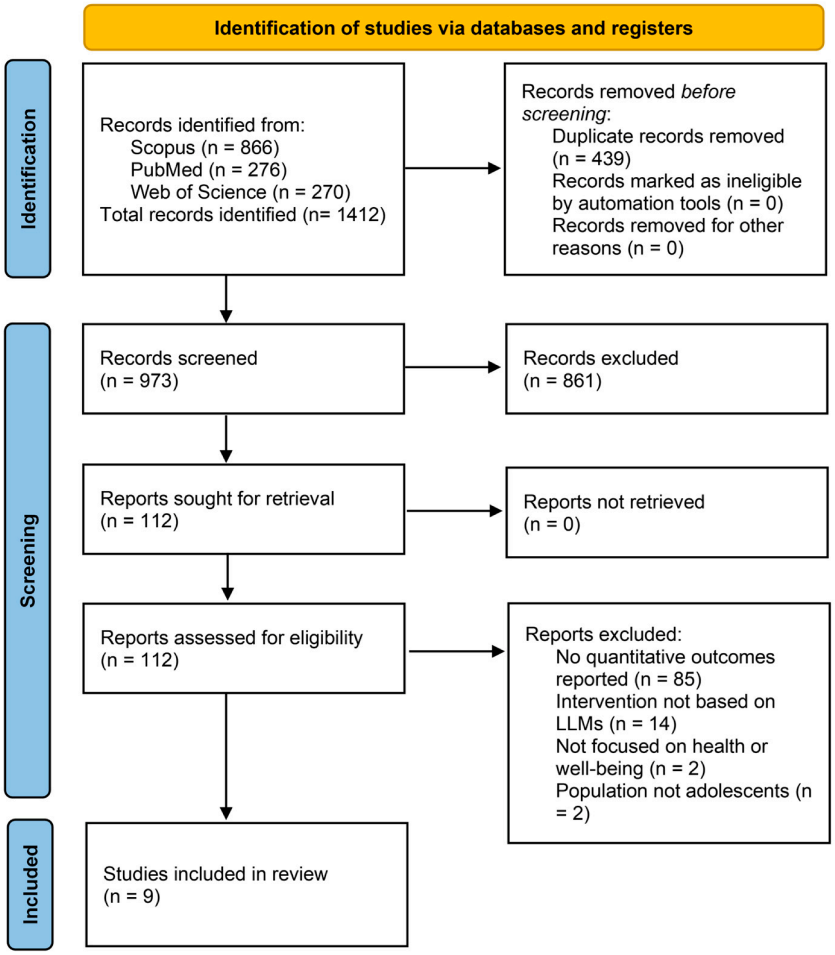

Figure 1. Prisma 2020 flow diagram for study selection.

We used a keyword query combining terms related to adolescence, large language models, AI assistance, as well as widely available chatbots, for search within the title, abstract and keywords of the manuscripts (Supplementary Material S2). Studies focusing primarily on unrelated domains—such as purely technical AI development, general education without a well-being component, or behaviours not explicitly linked to health or well-being (e.g., general media use, gaming, or non-health-related risk behaviours)—were excluded. Studies examining interventions or tools that were not focused on leveraging LLMs were excluded. Furthermore, studies focusing exclusively on the use of LLM-based tools by health professionals and not adolescents were not included. Studies not examining quantitative outcomes, surveys, and protocol papers were also excluded from the review.

Four reviewers (AT, SS, AA, and EEL) independently screened titles and abstracts against the predefined eligibility criteria. Full-text articles were subsequently assessed

independently by the same reviewers. Disagreements were resolved through discussion and consensus among the reviewers. No automation tools were used in the screening process. The four reviewers independently extracted data using a predefined extraction form. Information collected included publication details, study design, participant characteristics, intervention objectives, assistant features, LLMs used, data sources, study duration, outcomes, and reported limitations. Discrepancies were resolved through consensus.

Risk of bias was assessed using the Mixed Methods Appraisal Tool (MMAT, 2018 version) [33,34]. The MMAT was selected because of the heterogeneity of study designs included in this review, encompassing randomised, non-randomised, and descriptive studies. Two reviewers (AT, SS) independently assessed each study. Any disagreements were resolved through discussion and consensus. Due to the small number of studies, substantial heterogeneity in interventions and outcomes, and the absence of quantitative synthesis, formal assessment of certainty of evidence (e.g., GRADE) was not undertaken.

3. Results

3.1. Literature Search Outcomes

Our literature search was conducted in November 2025 using the Scopus, PubMed, and Web of Science databases to identify studies published from 2022 onward. This search retrieved 866 records from Scopus, 276 from PubMed, and 270 from Web of Science. Following the removal of duplicates using the Mendeley© reference management software (Elsevier B.V., Amsterdam, The Netherlands) [35] and the application of predefined inclusion and exclusion criteria, 112 articles were selected for full-text review. During the title and abstract screening stage, 861 records were excluded because they did not meet the eligibility criteria, most commonly due to the absence of an adolescent population, lack of an LLM-based intervention, or lack of relevance to health and well-being outcomes. The remaining 112 articles underwent full-text assessment. Of these, 103 articles were excluded for the following reasons: no quantitative outcomes reported (n = 85), intervention not based on LLMs (n = 14), no focus on adolescent health or well-being (n = 2), and population not involving adolescents (n = 2). Ultimately, 9 studies met the eligibility criteria and were included in the analysis [36–44]. A list of full-text articles excluded after eligibility assessment and the corresponding reasons for exclusion is provided in Supplementary Material S3.

3.2. Characteristics of Included Studies

The included studies (n = 9) demonstrate a diverse but still emerging body of evidence on LLM-based assistants for adolescent health and well-being. As shown in Table 1, the majority of studies focused on mental health support, while other application domains included promotion of physical activity, general well-being, vaccination uptake, and reduction in academic procrastination. Study designs were predominantly exploratory, with more than half employing proof-of-concept or experimental approaches, alongside a smaller number of randomised controlled trials and pre–post studies. Sample sizes were generally small, ranging from 3 to 40 participants, and several studies relied on very limited participant groups or even simulated cases. The age of participants typically spanned early to mid-adolescence (approximately 12–17 years), although one study [43] included younger participants (9–12 years). Intervention durations were also short, varying from single-session interactions lasting minutes to a maximum follow-up of six weeks. The included studies were conducted in a range of countries, including China, Japan, South Korea, Türkiye, Qatar, and the United States. Each intervention was evaluated within a single cultural and linguistic context, and no study reported cross-cultural or multinational evaluation.

Table 1. Characteristics of included studies.

