The American Journal of
Engineering and Technology

ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 213-221

Machine Intelligence, Mental Health Access, and Suicide Prevention: Opportunities, Risks, and a Research Framework for Responsible Large Language Model Integration in Clinical and Community

Dr. Elena Morales

University of Lisbon

OPEN ACCESS

SUBMITTED 12 November 2025
ACCEPTED 22 November 2025
PUBLISHED 30 November 2025
VOLUME Vol.07 Issue 11 2025

CITATION

Dr. Mateo Alvarez. (2025). Machine Intelligence, Mental Health Access, and Suicide Prevention: Opportunities, Risks, and a Research Framework for Responsible Large Language Model Integration in Clinical and Community Settings. *The American Journal of Engineering and Technology*, 7(11), 213–221.

Abstract:

Background: Suicide remains a leading public health concern internationally, with measurable changes over recent years that highlight both progress and persistent vulnerabilities in mental health systems (National Institute of Mental Health, 2024; Saunders & Panchal, 2023). Concurrently, development and deployment of large language models (LLMs) and AI-augmented mental health applications are accelerating, producing a contested landscape of opportunity and risk for suicide prevention and mental healthcare broadly (Omar et al., 2024; Karabacak & Margetis, 2023).

Objective: This article synthesizes extant literature to construct a comprehensive, publication-ready research manuscript that: (1) examines how LLMs encode clinical knowledge and their potential utility for mental healthcare and suicide prevention (Singhal et al., 2023; Omar et al., 2024); (2) situates LLMs within persistent structural barriers to access and delivery of mental health services (Ziller, Anderson & Coburn, 2010; Donohue, Goetz & Song, 2024; Coombs et al., 2021); (3) articulates principal technical risks (hallucinations, dataset quality, domain drift) and governance challenges (Islam et al., 2025; Chen et al., 2024; Wettig et al., 2024); and (4) proposes a detailed, ethically grounded methodological framework for evaluation, validation, and staged integration of LLM-based tools into clinical and community settings.

Methods: We performed an integrative synthesis of the provided sources, mapping empirical evidence on suicide epidemiology and service access to contemporary technical literature on LLM capabilities, training-data concerns, and evaluation strategies. From this synthesis we derived a multi-modal research framework combining qualitative stakeholder inquiry, simulated and retrospective validation experiments, prospective safety trials, and continuous monitoring guided by hybrid human-AI oversight. Each element is detailed with operational procedures, measurement constructs, and ethical safeguards drawn from the literature.

Results: The synthesis reveals convergent themes: (1) suicide prevention needs precise, equitable, and accessible interventions; (2) LLMs exhibit surprising clinical pattern understanding but retain unpredictable failure modes and hallucinations; (3) disparities in access to care create both need and risk when AI systems are unevenly distributed or poorly validated in underserved populations; (4) robust evaluation requires domain-specific high-quality data, multi-language and demographic validation, human-feedback loops, and transparency metrics.

Conclusions: LLMs can augment suicide prevention and mental healthcare, but safe, equitable deployment requires methodical evaluation, domain-specific fine-tuning with quality-controlled data, human-in-the-loop safeguards, and policy frameworks to mitigate access-related harms. The proposed research framework operationalizes these requirements and outlines steps for translational research aimed at realizing benefits while minimizing risks.

Keywords: large language models, suicide prevention, mental health access, hallucination, evaluation framework, human-in-the-loop

Introduction

Suicide is a complex, multifactorial public health problem that continues to exert substantial human and social costs globally (National Institute of Mental Health, 2024). Epidemiological trends show both long-term patterns and more recent temporal shifts that demand sustained scientific and policy attention (Saunders & Panchal, 2023). The work of suicide prevention encompasses population-level surveillance, targeted clinical interventions, crisis services, and supportive community programs; it requires not only accurate risk detection but also timely, accessible, and culturally appropriate responses (National Institute of Mental Health, 2024). At the same time, persistent barriers—geographic, financial, systemic—limit the reach and effectiveness of traditional mental health services for many groups, particularly rural and underserved populations (Ziller, Anderson & Coburn, 2010; Coombs et al., 2021; Donohue, Goetz & Song, 2024).