Study	Target Domain	Study Design	Participants	Age	Study Duration
Ahmed et al. [36]	Promotion of well-being	Proof-of-concept experimental study	12	13–17 years	6 weeks
Aydemir [37]	Promotion of physical activity	Pre-post experimental study	26	14 years on average	4 weeks
Clark [38]	Mental health support	Fictional case vignettes study	3 fictional adolescents	N/A	Single encounter study lasting 15–30 min
Hasei et al. [39]	Mental health support	Pre-post experimental study	5	13–20 years	2 weeks
Hu et al. [40]	Mitigation of academic procrastination and stress reduction	Proof-of-concept experimental study	12	12–14 years	2 weeks
Kim et al. [41]	Mental health support	Proof-of-concept experimental study	10	13–17 years	Single encounter study
Ni et al. [42]	Mental health support	Proof-of-concept experimental study	21	14.7 years on average	Single encounter study
Steenstra et al. [43]	Promotion of vaccination	Randomised between-subjects experimental study	13	9–12 years	Single encounter study
Ye et al. [44]	Mental health support	Randomised controlled trial	40	12–14 years	2 weeks

3.3. Quality Assessment

Overall, the methodological quality of the included studies based on MMAT was found to be moderate (see Appendix A: Tables A1–A3). Randomised studies [43,44] demonstrated limitations in blinding and reporting of randomisation procedures. Non-randomised and pre–post studies [37,39,42] showed consistent weaknesses related to the absence of control groups and limited handling of confounders. Exploratory studies [36,40,41] were constrained by small samples and limited representativeness. One simulation-based study [38] demonstrated high applicability bias due to the use of fictional scenarios rather than real users.

3.4. Overview of LLM Applications for Adolescence

The LLM applications of the included studies summarised in Table 2 illustrate a wide range of objectives, features, and outcomes for adolescent health and well-being. Most interventions aimed to provide personalised support, including mental health counselling, emotional regulation, behaviour change, and health education, often leveraging conversational interfaces to enhance engagement. Several applications incorporated advanced features such as multimodal data integration, empathetic dialogue, natural voice interaction, or gamified environments to improve user experience. GPT-based models (particularly GPT-4) were most frequently used [37–39,41–44], followed by LLaMA-based systems [36,40]. Data sources varied considerably, with some studies utilising real-world datasets (e.g., wearable data, counselling transcripts, or emotional regulation datasets), while others relied on minimal or unspecified inputs. Reported outcomes were generally positive, indicating improvements in mental health indicators (e.g., reduced depression, anxiety, and stress) [41,44], enhanced emotional regulation [42], increased well-being including physical activity [36,37], greater treatment engagement [39], improved knowledge (e.g., vaccination awareness) [43], and stronger academic self-efficacy [40]. However, the findings were not uniformly positive. Clark et al. [38] demonstrated that AI chatbots may endorse harmful or inappropriate suggestions in certain mental health scenarios, raising important safety concerns. Similarly, Ahmed et al. [36] reported that although recommendations were generally clear and actionable, their alignment with underlying user data was inconsistent.

Table 2. Characteristics of LLM applications in adolescence.

Study	Key Objective	Main Features	LLM Used	Data Sources	Outcomes/Performance
Ahmed et al. [36]	To generate personalised recommendations for improving sleep hygiene, increasing physical activity, and enhancing academic performance	Individualised recommendations through leveraging multi-modal data	LLaMa	Fitbit data, survey data for sleep quality, school reports on academic performance	Clear and actionable recommendations (average score ~4/5), variable alignment with data (score < 4)
Aydemir [37]	To evaluate the feasibility and potential effectiveness of ChatGPT-delivered physical activity intervention	Parental participation through training sessions, ChatGPT-delivered physical activities to increase the physical activity levels of children with ASD	GPT	N/A	Significant increase in the physical activity levels of children
Clark [38]	To determine the willingness of therapy and companion AI chatbots to endorse harmful or ill-advised ideas in adolescents with mental distress	Testing of fictional clinical scenarios	GPT, Gemini, and custom LLMs	N/A	Chatbots and companions endorsed teenagers' highly problematic proposals almost one-third of the time
Hasei et al. [39]	To alleviate psychological distress and enhance treatment engagement in young cancer patients	Use of visual characters and conversation principles to enhance user engagement	GPT-4	N/A	All participants noted improved treatment motivation, with 80% disclosing personal concerns to the chatbot they had not shared with healthcare providers.
Hu et al. [40]	To investigate the underlying causes and potential interventions for academic procrastination	Empathetic learning companion, support of natural voice interactions, 3D printing	LLaMa	Facial images, user photos, linguistic data	Significant effectiveness in helping adolescents mitigate academic procrastination behaviours, assisting students in alleviating anxiety and reducing stress, enhancing adolescents' academic self-efficacy
Kim et al. [41]	To provide a human-like AI counselor to overcome psychological barriers	Chunk-based streaming methodology to enable human-like dialogue, prompt engineering	GPT-4	Real counseling scripts and synthetic counseling scripts	90.0% of adolescents self-reported that their concerns were resolved following counseling, and 70.0%indicated self-reported mood change
Ni et al. [42]	To explore adolescents' emotional regulation needs and build an AI emotional regulation system	Use of a dialogue model to provide more personalised and effective emotional regulation support	GPT-2	Chinese Emotional Regulation Dataset with 400 dialogue samples between therapists and clients	Significant difference in negative affect scores between the pre-test and post-test, no difference in positive affect scores; the system provides improved emotional regulation support
Steenstra et al. [43]	To assess HPV knowledge and attitudes, HPV vaccine knowledge and attitudes, general vaccine knowledge and acceptability, intent to vaccinate, usability and satisfaction, engagement, and entertainment	Use of a fantasy narrative game	GPT-4	N/A	Statistically significant pre-to-post improvements in HPV knowledge
Ye et al. [44]	To assess the effect of an LLM on alleviating depression and anxiety	WarmGPT: A conversational mental health service robot	GPT-4	N/A	Statistically significant reduction in depression, anxiety and negative emotions, and improvement of positive affect

3.5. Description of Applications

We briefly describe all the included studies below in terms of intervention purpose and content, evaluation outcomes, and implications for clinical practice. For presentation purposes, the included studies were grouped according to their primary intervention objective and main reported outcomes. Studies addressing depression, anxiety, counseling, or emotional regulation were categorized under Mental Health and Emotional Regulation; studies focusing on well-being promotion, academic functioning, or physical activity were grouped under Well-being, Academic Performance, and Physical Activity; studies aimed at improving treatment adherence or engagement were classified under Treatment Engagement; and studies targeting vaccination-related knowledge, attitudes, or intentions were categorized under Vaccination Awareness.

3.5.1. Mental Health and Emotional Regulation

Kim et al. [41] developed BetterMood, an LLM-based AI counseling system designed to provide accessible, human-like mental health support for adolescents by incorporating psychotherapeutic techniques and multimodal interaction. The system was evaluated through a user study involving adolescents, young adults, and clinicians, showing high satisfaction in interaction quality, usability, and safety, with adolescents reporting improved mood and perceived resolution of concerns. The findings suggest that such tools can serve as effective supplementary support for early psychological intervention, although they should complement rather than replace professional clinical care.

Ye et al. [44] evaluated WarmGPT, an LLM-based conversational agent integrating cognitive-behavioural therapy to support adolescent mental health through voice-based interactions. In a randomised controlled trial with 40 secondary school students over two weeks, the intervention significantly reduced depression, anxiety, and negative affect while enhancing positive emotions compared to a control group receiving educational videos on mental health. These findings indicate that LLM-based assistants can provide effective short-term mental health support for adolescents and may serve as accessible, scalable complements to traditional care.

Ni et al. [42] developed an LLM-based chatbot for adolescent emotional regulation, designed using a user-centered approach that incorporates adolescents' real-world emotional needs and a custom dialogue dataset tailored to this population. The system, built on a fine-tuned GPT-2 model with an emotion-weighting mechanism, demonstrated improved performance in generating personalised emotional support, and evaluation showed a significant reduction in negative emotions among adolescent users, although effects on positive affect were limited. These findings suggest that LLM-based tools can effectively support emotional regulation in adolescents, while highlighting the importance of personalisation and population-specific design for real-world applications.

Clark [38] investigated the safety of AI therapy chatbots by testing whether they would appropriately reject harmful or ill-advised suggestions from fictional adolescents experiencing mental health distress. In a simulation study involving 10 publicly available chatbots and 60 scenarios, the bots endorsed harmful proposals in 32% of cases, with none consistently rejecting all dangerous suggestions, highlighting variability in their ability to set appropriate boundaries. These findings raise significant concerns about the reliability and safety of AI-based mental health support for adolescents, emphasizing the need for stronger oversight, ethical safeguards, and further research before widespread adoption.

3.5.2. Well-Being, Academic Performance, and Physical Activity

Ahmed et al. [36] developed a proof-of-concept LLM-based system that integrates wearable data, survey measures, and academic reports to generate personalised recom-

mendations aimed at improving adolescents' well-being and academic performance. The intervention used multimodal student data (e.g., sleep, activity, stress, and school feedback) to produce tailored advice on sleep hygiene, stress management, physical activity, and study habits, with evaluation by independent raters showing that recommendations were generally clear and actionable, though alignment with individual data varied. These findings suggest that LLM-driven, data-informed recommendations hold potential for personalised adolescent health support, but further validation, larger studies, and real-world interventions are needed to establish their effectiveness and reliability.

Hu et al. [40] introduced LittleToDo, an LLM-driven intervention system that uses affective computing and an AI companion to help adolescents reduce academic procrastination through personalised task planning, emotional support, and therapeutic activities. The system monitors users' emotional states and attention, dynamically adapts study schedules, and provides motivational feedback. An experimental 2-week user study with 12 adolescents showed significant reductions in procrastination and stress alongside improvements in self-efficacy and task completion. These findings suggest that emotionally aware, LLM-based tools could potentially support adolescent learning behaviours, though further longitudinal research with larger samples is needed to confirm those findings.

Aydemir [37] investigated a ChatGPT-delivered physical activity intervention aimed at increasing activity levels in children with autism spectrum disorder, with parents using the system to deliver tailored, home-based exercise programs. In a 4-week pre-post experimental study with parent–child dyads, the intervention was found to be acceptable and engaging, and resulted in improved children's physical activity levels. These findings suggest that LLM-based interventions can support accessible, home-based health promotion for special populations, although further research is needed to confirm long-term effectiveness and scalability in clinical practice.

3.5.3. Treatment Engagement

Hasei et al. [39] evaluated GPT-4–based conversational AI chatbots designed to provide empathetic, continuous psychological support for adolescent cancer patients. In a 2-week pre-post pilot study with five patients, four participants reported reduced anxiety and stress, while change in treatment engagement and increased willingness to disclose personal concerns were also noted. The findings from this early experimental study suggest that generative AI chatbots may serve as accessible complementary tools for emotional support in vulnerable clinical populations.

3.5.4. Vaccination Awareness

Steenstra et al. [43] explored the use of LLMs and virtual agents to develop gamified, narrative-based health education interventions for adolescents, focusing on HPV vaccination knowledge and attitudes. This single-encounter experimental study with 13 participants demonstrated that both an LLM-empowered fantasy narrative game and a traditional pedagogical virtual agent improved adolescents' knowledge and attitudes toward HPV; however, the LLM game-based approach showed higher engagement and entertainment than the traditional virtual agent. These findings suggest that LLM-enabled gamification may enrich the educational experience for adolescents.

3.6. Challenges of LLM-Based Assistants in Adolescence

The challenges presented in this section emerged from a narrative synthesis of the limitations, risks, and implementation barriers reported across the included studies. Similar issues were grouped into broader thematic categories to facilitate interpretation of the evidence, resulting in themes related to clinical reliability, hallucinations and accuracy,

data-related challenges, technical constraints, user engagement, cultural adaptation, ethical concerns, and limitations of the current evidence base.

3.6.1. Clinical Reliability

The issue of safety and clinical reliability remains one of the most critical concerns in LLM-based interventions. Clark's study [38] highlights that AI therapy chatbots can endorse harmful or ill-advised behaviours in a substantial proportion of cases, reflecting weak boundary-setting and inadequate risk recognition. In educational contexts, LLM-generated content can contain inaccuracies or misleading information, requiring expert review to ensure validity [37,41,42]. Even in more controlled systems such as WarmGPT [44], safeguards—such as expert review of conversations and integration with human oversight—are necessary to mitigate risks. Across studies, there is a clear implication that AI systems cannot yet independently manage high-risk psychological situations, reinforcing the need for human supervision and escalation mechanisms. Furthermore, the lack of contextualization and personalisation can reduce perceived usefulness and weaken the therapeutic or educational impact of LLM interventions [37,39].

3.6.2. Hallucinations

Another recurring challenge is hallucinations and variability in accuracy. LLMs may produce outputs that are coherent but factually incorrect or poorly aligned with user data [43]. This issue is evident both in health education contexts and in personalised recommendation systems. In Ahmed's study [36], although recommendations were generally clear and actionable, alignment with actual student data was inconsistent, and some outputs were unusable due to missing or unreliable inputs. This highlights that accuracy depends not only on the model but also on data quality and prompt structure, making validation a critical requirement.

3.6.3. Data-Related Challenges

Data-related challenges extend further to input reliability and integration of multimodal data. Systems that rely on wearables, surveys, and behavioural data face issues such as missing values, device errors, and discrepancies between objective and self-reported measures. For example, extreme or inaccurate sleep data from wearable devices led to questionable outputs, emphasizing that poor-quality data can propagate errors into AI-generated recommendations. More broadly, these findings show that LLM performance is fundamentally constrained by the quality, consistency, and completeness of input data [36].

3.6.4. Technical Constraints of LLM Tools

From an implementation perspective, studies also highlight technical and usability constraints for the LLM tools. These include token limits, latency issues, and dependence on structured prompts or multi-stage pipelines. In addition, systems may require significant training, parental involvement, or system scaffolding to be used effectively, limiting their accessibility as standalone tools. This challenges the assumption that AI interventions can be easily deployed at scale without additional support mechanisms [38].