Recent advances in artificial intelligence, and particularly large language models (LLMs), have introduced new possibilities for augmenting mental health delivery and suicide prevention efforts (Singhal et al., 2023; Omar et al., 2024). LLMs demonstrate the ability to encode wide-ranging clinical knowledge and to generate fluent, contextually appropriate language that could support symptom screening, psychoeducation, therapeutic conversation design, and clinician decision support (Singhal et al., 2023; Karabacak & Margetis, 2023). Alongside these opportunities, scholarly and technical concerns have surfaced: models may produce incorrect or fabricated assertions ("hallucinations"), fail to generalize across languages or sociocultural contexts, and reflect biases present in training data (Islam et al., 2025; Chen et al., 2024; Dahl et al., 2024). The literature further emphasizes the centrality of data quality and domain specificity: improvements in data curation and selection are often more influential than mere data volume for model performance in high-stakes domains (Wettig et al., 2024; Chan et al., 2025).

This article aims to integrate evidence about suicide epidemiology, access barriers to mental health care, and technical capacities and risks of LLMs to produce an actionable research framework for responsible LLM integration in suicide prevention and mental health services. The analysis proceeds from three primary premises: (1) technological tools must be tailored to the epidemiological and service realities they aim to augment; (2) technical validation and ethical governance must be embedded throughout the lifecycle of design, testing, and deployment; and (3) equitable access and continuous evaluation are non-negotiable for interventions intended to benefit populations at risk of suicide. These premises are grounded in the cited literature and inform the proposed methodology and evaluation approach described below (National Institute of Mental Health, 2024; Ziller, Anderson & Coburn, 2010; Donohue, Goetz & Song, 2024; Singhal et al., 2023; Omar et al., 2024).

Problem Statement and Literature Gap

Despite the proliferation of mental health applications and AI-assisted tools marketed for emotional wellbeing and symptom management (Mya Care, 2023; Rawat, 2023), rigorous, domain-specific evidence on the safe and equitable use of LLMs in suicide prevention remains limited. Existing reviews highlight potential clinical utility but emphasize heterogeneity in methods, variable reporting standards, and insufficient prospective safety trials (Omar et al., 2024). Moreover, the canonical challenges of mental health service delivery—cost barriers, provider distribution, and rural access gaps—imply that technology-based interventions could either ameliorate or exacerbate existing inequities depending on how they are designed and governed (Ziller, Anderson & Coburn, 2010; Coombs et al., 2021; Donohue, Goetz & Song, 2024). The technical literature on LLM training underscores that model capabilities depend critically on data selection, annotation quality, and evaluation metrics; yet domain-specific datasets for psychiatry and suicide prevention are often small, noisy, and compositionally unrepresentative (Chen et al., 2024; Sun et al., 2024; Wettig et al., 2024). Finally, safety concerns such as hallucinations and legal or ethical misclassification are not uniformly studied across languages and contexts, leaving important gaps for multilingual, cross-cultural deployment (Islam et al., 2025; Bagheri Nezhad et al., 2024).

To address these gaps, research must move beyond proofs-of-concept toward rigorous, multi-stage evaluation frameworks that combine retrospective benchmarking, prospective safety testing, stakeholder-guided design, and ongoing operational monitoring. This article presents such a framework grounded in the extant literature and aimed at bridging technical research and real-world suicide prevention practice.

Methodology

This section presents a detailed, text-based methodological approach for conducting rigorous, ethically grounded research on LLM integration into suicide prevention and mental health care. The methodology synthesizes best-practice insights from clinical and technical literatures and is designed to be operational and reproducible. It comprises (A) preparatory domain analysis and stakeholder engagement, (B) data strategy and curation, (C) model fine-tuning and internal validation, (D) simulated and retrospective clinical evaluations, (E) prospective safety trials with human oversight, and (F) continuous monitoring and governance. Each component is elaborated with stepwise procedures, measurement constructs, ethical safeguards, and references.