3.6.5. User Engagement

From a design perspective, balancing engagement, usability, and effectiveness remains a challenge. While features such as gamification, conversational interfaces, and human-like avatars may be engaging [43], sustaining long-term engagement is difficult, as initial user interest may be influenced by novelty effects rather than lasting value. Systems must therefore strike a balance between being engaging and being clinically or educationally effective [41].

3.6.6. Cultural Adaptation

Cultural and contextual adaptation is another cross-cutting issue. Several studies show that systems must be localized to specific languages, educational systems, and cultural norms to be effective [42]. Without such adaptation, AI tools risk being irrelevant or ineffective for diverse populations, raising concerns about equity and inclusivity [44].

3.6.7. Ethical Concerns

Ethical concerns remain central, particularly regarding privacy, consent, and data governance for minors. Many systems collect and process sensitive information such as emotional disclosures, behavioural patterns, and health metrics. While measures such as anonymization, local data storage, and restricted access are implemented, risks remain—especially when systems store longitudinal interaction data for personalisation. Additionally, the need to balance privacy with safety monitoring (e.g., reviewing conversations to detect risk) creates ongoing ethical tension [36,38].

3.6.8. Limited Evidence Base on the Effectiveness of LLM Tools in Adolescent Health

An important challenge is the limited and early-stage evidence base on the effectiveness of LLM-based tools in adolescent health and well-being. The included studies relied on small samples, short intervention periods, and self-reported outcomes. Even in stronger designs, such as Ye et al.'s randomised controlled trial [44], the intervention lasted only two weeks, leaving long-term effectiveness, sustainability, and generalizability uncertain. Similarly, proof-of-concept studies emphasize feasibility rather than causal impact, underscoring the need for larger-scale and longitudinal research [36–38,41,43,44].

4. Discussion

4.1. Main Findings

We conducted a systematic review of AI assistants based on LLMs for adolescent health and well-being. Overall, the findings indicate that these systems demonstrate promising early effectiveness across multiple domains, particularly in mental health, but remain at an early stage of development with significant methodological, clinical, and ethical limitations.

A consistent finding across studies is that LLM-based assistants can produce positive short-term outcomes, particularly in mental health contexts. Interventions such as BetterMood [41] and WarmGPT [44] showed improvements in mood, reductions in depression and anxiety, and enhanced emotional regulation, while other systems supported treatment engagement and self-disclosure among vulnerable populations such as adolescent cancer patients [39]. Similarly, LLM-driven tools demonstrated benefits beyond mental health, including increased physical activity [36,37], improved vaccination knowledge [43], and enhanced academic self-efficacy [40], highlighting the cross-domain applicability of these systems.

These findings suggest that LLM conversational assistants may be particularly well-suited for preventive and supportive roles, where accessibility, 24/7 availability, and user engagement are critical. The ability of these systems to provide immediate, always-available support represents a significant advantage compared to traditional services, particularly in settings where access to professional care is limited [45].

However, despite these promising outcomes, the review also highlights that the current evidence base is limited in scope, scale, and robustness. Most studies involved small samples (often fewer than 20 participants), short intervention periods (frequently $\leq$2–4 weeks), and exploratory or proof-of-concept designs. Even in randomised trials, follow-up durations were brief, restricting conclusions about the long-term sustainability, clinical signifi-

cance, and generalisability of observed effects. This indicates that current findings should be interpreted as preliminary rather than definitive evidence of effectiveness.

Beyond effectiveness, the review identifies a set of multi-dimensional challenges that constrain the safe and reliable deployment of LLM assistants for adolescents. First, safety and clinical reliability emerge as critical concerns. Evidence shows that AI systems may fail to appropriately handle high-risk situations or may even endorse harmful suggestions, raising serious concerns for their use in mental health contexts [20,38]. In addition, hallucinations and inaccuracies—especially in medical or educational contexts—can undermine trust and potentially lead to harmful or misleading guidance [46]. These risks emphasize that current systems lack the robustness required for unsupervised clinical use.

The review also highlights challenges related to data quality and multimodal integration. Systems that rely on wearable data, behavioural inputs, or self-reported measures are vulnerable to missing, noisy, or inconsistent data, which can negatively affect output quality. This underscores a fundamental limitation: LLM outputs are only as reliable as the data they are based on, making data validation and preprocessing essential components of system design [47,48].

Ethical concerns are pervasive. The collection and processing of sensitive adolescent data—including emotional states, behavioural patterns, and health information—raise issues related to privacy, consent, and data governance. Moreover, the potential for emotional dependence on AI and the absence of clear regulatory frameworks further complicate the responsible deployment of these systems [19,49].

Finally, the review underscores broader implementation and role-related challenges. LLM-based assistants often require structured prompts, human oversight, or training support, limiting their scalability as fully autonomous tools. At the same time, there is a risk that adolescents may over-rely on AI systems or use them as substitutes for professional care [50]. Across studies, there is strong consensus that these systems should be positioned as complementary tools within human-centered care models, rather than replacements for clinicians, educators, or caregivers.

An important consideration when interpreting these findings is the heterogeneity of the AI systems included in this review. Although all the interventions were based on LLM technologies, they ranged from general-purpose foundation models such as GPT-4 and LLaMA, to fine-tuned dialogue models trained for specific emotional regulation tasks, as well as publicly available conversational agents. These systems differ substantially in their underlying architecture, training procedures, degree of customisation, interaction modality (e.g., text-only, voice-based, multimodal, or game-based), and level of autonomy. Such differences can influence key outcomes including personalisation, clinical reliability, explainability, user engagement, and safety. Consequently, conclusions regarding effectiveness and safety should be interpreted within the context of the specific type of system evaluated, and findings from one class of LLM-based assistants may not be directly transferable to another.

Another consideration when interpreting these findings is the developmental heterogeneity of the populations studied. Although this review focused predominantly on adolescence, participants in the included studies ranged from children aged 9–12 years to older adolescents and young adults up to 20 years of age. These groups differ in cognitive maturity, emotional regulation, health literacy, autonomy, parental involvement, and consent requirements, all of which may influence the effectiveness, acceptability, and safety of LLM-based interventions. Consequently, findings observed in one age group should not be assumed to generalise to other developmental stages, highlighting the need for age-specific design and evaluation of future systems [51].

The generalisability of the findings is further limited by the diversity of the populations studied, which ranged from generally healthy adolescents to young people with psychological distress, autism spectrum disorder, and cancer. Given differences in clinical needs, vulnerability, and support requirements, findings related to effectiveness and safety should not be assumed to transfer directly across different adolescent groups.

Cultural and linguistic considerations also warrant attention when interpreting the findings. The interventions in the included studies were evaluated within a single cultural and linguistic setting. Conversational norms, emotional expression, attitudes toward mental health, family involvement, and access to healthcare services vary across cultures and may influence user engagement, trust, and intervention effectiveness. Consequently, findings from one cultural context should not be assumed to generalise directly to others, highlighting the need for cross-cultural co-creation, evaluation and adaptation of future LLM-based interventions [52].