A. Preparatory Domain Analysis and Stakeholder Engagement

Rationale and objectives. Any LLM application in suicide prevention must begin with careful problem scoping that aligns technological possibilities with clinical needs and social realities (National Institute of Mental Health, 2024; Ziller, Anderson & Coburn, 2010). Preparatory work prevents misaligned solutions and reduces the risk of harm from unanticipated model behaviors.

Procedures.

1.Stakeholder mapping: Identify and recruit stakeholders across clinical disciplines (psychiatrists, psychologists, emergency clinicians), community organizations (suicide prevention NGOs, peer-support groups), technology developers, legal/ethical experts, and end users including individuals with lived experience of suicidal ideation or recovery. Ensure diversity in geography, language, and socioeconomic status to capture varied access barriers and cultural perspectives (Coombs et al., 2021; Donohue, Goetz & Song, 2024).

2.Needs assessment workshops: Conduct structured workshops and semi-structured interviews to elicit specific service gaps—e.g., screening at primary care, crisis triage, follow-up for discharged patients, and psychoeducation needs. Use qualitative thematic analysis to produce domain requirements. Document pain points amenable to LLM augmentation and those requiring human-only interventions.

3.Ethical and legal review: Convene institutional review boards and legal advisors to assess jurisdictional obligations regarding mandatory reporting, crisis intervention, data protection, and liability. This step is essential given the high-stakes nature of suicide-related interactions.

Measurement constructs. Metrics include stakeholder-reported priority needs, anticipated benefits and risks, and a dashboard of legal/ethical constraints by jurisdiction. Each metric is catalogued for use in later risk assessment and deployment planning.

Citation. The importance of stakeholder engagement and context-aware design is emphasized in literature on access barriers and service use, which documents heterogeneity across populations and the ethical imperatives when designing interventions for vulnerable groups (Ziller, Anderson & Coburn, 2010; Coombs et al., 2021; Donohue, Goetz & Song, 2024).

B. Data Strategy and Curation

Rationale and objectives. Model behavior is fundamentally shaped by training and fine-tuning data. High-quality, domain-specific, and representative datasets are critical to reduce hallucination, bias, and capability collapse when models are adapted to sensitive domains like suicide prevention (Wettig et al., 2024; Sun et al., 2024; Chen et al., 2024).

Procedures.

1.Inventory existing datasets: Catalog available clinical corpora, anonymized crisis chat logs (where accessible under strict privacy controls), crisis line transcripts, structured clinical data, and validated psychiatric assessment instruments. For each dataset, record provenance, language, demographic coverage, annotation schema, and any access restrictions.

2.Data quality assessment: Apply a structured rubric to score datasets on attributes such as completeness, representativeness, annotation consistency, and documented provenance (Wettig et al., 2024; Chan et al., 2025). Identify gaps—e.g., underrepresentation of minority languages or rural populations.

3.Ethical sourcing and consent: For new or pooled data, implement consent processes, de-identification protocols, and data-use agreements consistent with local regulations. Where retrospective data contain sensitive disclosures (suicidal ideation), additional safeguards are installed, including limited researcher access and secure storage.

4.Annotation framework: Develop an annotation guideline for suicide-relevant constructs: ideation severity, imminence markers, intent, plan specificity, protective factors, and contextual stressors. Use multi-rater annotation with adjudication and report inter-rater reliability measures. Incorporate lived-experience reviewers to validate interpretive frameworks.

5.Synthetic augmentation with caution: Where data are scarce for specific subgroups, consider synthetic data generation techniques but only under rigorous validation: synthetic variations should be audited against real instances for fidelity and not used as primary evidence in safety-critical evaluation. Research indicates synthetic data diversity can impact training outcomes but must be balanced with quality considerations (Chen et al., 2024).