4.2. Implications for Research and Practice

Taken together, these findings have several important implications for both future research and real-world implementation of LLM-based assistants for adolescent health and well-being. From a research perspective, there is a clear need for larger, longitudinal, and rigorously designed studies, including randomized controlled trials with extended follow-up, to establish the long-term effectiveness, safety, and sustainability of LLM-based interventions. Beyond study size and duration, future research should also prioritise methodological standardisation, including the adoption of validated outcome measures and established reporting frameworks such as CONSORT-AI [53], TIDieR [54], and TRIPOD-LLM [55]. Such standardisation would improve the transparency and reproducibility of intervention reporting, enhance comparability across studies, and facilitate future evidence synthesis and meta-analytical approaches. As the evidence base matures and a greater number of comparable studies become available, future reviews should consider conducting meta-analyses to quantify the effectiveness of LLM-based interventions across different health and well-being domains. Furthermore, given the rapid pace of development in generative AI technologies, a living systematic review approach may be particularly valuable for maintaining an up-to-date synthesis of the evolving evidence base and informing research, practice, and policy [56].

From a design and technical perspective, future systems must go beyond general-purpose conversational capabilities and focus on enhanced personalisation, contextual awareness, and emotional intelligence [57]. Adolescents require developmentally appropriate, context-sensitive interactions, which cannot be achieved through generic LLM outputs alone. This will likely require: fine-tuning on population-specific datasets, integration of context-aware inputs (e.g., behavioural or environmental data), and continuous adaptation through feedback loops and user modeling [58,59].

At the same time, the findings highlight the need for robust quality assurance mechanisms. Current studies are often unclear about how output quality is ensured, particularly regarding factual accuracy and clinical appropriateness. In this direction future systems should incorporate: human-in-the-loop validation (e.g., clinician review of outputs), safety filters and guardrails to prevent harmful or inappropriate responses, structured testing frameworks, including scenario-based evaluation of high-risk situations (e.g., crisis or self-harm dialogues), as well as monitoring and auditing mechanisms to detect hallucinations, bias, or unsafe outputs [60,61]. In addition, integration with validated clinical frameworks (e.g., CBT, motivational interviewing) is essential to ensure that outputs are not only engaging but also therapeutically meaningful and evidence-based [62].

From an ethical and regulatory standpoint, the findings underscore the urgent need for clear, adolescent-specific guidelines for generative AI use [63]. While many studies mention anonymisation or basic privacy protections, they often lack detailed descriptions of data governance practices. Future research and implementations should explicitly address: data collection, storage, and retention policies; informed consent procedures for minors and guardians; transparency regarding how user data are used and processed; and mechanisms for safeguarding vulnerable users, particularly in mental health contexts. Moreover, ethical design should extend beyond privacy to include risk mitigation and accountability [64]. This includes defining when and how systems should escalate to human support, how responsibility is assigned for AI-generated advice, and how users are informed about system limitations. The development of regulatory frameworks and certification standards for AI tools in adolescent health will be critical to ensure safe and trustworthy deployment [65].

From a practical and implementation perspective, preliminary findings support the positioning of LLM-based assistants as accessible, scalable support tools that complement existing health and educational services [66]. Their strengths lie in providing on-demand, continuous, and non-judgmental support, which is particularly valuable for early intervention, self-management, and engagement. However, effective implementation will require integration within hybrid care models involving clinicians, educators, and caregivers, user training and guidance to ensure appropriate use and avoid overreliance, adaptation to local cultural, linguistic, and institutional contexts, and evaluation of long-term engagement and real-world adoption barriers [67].

Addressing the current gaps in transparency, safety validation, and ethical design will be essential to move from promising prototypes to trustworthy, real-world applications. An additional direction for future research is the exploration of multi-agent LLM architectures. While most systems identified in this review relied on a single conversational agent, recent advances in AI have demonstrated the potential of coordinated teams of specialised LLM agents for collaborative problem solving, information retrieval, verification, and decision support [68,69]. In the context of adolescent health, multi-agent approaches could enable the separation of key functions across dedicated agents, such as symptom assessment, evidence retrieval, personalisation, safety monitoring, and coordination by human professionals (human-in-the-loop). Such architectures may improve reliability, reduce the risk of hallucinations, and provide additional safeguards through internal verification and review processes. Although these approaches have not yet been systematically evaluated in adolescent health settings, they represent a promising technical foundation for the development of more trustworthy, explainable, and human-centred AI health assistants.

4.3. Limitations

This review should be interpreted within the context of several limitations. First, the literature search was based on a predefined set of keywords related to LLMs, conversational assistants, and adolescent populations, which may have led to the omission of relevant studies that used alternative terminology (e.g., broader AI terms or domain-specific descriptions) or did not explicitly mention LLMs in titles, abstracts, or keywords.

Second, the database search was limited to PubMed, Scopus, and Web of Science. Although these are widely used and comprehensive databases, this methodological choice might have introduced selection bias and limited the completeness of the evidence base. Future updates of this review could expand the search to additional databases (e.g., Embase, PsycINFO, and Cochrane Library) to further enhance coverage and comprehensiveness.

Third, the included studies were highly heterogeneous in terms of design, population characteristics, intervention types, outcome measures, and target domains (e.g., mental

health, physical activity, education). This heterogeneity prevented the conduction of a meta-analysis and limited the ability to draw strong quantitative conclusions or directly compare effect sizes across studies. The diversity of outcome measures and study designs limits the ability to draw definitive or generalisable conclusions.

Fourth, the review is constrained by the limitations of the included studies themselves. Most studies relied on small sample sizes, short intervention durations, and in some cases experimental or simulated settings rather than real-world environments, which may have affected the generalisability and robustness of the findings.

Fifth, the review focused exclusively on studies published after 2022, reflecting the rapid emergence of LLM technologies following the introduction of ChatGPT. While appropriate for capturing the most recent developments, this restriction may have excluded earlier research on conversational agents or related AI systems that could provide valuable contextual insights.

Finally, another limitation relates to the potential for publication bias and selective outcome reporting. Most included studies reported positive findings regarding feasibility, acceptability, engagement, or short-term effectiveness despite being based on relatively small samples and exploratory designs. It is possible that studies reporting null results, implementation failures, safety concerns, or commercially sensitive findings are less likely to be published or identified through the literature search. Consequently, the generally positive picture presented in the current evidence base may not fully reflect the true balance of benefits, risks, and limitations of LLM-based interventions for adolescent health and well-being. The generalisability of the findings is further limited by the small number of eligible studies (n = 9) identified in this emerging field. As a result, the conclusions drawn should be considered preliminary, and there is a clear need for further research to provide a more comprehensive and representative evidence base. The protocol of this study was not pre-registered.