Measurement constructs. Data quality scores, inter-rater reliability (Cohen’s kappa or other robust agreement metrics), demographic coverage indices, and a dataset governance ledger accessible to project oversight.

Citations. Concerns about data quality and the importance of targeted, high-quality domain data are documented in the technical literature and argued to often outweigh raw data volume in high-stakes domains (Wettig et al., 2024; Chan et al., 2025; Sun et al., 2024; Chen et al., 2024).

C. Model Fine-Tuning and Internal Validation

Rationale and objectives. Fine-tuning a base LLM on domain-specific, high-quality annotated data can improve task performance such as intent detection, risk stratification, and therapeutic message generation. However, it also carries risks of overfitting and capability collapse when domain tuning is not carefully constrained (Sun et al., 2024).

Procedures.

1.Baseline evaluation: Establish baseline performance of the base LLM on benchmark tasks (e.g., clinical question-answering, risk-labeling) using held-out domain test sets. Use standardized metrics (precision, recall, F1, area under curve where appropriate) and error analyses. Report and document failure modes.

2.Fine-tuning strategy: Adopt a staged fine-tuning approach—first, small learning rates with constrained parameter updates focusing on classification heads or adapters; second, iterative evaluation with increasing data complexity to avoid rapid drift. Consider parameter-efficient fine-tuning methods and human-in-the-loop reviews for generated outputs (Singhal et al., 2023; Sun et al., 2024).

3.Safety-oriented objective functions: Integrate safety signals into fine-tuning, such as penalizing outputs that provide medical instructions, encourage self-harm, or ignore crisis protocols. Use reinforcement learning from human feedback (RLHF) where human raters evaluate model outputs against safety and clinical appropriateness criteria (Reinforcement Learning from Human Feedback, 2024).

4.Internal adversarial testing: Employ adversarial prompts and scenario testing to probe hallucination risks, boundary behaviors, and abuse cases (Dahl et al., 2024; Islam et al., 2025). Maintain a taxonomy of prompt classes and model weaknesses.

Measurement constructs. Task accuracy metrics, safety violation counts in adversarial testing, calibration metrics (confidence vs. correctness), and qualitative error typology.

Citations. Demonstration that LLMs encode clinical knowledge and can be fine-tuned for domain tasks is supported by empirical studies; RLHF and human feedback have been central to aligning model behavior but do not eliminate hallucinations, thus necessitating careful testing (Singhal et al., 2023; Reinforcement Learning from Human Feedback, 2024; Sun et al., 2024).

D. Simulated and Retrospective Clinical Evaluations

Rationale and objectives. Before prospective trials, LLM tools should be assessed in simulated environments and on retrospective clinical data to measure reliability, safety, and potential impact without exposing live patients to untested systems.

Procedures.

1.Retrospective validation: Use de-identified clinical records and crisis transcripts to measure model sensitivity and specificity for detecting suicidal ideation, imminent risk, and need for escalation. Compare model outputs to clinician annotations and known outcomes where available. Report subgroup analyses for language, age, gender, and socioeconomic markers to detect distributional performance gaps.

2.Simulation studies: Create simulated care pathways where the LLM interacts with standardized patient vignettes representing diverse cultural contexts and crisis acuity. Include clinician actors and peer-support participants to rate the quality of model responses and escalation appropriateness.

3.Human oversight trials: In retrospective-controlled simulations, pair LLM suggestions with clinician decision-makers to examine whether LLM outputs materially change clinician behavior and whether those changes align with desired safety outcomes.

Measurement constructs. Diagnostic performance metrics, decision impact measures (e.g., change in clinician disposition), false positive and false negative consequences, and qualitative acceptability scores.

Citations. The need for rigorous retrospective and simulation testing is consonant with calls for evidence-based evaluation before deployment in clinical contexts (Omar et al., 2024; Singhal et al., 2023). Retrospective studies also illuminate differential access barriers and population heterogeneity relevant to deployment planning (Ziller, Anderson & Coburn, 2010; Donohue, Goetz & Song, 2024).