5. Conclusions

In summary, this review demonstrates that LLM-based assistants hold significant potential to transform adolescent health support, particularly through their scalability, accessibility, and ability to engage users in personalised interactions. However, this potential is currently constrained by technical limitations, safety concerns, ethical challenges, and a limited evidence base. At present, the most appropriate role of these systems is as supervised, complementary tools embedded within broader human support systems, rather than as standalone solutions. Continued interdisciplinary research—combining advances in AI, clinical science, and ethics—will be essential to realise their safe and effective integration into adolescent health and well-being.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info17080770/s1, Supplementary Material S1: PRISMA 2020 Checklist; Supplementary Material S2: Search Strategy; Supplementary Material S3: Full-Text Articles Excluded after Eligibility Assessment with Reasons for Exclusion.

Author Contributions: Conceptualization, A.T.; methodology, A.T., S.S., A.A. and E.E.L.; writing—original draft preparation, A.T.; writing—review and editing, A.T., S.S., E.E.L., A.A., K.V., K.K., V.K., H.K., E.A.-A., S.H., N.B., M.K., L.H. and D.T.; supervision, A.T.; project administration, A.T.; funding acquisition, A.T. All authors have read and agreed to the published version of the manuscript.

Funding: All authors were supported through funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101136829 (SUNRISE).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data extracted and analyzed during this review are contained within the article and its Supplementary Materials. Additional information may be obtained from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. Quality assessment of randomized studies based on MMAT.

Study	Randomisation	Baseline Comparable	Complete Data	Blinding	Adherence	Overall
Steenstra et al. [43]	Somewhat	Yes	Yes	No	Yes	Moderate
Ye et al. [44]	Yes	Yes	Yes	No	Yes	Moderate

Table A2. Quality assessment of non-randomised studies based on MMAT.

Study	Representative	Measurement Appropriate	Complete Data	Confounders Handled	Intervention Fidelity	Overall
Aydemir [37]	Yes	Yes	Somewhat	No	Yes	Moderate
Hasei et al. [39]	Yes	Yes	No	No	Yes	Moderate
Ni et al. [42]	Somewhat	Yes	No	No	Yes	Moderate

Table A3. Quality assessment of descriptive exploratory studies.

Study	Sampling Strategy	Representativeness	Measurement Appropriate	Non-Response Bias	Analysis Appropriate	Overall
Ahmed et al. [36]	Yes	No	Yes	Unclear	Yes	Moderate
Clark [38]	No	No	No	Unclear	Somewhat	High
Hu et al. [40]	Yes	Somewhat	Yes	Unclear	Yes	Moderate
Kim et al. [41]	Yes	Somewhat	Yes	Unclear	Yes	Moderate

References

1. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef] [PubMed]
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
3. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2022**, arXiv:2108.07258.
4. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
5. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–72. [CrossRef]
6. Jiang, L.Y.; Liu, X.C.; Nejatian, N.P.; Nasir-Moin, M.; Wang, D.; Abidin, A.; Eaton, K.; Riina, H.A.; Laufer, I.; Punjabi, P.; et al. Health System-Scale Language Models Are All-Purpose Prediction Engines. *Nature* **2023**, *619*, 357–362. [CrossRef] [PubMed]
7. Triantafyllidis, A.; Segkouli, S.; Kokkas, S.; Alexiadis, A.; Lithoxoidou, E.E.; Manias, G.; Antoniades, A.; Votis, K.; Tzovaras, D. Large Language Models for Cardiovascular Disease, Cancer, and Mental Disorders: A Review of Systematic Reviews. *Healthcare* **2026**, *14*, 45. [CrossRef] [PubMed]
8. Maity, S.; Saikia, M.J. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* **2025**, *12*, 631. [CrossRef] [PubMed]
9. Karanikolas, N.N.; Manga, E.; Samaridi, N.; Stergiopoulos, V.; Tousidou, E.; Vassilakopoulos, M. Strengths and Weaknesses of LLM-Based and Rule-Based NLP Technologies and Their Potential Synergies. *Electronics* **2025**, *14*, 3064. [CrossRef]
10. Sawyer, S.M.; Afifi, R.A.; Bearinger, L.H.; Blakemore, S.-J.; Dick, B.; Ezeh, A.C.; Patton, G.C. Adolescence: A Foundation for Future Health. *Lancet* **2012**, *379*, 1630–1640. [CrossRef] [PubMed]