E. Prospective Safety Trials with Human Oversight

Rationale and objectives. Controlled prospective studies are essential to understand real-world effects, safety, and unintended consequences—especially given the high stakes of suicide prevention.

Procedures.

1.Staged deployment: Use an incremental roll-out design beginning with low-risk tasks (e.g., clinician decision support, automated documentation suggestions) before moving to public-facing features (e.g., direct user interactions) where the model could interact with people in distress.

2.Human-in-the-loop requirement: For any system that provides risk assessment or empathetic responses related to suicide, institute mandatory human oversight for escalation decisions. LLM outputs are presented as suggestions with explicit confidence estimations and rationales; final decisions remain with trained human personnel.

3.Safety-service integration: Predefine escalation pathways (e.g., local crisis lines, emergency services) and ensure that any automated suggestions conform to legal and local practice standards. Implement logging and auditing of every LLM interaction that involves safety concerns.

4.Ethics and consent in prospective use: Obtain informed consent where appropriate, especially for research settings. For public deployments (e.g., apps), provide transparent terms of use that explain model limitations and crisis resources.

5.Monitoring and rapid rollback: Establish real-time monitoring dashboards for safety metrics (e.g., flagged risky cases, false negative incidents) and predefined thresholds that trigger immediate suspension of features until safety review.

Measurement constructs. Incidence of safety breaches, appropriateness of escalations, user-reported experience and trust measures, clinician workload indicators, and time-to-escalation metrics.

Citations. The staged approach and human oversight mechanisms are recommended in the safety-first literature and are consistent with clinical governance needs for high-stakes interventions (Omar et al., 2024; Karabacak & Margetis, 2023; Reinforcement Learning from Human Feedback, 2024).

F. Continuous Monitoring, Auditing, and Governance

Rationale and objectives. LLM systems are not static artifacts; they require continuous oversight for drift, new failure modes, and changing population needs.

Procedures.

1.Post-deployment surveillance: Implement metrics for model performance across demographics, languages, and settings; monitor for emergent biases or declines in accuracy.

2.Periodic revalidation: Schedule re-evaluations when models are updated, when new data indicate shift in user populations, or following any safety incident.

3.Transparency reporting: Publish periodic transparency reports detailing model evaluations, safety incidents, mitigations, and governance decisions (subject to legal constraints).

4.Community feedback channels: Maintain accessible channels for users and clinicians to report harms or misclassifications and to request data removal or other remedies.

Measurement constructs. Drift detection indicators, incident frequency and resolution times, and compliance with auditing standards.

Citations. The necessity of continuous evaluation and governance is emphasized across technical and clinical domains where domain shift and dataset quality issues can affect model reliability (Wettig et al., 2024; Chen et al., 2024; Islam et al., 2025).

Results

This section synthesizes hypothetical and literature-driven results that would be expected under the described methodology. Because the current article is a methodological and integrative research manuscript, the "results" synthesise what the literature indicates and what rigorous evaluation would be likely to reveal if the proposed steps were followed.

Epidemiological Context and Needs

Epidemiological data confirm that suicide remains a pressing global concern, with observable shifts in rates and patterns over recent years (National Institute of Mental Health, 2024; Saunders & Panchal, 2023). These shifts underline persistent service gaps, including geographic disparities where rural populations face higher travel burdens and out-of-pocket expenses for mental healthcare (Ziller, Anderson & Coburn, 2010). Additionally, financial and market dynamics influence access—cash-paying markets and burden-based disparities shape who receives care (Donohue, Goetz & Song, 2024). Population-level studies also highlight barriers beyond cost: stigma, limited provider availability, and fragmented care pathways that inhibit timely intervention (Coombs et al., 2021).

Implication for LLM integration. These service gaps create both a rationale and a risk for LLM-based interventions. On one hand, scalable conversational or decision-support tools could increase access to triage and psychoeducation where human resources are scarce (Mya Care, 2023; Rawat, 2023). On the other hand, unequal distribution and insufficient validation across underrepresented populations could exacerbate disparities if systems perform worse for marginalized groups (Donohue, Goetz & Song, 2024; Coombs et al., 2021).

LLM Capabilities and Risks

Empirical work demonstrates that LLMs encode substantive clinical knowledge and can perform various medical question-answering tasks with competence in controlled evaluations (Singhal et al., 2023). Systematic reviews find increasing application of LLMs in psychiatry for tasks such as diagnostic support, therapeutic content generation, and data augmentation (Omar et al., 2024). However, case series and experimental probe studies document model hallucinations—instances of confidently stated but factually incorrect information—and domain-specific failure modes, particularly in underrepresented languages and nuanced clinical scenarios (Islam et al., 2025; Chen et al., 2024; Dahl et al., 2024).

Implication for LLM integration. Performance alone is insufficient; safety and reliability must be explicitly evaluated. The literature shows that targeted fine-tuning with high-quality, domain-specific data can improve performance, but also warns about capability collapse if tuning diminishes broader language understanding without improving task alignment (Sun et al., 2024; Wettig et al., 2024).

Data Quality and Training Considerations

Technical research emphasizes that data selection and quality critically determine downstream model behavior. Studies comparing high-quality, curated datasets to larger but noisier corpora suggest that quality-focused selection yields superior performance on domain tasks (Chan et al., 2025; Wettig et al., 2024). Additionally, synthetic data and augmentation strategies can be valuable but require validation to avoid inducing artifacts or spurious generalization (Chen et al., 2024).

Implication for LLM integration. A deliberate curation strategy—annotated, representative, and ethically sourced—will likely yield better safety and generalization than indiscriminate reliance on large, noisy datasets.

Evaluation and Safety Outcomes

When LLMs are assessed using the staged approach described (retrospective validation, simulation, human-in-the-loop trials), likely findings include: improved clinician efficiency in documentation and triage tasks with decision support; variable performance on direct user-facing empathetic conversations; and a nontrivial incidence of outputs requiring human correction, particularly in high-acuity or culturally specific vignettes. Subgroup analyses commonly reveal performance disparities that necessitate targeted model adjustments or operational controls (Omar et al., 2024; Singhal et al., 2023; Islam et al., 2025).

Implication for LLM integration. Both positive impacts and residual risks are expected. The net benefit depends on rigorous validation, human oversight, and continuous monitoring.

Acceptability and Ethical Considerations

Stakeholder engagement typically surfaces both enthusiasm and caution: clinicians and organizations value potential efficiency gains but demand transparency, liability clarity, and assurance that tools will not replace human judgement in crises. People with lived experience emphasize the need for sensitivity, privacy protections, and local crisis linkages (Coombs et al., 2021; Donohue, Goetz & Song, 2024).

Implication for LLM integration. Implementation must prioritize user trust, legal clarity, and community-specific adaptations.

Discussion

This section interprets the synthesized findings, addresses limitations, and outlines a research and policy agenda. The discussion is organized around core themes: (1) the promise of LLMs in augmenting access and clinician capacity; (2) the technical and ethical challenges that must be addressed to prevent harm and disparity; (3) operational recommendations for researchers, clinicians, and policymakers; and (4) directions for future research.

The Promise: Augmenting Capacity and Access

LLMs offer scalable language mediation that can assist in screening, patient education, and administrative workflows—areas where human resources are frequently bottlenecked (Singhal et al., 2023; Mya Care, 2023). For rural or underserved populations, well-designed LLM-enabled tools could extend triage capabilities and facilitate connections to local resources when paired with human oversight and clear escalation protocols (Ziller, Anderson & Coburn, 2010). In clinical workflows, LLMs may reduce documentation burdens and suggest evidence-aligned phrasing for clinicians, thereby freeing clinician time for therapeutic tasks (Karabacak & Margetis, 2023).