11. Tarasenko, A.; Josy, G.; Minnis, H.; Hall, J.; Danese, A.; Lau, J.Y.F.; Cortese, S.; Stringaris, A.; Redlich, C.; Ougrin, D. Mental Health of Children and Young People in the WHO Europe Region. *Lancet Reg. Health Eur.* **2025**, *57*, 101459. [CrossRef] [PubMed]
12. Ricoy, M.-C.; Martínez-Carrera, S.; Martínez-Carrera, I. Social Overview of Smartphone Use by Teenagers. *Int. J. Environ. Res. Public Health* **2022**, *19*, 15068. [CrossRef] [PubMed]
13. Hong, S.H.; Chun, T.K.; Nam, Y.J.; Kim, T.W.; Cho, Y.H.; Son, S.J.; Roh, H.W.; Hong, C.H. Digital Mental Health Interventions for Adolescents: An Integrative Review Based on the Behavior Change Approach. *Children* **2025**, *12*, 770. [CrossRef] [PubMed]
14. Fostier, J.; Leemans, E.; Meeussen, L.; Wulleman, A.; Van Doren, S.; De Coninck, D.; Toelen, J. Dialogues with AI: Comparing ChatGPT, Bard, and Human Participants' Responses in In-Depth Interviews on Adolescent Health Care. *Future* **2024**, *2*, 30–45. [CrossRef]
15. Wu, Y.; Wu, T.; Zhu, K.; Liu, X.; Liu, J.; Wang, Y.; Shi, R.; Xiong, J.; Xing, X.; Lv, Y.; et al. The Impact of Chatbots on Adolescent Mental Health Development: A Comprehensive Literature Review. *J. Multidiscip. Healthc.* **2026**, *19*, 579872. [CrossRef] [PubMed]
16. Matthews, E.; Cleary, F.; Firth, J. Considerations about the Proliferation of Large Language Model Chatbots and Youth Mental Health. *Ir. J. Psychol. Med.* **2026**, 1–4. [CrossRef] [PubMed]
17. Fulmer, R.; Joerin, A.; Gentile, B.; Lakerink, L.; Rauws, M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment. Health* **2018**, *5*, e9782. [CrossRef] [PubMed]
18. Voultsiou, E.; Moussiades, L. A Systematic Review of Large Language Models in Mental Health: Opportunities, Challenges, and Future Directions. *Electronics* **2026**, *15*, 524. [CrossRef]
19. Saracini, C.; Cornejo-Plaza, M.I.; Cippitani, R. Techno-Emotional Projection in Human–GenAI Relationships: A Psychological and Ethical Conceptual Perspective. *Front. Psychol.* **2025**, *16*, 1662206. [CrossRef] [PubMed]
20. Harrer, S. Attention Is Not All You Need: The Complicated Case of Ethically Using Large Language Models in Healthcare and Medicine. *eBioMedicine* **2023**, *90*, 104512. [CrossRef] [PubMed]
21. Nagata, J.M.; Memon, Z.; Huang, O.; Moreno, M.A. Adolescent Health and Generative AI-Risks and Benefits. *JAMA Pediatr.* **2026**, *180*, 7–8. [CrossRef] [PubMed]
22. Laux, J.; Wachter, S.; Mittelstadt, B. Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk. *Regul. Gov.* **2024**, *18*, 3–32. [CrossRef] [PubMed]
23. Zhao, Y.; Guo, R.; Miao, Y.; Luo, Y.; Wang, H.; Wu, Y. Behavior Change Content and Implementation of Large Language Model–Driven Conversational Agents in Cardiometabolic Care: Scoping Review. *J. Med. Internet Res.* **2026**, *28*, e89190. [CrossRef] [PubMed]
24. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef] [PubMed]
25. Li, H.; Zhang, R.; Lee, Y.-C.; Kraut, R.E.; Mohr, D.C. Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being. *npj Digit. Med.* **2023**, *6*, 236. [CrossRef] [PubMed]
26. Laranjo, L.; Dunn, A.G.; Tong, H.L.; Kocaballi, A.B.; Chen, J.; Bashir, R.; Surian, D.; Gallego, B.; Magrabi, F.; Lau, A.Y.S.; et al. Conversational Agents in Healthcare: A Systematic Review. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1248–1258. [CrossRef] [PubMed]
27. Milne-Ives, M.; de Cock, C.; Lim, E.; Shehadeh, M.H.; de Pennington, N.; Mole, G.; Normando, E.; Meinert, E. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J. Med. Internet Res.* **2020**, *22*, e20346. [CrossRef] [PubMed]
28. Li, Y.; Liang, S.; Zhu, B.; Liu, X.; Li, J.; Chen, D.; Qin, J.; Bressington, D. Feasibility and Effectiveness of Artificial Intelligence-Driven Conversational Agents in Healthcare Interventions: A Systematic Review of Randomized Controlled Trials. *Int. J. Nurs. Stud.* **2023**, *143*, 104494. [CrossRef] [PubMed]
29. Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J.A.; Wornow, M.; Swaminathan, A.; Lehmann, L.S.; et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* **2025**, *333*, 319–328. [CrossRef] [PubMed]
30. Li, J.; Dada, A.; Puladi, B.; Kleesiek, J.; Egger, J. ChatGPT in Healthcare: A Taxonomy and Systematic Review. *Comput. Methods Programs Biomed.* **2024**, *245*, 108013. [CrossRef] [PubMed]
31. Omar, M.; Levkovich, I. Exploring the Efficacy and Potential of Large Language Models for Depression: A Systematic Review. *J. Affect. Disord.* **2025**, *371*, 234–244. [CrossRef] [PubMed]
32. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
33. Hong, Q.N.; Pluye, P.; Fàbregues, S.; Bartlett, G.; Boardman, F.; Cargo, M.; Dagenais, P.; Gagnon, M.-P.; Griffiths, F.; Nicolau, B.; et al. Improving the Content Validity of the Mixed Methods Appraisal Tool: A Modified e-Delphi Study. *J. Clin. Epidemiol.* **2019**, *111*, 49–59.e1. [CrossRef] [PubMed]
34. Souto, R.Q.; Khanassov, V.; Hong, Q.N.; Bush, P.L.; Vedel, I.; Pluye, P. Systematic Mixed Studies Reviews: Updating Results on the Reliability and Efficiency of the Mixed Methods Appraisal Tool. *Int. J. Nurs. Stud.* **2015**, *52*, 500–501. [CrossRef] [PubMed]