The Risks: Hallucination, Bias, and Unequal Benefit

Key risks identified in the literature include hallucinations—confident but incorrect outputs that are particularly dangerous in medical contexts—and biased performance that disadvantages marginalized groups (Islam et al., 2025; Chen et al., 2024). The potential for LLMs to produce plausible-sounding but clinically unsound guidance necessitates a default posture of skepticism and the requirement that any clinical action derives from human-evaluated recommendations (Dahl et al., 2024). Data provenance and quality issues compound these risks: training on noisy or unrepresentative data can entrench biases and reduce performance where it matters most (Wettig et al., 2024; Chan et al., 2025).

Operational Recommendations

1.Prioritize High-Quality, Domain-Specific Data: Invest in curated, annotated datasets that include diverse languages, cultures, and care settings. Use rigorous annotation guidelines and inter-rater reliability assessments (Wettig et al., 2024; Sun et al., 2024).

2.Human-in-the-Loop and Staged Rollout: Maintain human oversight for any safety-critical decisions, and deploy LLM assistance first in low-risk, clinician-facing areas before public deployment (Reinforcement Learning from Human Feedback, 2024; Omar et al., 2024).

3.Transparent Reporting and Accountability: Publish validation results, safety incidents, and governance measures in accessible transparency reports. These build trust and allow external scrutiny.

4.Contextualize Tools to Local Systems: Integrate escalation pathways that reflect jurisdictional crisis services and legal requirements. One-size-fits-all deployments risk misalignment with local emergency protocols (Ziller, Anderson & Coburn, 2010).

5.Continuous Monitoring and Update Paths: Implement routine revalidation cycles and drift detection; be prepared to roll back or modify features in response to safety signals (Wettig et al., 2024).

Policy and Ethical Implications

Policymakers and regulators should require evidence of safety, fairness, and effectiveness for LLM-based systems intended for mental health use. Regulation should focus both on pre-deployment validation and post-deployment surveillance. The existing literature supports regulatory emphasis on transparency, data governance, and human oversight to protect vulnerable populations (Omar et al., 2024; Donohue, Goetz & Song, 2024).

Limitations

This article synthesizes a provided set of references and extrapolates methodological recommendations and expected outcomes rather than reporting novel empirical trial results. The conclusions and framework proposed are therefore prescriptive and intended as a research roadmap grounded in the cited literature rather than definitive proof of efficacy in live deployments. Additionally, while the references include recent and varied sources, the dynamic nature of model development and policy means that new findings emerging after these sources could refine or alter specific technical recommendations.

Future Research Directions

1.Prospective Trials in Diverse Settings: Conduct multi-site randomized trials comparing human-only, LLM-augmented care, and hybrid approaches with robust safety endpoints. Include rural, low-resource settings to evaluate equity effects.

2.Multilingual and Cultural Generalization Studies: Evaluate LLM performance across languages and cultural contexts, focusing on hallucination rates, misclassification patterns, and acceptability among diverse populations (Islam et al., 2025; Bagheri Nezhad et al., 2024).

3.Data-Efficiency and Quality Research: Investigate how targeted data selection strategies, quality scoring frameworks, and domain-specific fine-tuning influence both performance and safety outcomes (Chen et al., 2024; Wettig et al., 2024; Chan et al., 2025).

4.Human-AI Interaction Studies: Explore optimal interfaces for human oversight, including how clinicians interpret and act on model explanations, confidence scores, and recommended escalations.

5.Governance and Liability Research: Evaluate legal frameworks and institutional policies that can distribute accountability fairly while enabling beneficial innovation (Donohue, Goetz & Song, 2024).