35. Mohammadi, E.; Thelwall, M.; Haustein, S.; Larivière, V. Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 1832–1846. [CrossRef]
36. Ahmed, A.; Aziz, S.; Abd-alrazaq, A.A.; AlSaad, R.; Sheikh, J.I. Leveraging LLMs and Wearables to Provide Personalized Recommendations for Enhancing Student Well-Being and Academic Performance through a Proof of Concept. *Sci. Rep.* **2025**, *15*, 4591. [CrossRef] [PubMed]
37. Aydemir, U. ChatGPT-Delivered Physical Activity Intervention for Children With Autism Spectrum Disorder: Pre-Post Feasibility Study. *JMIR Hum. Factors* **2025**, *12*, e71119. [CrossRef] [PubMed]
38. Clark, A.B. The Ability of AI Therapy Bots to Set Limits with Distressed Adolescents: Simulation-Based Comparison Study. *JMIR Ment. Health* **2025**, *12*, e78414. [CrossRef] [PubMed]
39. Hasei, J.; Hanzawa, M.; Nagano, A.; Maeda, N.; Yoshida, S.; Endo, M.; Yokoyama, N.; Ochi, M.; Ishida, H.; Katayama, H.; et al. Empowering Pediatric, Adolescent, and Young Adult Patients with Cancer Utilizing Generative AI Chatbots to Reduce Psychological Burden and Enhance Treatment Engagement: A Pilot Study. *Front. Digit. Health* **2025**, *7*, 1543543. [CrossRef] [PubMed]
40. Hu, X.; Yang, X.; Lou, X.; He, T. LittleToDo: Large Language Model Driven Intervention Tool for Adolescent Academic Procrastination with Affective Computing. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2025.
41. Kim, D.H.; Baek, S.; Lee, J.; Lee, T.; Park, S.; You, B.; Hur, J.-W.; Kim, M.; Lee, C.-G. BetterMood: A Human-like AI Counseling Service for Adolescents and Young Adults. *Digit. Health* **2025**, *11*, 20552076251392294. [CrossRef] [PubMed]
42. Ni, Y.; Ding, R.; Chen, Y.; Hou, H.; Ni, S. Focusing on Needs: A Chatbot-Based Emotion Regulation Tool for Adolescents. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oahu, HI, USA, 1–4 October 2023; pp. 2295–2300.
43. Steenstra, I.S.; Murali, P.; Perkins, R.B.; Pierre-Joseph, N.; Paasche-Orlow, M.K.; Bickmore, T.W. Engaging and Entertaining Adolescents in Health Education Using LLM-Generated Fantasy Narrative Games and Virtual Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2024.
44. Ye, X.; Shan, X.; Tu, Y.; Zhang, Y. Examining the Efficacy of Large Language Models for Mitigating Depression and Anxiety Among Chinese Students: A Randomized Controlled Trial. *CIN—Comput. Inform. Nurs.* **2025**, *43*, e01349. [CrossRef] [PubMed]
45. Potts, C.; Ennis, E.; Bond, R.B.; Mulvenna, M.D.; McTear, M.F.; Boyd, K.; Broderick, T.; Malcolm, M.; Kuosmanen, L.; Nieminen, H.; et al. Chatbots to Support Mental Wellbeing of People Living in Rural Areas: Can User Groups Contribute to Co-Design? *J. Technol. Behav. Sci.* **2021**, *6*, 652–665. [CrossRef] [PubMed]
46. Bélisle-Pipon, J.-C. Why We Need to Be Careful with LLMs in Medicine. *Front. Med.* **2024**, *11*, 1495582. [CrossRef] [PubMed]
47. MacMaster, S.; Sinistore, J. Testing the Use of a Large Language Model (LLM) for Performing Data Quality Assessment. *Int. J. Life Cycle Assess.* **2025**, *30*, 2349–2360. [CrossRef]
48. Han, X.; Zhan, Y.; Ni, J.; Yu, B.; Tao, D. Mind the Data: Evaluating Data Quality Sensitivity in Medical LLMs. *Pattern Recognit. Lett.* **2026**, *199*, 68–74. [CrossRef]
49. Ramello, S. Emotional Dependence in Long-Term Use of Large Language Models. *Digit. Soc.* **2025**, *5*, 1. [CrossRef]
50. Namvarpour, M.; Brofsky, B.; Medina, J.Y.; Akter, M.; Razi, A. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2026; pp. 1–19.
51. Lee, S.; Choi, D.; Im, H.; Choi, Y.J.; Lee, M.; Hong, H.; Lee, S. Uncovering Youth's Needs and Concerns About AI Through Art-Based Participatory Design. *Int. J. Hum.–Comput. Interact.* **2026**, 1–25. [CrossRef]
52. Koutra, K.; Kondylakis, H.; Kilintzis, V.; Guerra, B.; Triantafyllidis, A.; Tziraki, C.; Tsiknakis, E. MapStakes-PH: A Methodological Framework for Co-Creation in Preventive Public Health. *Int. J. Qual. Methods* **2026**, *25*, 16094069261435583. [CrossRef]
53. Liu, X.; Rivera, S.C.; Moher, D.; Calvert, M.J.; Denniston, A.K.; Ashrafian, H.; Beam, A.L.; Chan, A.-W.; Collins, G.S.; Deeks, A.D.J.; et al. Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension. *Lancet Digit. Health* **2020**, *2*, e537–e548. [CrossRef] [PubMed]
54. Hoffmann, T.C.; Glasziou, P.P.; Boutron, I.; Milne, R.; Perera, R.; Moher, D.; Altman, D.G.; Barbour, V.; Macdonald, H.; Johnston, M.; et al. Better Reporting of Interventions: Template for Intervention Description and Replication (TIDieR) Checklist and Guide. *BMJ* **2014**, *348*, g1687. [CrossRef] [PubMed]
55. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM Reporting Guideline for Studies Using Large Language Models. *Nat. Med.* **2025**, *31*, 60–69. [CrossRef] [PubMed]
56. Elliott, J.H.; Synnot, A.; Turner, T.; Simmonds, M.; Akl, E.A.; McDonald, S.; Salanti, G.; Meerpohl, J.; MacLehose, H.; Hilton, J.; et al. Living Systematic Review: 1. Introduction-the Why, What, When, and How. *J. Clin. Epidemiol.* **2017**, *91*, 23–30. [CrossRef] [PubMed]

57. Souza, G.d.P.; Melo, G.; Schneider, D. Tailoring Treatment in the Age of AI: A Systematic Review of Large Language Models in Personalized Healthcare. *Informatics* **2025**, *12*, 113. [CrossRef]
58. Abbasian, M.; Azimi, I.; Rahmani, A.M.; Jain, R. Conversational Health Agents: A Personalized Large Language Model-Powered Agent Framework. *Jamia Open* **2025**, *8*, ooaf067. [CrossRef] [PubMed]
59. Neveditsin, N.; Lingras, P.; Mago, V. Clinical Insights: A Comprehensive Review of Language Models in Medicine. *PLoS Digit. Health* **2025**, *4*, e0000800. [CrossRef] [PubMed]
60. Olawade, D.B.; Plabon, S.B.; Ojo, A.; Ogunbona, M.A.; Makanjuola, B.D.; Olasilola, O.R. Human in the Loop Artificial Intelligence in Healthcare: Applications, Outcomes, and Implementation Challenges. *Int. J. Med. Inform.* **2026**, *213*, 106362. [CrossRef] [PubMed]
61. Templin, T.; Fort, S.; Padmanabham, P.; Seshadri, P.; Rimal, R.; Oliva, J.; Hassmiller Lich, K.; Sylvia, S.; Sinnott-Armstrong, N. Framework for Bias Evaluation in Large Language Models in Healthcare Settings. *npj Digit. Med.* **2025**, *8*, 414. [CrossRef] [PubMed]
62. Kim, Y.; Choi, C.-H.; Cho, S.; Sohn, J.; Kim, B.-H. Aligning Large Language Models for Cognitive Behavioral Therapy: A Proof-of-Concept Study. *Front. Psychiatry* **2025**, *16*, 1583739. [CrossRef] [PubMed]
63. Freyer, O.; Wiest, I.C.; Kather, J.N.; Gilbert, S. A Future Role for Health Applications of Large Language Models Depends on Regulators Enforcing Safety Standards. *Lancet Digit. Health* **2024**, *6*, e662–e672. [CrossRef] [PubMed]
64. Bailo, P.; Nittari, G.; Pesel, G.; Basello, E.; Spasari, T.; Ricci, G. Governing Healthcare AI in the Real World: How Fairness, Transparency, and Human Oversight Can Coexist: A Narrative Review. *Sci* **2026**, *8*, 36. [CrossRef]
65. Meskó, B.; Topol, E.J. The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare. *npj Digit. Med.* **2023**, *6*, 120. [CrossRef] [PubMed]
66. Wah, J.N.K. Revolutionizing E-Health: The Transformative Role of AI-Powered Hybrid Chatbots in Healthcare Solutions. *Front. Public Health* **2025**, *13*, 1530799. [CrossRef] [PubMed]
67. Shankar, R.; Lim, A.; Xu, Q. Barriers and Facilitators to the Use of Large Language Model-Based Conversational Agents in Mental Healthcare: A Systematic Review. *Healthcare* **2026**, *14*, 1267. [CrossRef] [PubMed]
68. Hang, C.N.; Tan, C.W.; Chiu, D.M. LLM Teams: Harnessing Large Language Models as Multi-Agent Teammates for Joint Problem-Solving. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale*; Association for Computing Machinery: New York, NY, USA, 2026; pp. 423–427.
69. Chen, G.; Yang, S.; Li, C.; Liu, W.; Luan, J.; Xu, Z. End-to-End Optimization of LLM-Driven Multi-Agent Search Systems via Heterogeneous-Group-Based Reinforcement Learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Liakata, M., Moreira, V.P., Zhang, J., Jurgens, D., Eds.; Association for Computational Linguistics: San Diego, CA, USA, 2026; pp. 30319–30338.