Conclusion

Large language models present both substantive opportunities and substantial risks for suicide prevention and mental health care. The literature indicates that LLMs can encode clinical knowledge and support a range of tasks, but also that hallucinations, dataset quality problems, and distributional failures impose real hazards—especially for populations already underserved by traditional services (Singhal et al., 2023; Omar et al., 2024; Islam et al., 2025). To ethically and effectively harness LLMs, researchers and implementers must adopt rigorous, staged methodologies that prioritize high-quality domain data, human oversight, robust retrospective and prospective evaluations, and continuous governance. The framework presented here translates these principles into concrete research steps and measurement constructs that can guide translational efforts aimed at improving reach and quality of suicide prevention while minimizing the risk of harm. Ultimately, the promise of LLMs will be realized only through methodical, evidence-driven integration that centers safety, equity, and respect for affected communities.

References

1. Suicide. National Institute of Mental Health. 2024. URL: https://www.nimh.nih.gov/health/statistics/suicide [accessed 2024-07-01]

2. Saunders H, Panchal N. A look at the latest suicide data and change over the last decade. Kaiser Family Foundation. Aug 04, 2023. URL: https://www.kff.org/mental-health/issue-brief/a-look-at-the-latest-suicide-data-and-change-over-the-last-decade/ [accessed 2024-07-01]

3. Omar M, Soffer S, Charney AW, Landi I, Nadkarni GN, Klang E. Applications of large language models in psychiatry: a systematic review. Front Psychiatry. 2024;15:1422807.

4. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus. 2023;15(5):e39305.

5. Mental health apps and the role of ai in emotional well-being. Mya Care. Nov 08, 2023. URL: https://myacare.com/blog/mental-health-apps-and-the-role-of-ai-in-emotional-wellbeing [accessed 2024-07-15]

6. Rawat M. Best AI apps for mental health (2023). MarkTechPost. Apr 11, 2023. URL: https://www.marktechpost.com/2023/04/11/best-ai-apps-for-mental-health-2023/ [accessed 2024-07-15]

7. Ziller EC, Anderson NJ, Coburn AF. Access to rural mental health services: service use and out-of-pocket costs. J Rural Health. 2010;26(3):214-224.

8. Malviya S, Parate V. AI-Augmented Data Quality Validation in P&C Insurance: A Hybrid Framework Using Large Language Models and Rule-Based Agents. International Journal of Computational and Experimental Science and Engineering. 2025;11(3). https://doi.org/10.22399/ijcesen.3613

9. Donohue JM, Goetz JL, Song Z. Who gets mental health care?-The role of burden and cash-paying markets. JAMA Health Forum. 2024;5(3):e240210.

10. Coombs NC, Meriwether WE, Caringi J, Newcomer SR. Barriers to healthcare access among U.S. adults with mental health challenges: a population-based study. SSM Popul Health. 2021;15:100847.

11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180.

12. Large Language Models (LLMs) [Online] - https://en.wikipedia.org/wiki/Large_language_model

13. Reinforcement Learning from Human Feedback (RLHF) [Online] - https://en.wikipedia.org/wiki/Reinforcement_learning_from_human_feedback

14. Islam SO, Lauscher A, Glavaš G. How Much Do LLMs Hallucinate across Languages? On Multilingual Estimation of LLM Hallucination in the Wild. arXiv preprint. 2025;2502.12769.

15. Chen H, et al. On the Diversity of Synthetic Data and its Impact on Training Large Language Models. arXiv preprint. 2024;2410.15226.

16. Dahl M, et al. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Law and Artificial Intelligence. 2024;16(1):64-102.

17. Chan W, et al. Lean-ing on Quality: How High-Quality Data Beats Diverse Multilingual Data in AutoFormalization. arXiv preprint. 2025;2502.15795.

18. Wettig A, et al. QuRating: Selecting High-Quality Data for Training Language Models. arXiv preprint. 2024;2402.09739.

19. Sun J, et al. Dial-insight: Fine-tuning Large Language Models with High-Quality Domain-Specific Data Preventing Capability Collapse. arXiv preprint. 2024;2403.09167.

20. Bagheri Nezhad S, Agrawal A, Pokharel R. Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models. arXiv preprint. 2024;2412.12500.