**Project Final Report**

Olive: Data-Driven, AI-Powered Mental Health Diagnosis

James Kemerait

CS 4365: Intro. Enterprise Computing

Prof. Calton Pu

21st July, 2025

**Table of Contents**

1. **Introduction…………………………………………………………2**
2. **Motivation……………………………………………………………4**
3. **Related Work………………………………………………………..7**
4. **Methodology…………………………………………………….…10**

Context…………………………………………………….……………………..10

User Interaction……………………………………………………………….10

Experimental Design....…………………………………………………….10

1. **Results…………………………………………………………3**
2. **Discussion…………………………………………………………3**
3. **Future Work…………………………………………………………3**
4. **References…………………………………………………………3**

# **1. Introduction**

We live in a time of globally unprecedented mental health challenges. The World Health Organization (2025) reports that suicide is now the third leading cause of death among 15-29 year-olds worldwide and that “more than 720,000 people die due to suicide every year”. The Jed Foundation (2024) reports that, in the United States alone, approximately 33% of adults aged 18-25 experienced some mental health issue in the past year, which is a significant increase from just 22.1% in 2016. This alarming trend is particularly pronounced among adolescents: a staggering “40% of high school students reported persistent feelings of sadness or hopelessness” (The Jed Foundation, 2024.)

These difficulties are only exacerbated further by trends of loneliness and isolation blanketing the globe. Approximately 20% of U.S. adults experience daily loneliness, a trend motivating “the Surgeon General [to declare] a loneliness epidemic in the U.S.” in 2023 (Gallup News, 2025).  while 33% of individuals globally report feelings of loneliness (Yellow Bus ABA, 2024). Harvard Researchers found that the COVID-19 pandemic accelerated these trends, particularly among young people, with “61% of young adults aged 18–25 reporting high levels of loneliness,” and “young adults were four to five times more likely to report loneliness compared with older adults” (Harvard Gazette, 2021). The gravity of these findings is not to be understated: Holt-Lunstad et al. (2015) found that “actual and perceived social isolation” both correspond to a “26% increased likelihood of [early] mortality.”

Despite the sheer scale and magnitude of the global mental health crisis, “85% of individuals with mental health disorders in low- and middle-income countries receive no treatment;” even in developed nations, “the supply-and-demand gap exceeds one-third of need” (Project HOPE, 2024). This discrepancy in available versus needed care, brought about due to geographic barriers, social stigma, and workforce shortages, emphasizes an “urgent imperative for scalable, accessible, and culturally adaptable interventions” (U.S. Surgeon General, 2023).

# **2. Motivation** The advent and widespread adoption of AI technologies, particularly Large Language Models (LLMs), has been unignorably disruptive, seemingly in all facets of our lives. The “exponential growth in large language model (LLM) deployment has fundamentally transformed how humans interact with artificial intelligence,” and this momentum is not slowing down: analysts predict an “80% compound annual growth rate for LLM products through 2030,” with “over 750 million LLM-powered applications projected by 2025” (Springs Analytics, 2025; Hostinger, 2025). This growth stems from the “unprecedented human-like conversational fluency” boasted by LLMs, which “has demonstrated potential for fostering therapeutic alliance levels comparable to outpatient psychotherapy in early clinical trials” (Jacobson, Heinz, & Winters, 2025).

Given the imprecise state of diagnostic practices in mental health, the widespread adoption of emerging AI technologies introduces a unique opportunity to leverage AI tools to provide a data-driven, statistically precise perspective to the practice of diagnosing mental illnesses. Bradford A, Meyer AND, Khan S, et al. (2024) explore diagnostic error in mental health settings, asserting that it “is well understood to be a problem.” While “various avenues for future research and development” have been suggested regarding “missed, wrong, delayed and disparate diagnosis of common mental disorders,” “a lack of clear consensus on how to conceptualise, define and measure errors in diagnosis will pose a barrier to advancement.” They then point to numerous statistics concerning the accuracy of diagnosis and findings on intersectional dynamics concerning age, gender, race, and other factors that may affect accuracy in diagnosing anxiety disorders, ADHD, ASD, mood disorders, and schizophrenia.

Martin-Key N. et al. (2021) explores digital assessment tools for mental health screening and diagnosis, noting that, in spite of their rapid growth in number, “little is known about their diagnostic accuracy.” The authors cite challenges patients may face “disclosing mental health difficulties in person” due to social stigma and geographical barriers. They also note the “acceptability and efficacy of digital platforms for improving the reach, quality, and impact of mental health care,” and, coupled with “strong interest” from patients in engaging with digital assessment tools, digital mental health assessment options appear to be growing in feasibility. Because “little attention and effort have been put into establishing their diagnostic accuracy,” it is critical that new digital tools involve an earnest, data-driven approach in their creation.

Smith et al. (2023) offer a framework followed closely in this project by which digital mental health innovations maintain efficacy and feasibility of introduction in the broader clinical ecosystem. It notes several key outcomes: “traditional diagnostic systems” may best suit digital approaches, digital approaches “require organisational change” rather than total replacement/automation of traditional methods, “unique ethical issues” warrant the design of “appropriate studies to measure the effectiveness of implementation” of digital solutions, “[a]ccessibility and codesign” of solutions should be considered with regards to their longevity, and “[s]tandardised guidelines” will ensure effective clinical implementation.

# **3. Related Work**

Digital tools for mental health diagnosis and treatment show promise concerning viable deployment, with Linardon et al. (2023) finding that “digital mental health tools are moderately to highly effective in reducing depression and anxiety symptoms.” Post-pandemic, one trial found digital mental health services to be “effective in improving subjective well-being and clinical improvement in depressive symptoms” (Prescott et al., 2022). Furthermore, Iyortsuun et al. (2023) demonstrate the potential of Machine Learning (ML) and Deep Learning (DL) methods to aid in diagnosing mental illnesses. The authors cite “the wide possibilities of using DL methods for mental health diagnosis with good results”, as one team developed “end-to-end CNN architecture show[ing] excellent precision (99.76%), Recall (99.74%), F1-Score (99.75%), accuracy (99.72%) and AUC (99.75%) in a three-way classification task.”

Systematic reviews now report LLM applications “ranging from suicide-risk triage to psychodynamic case formulation” (Omar, Soffer, & Charney, 2024). However, Iyortsuun et al. (2023) caution that “the wide possibilities...are tempered by significant limitations,” noting that previous reviews “used only four search databases” and “focused on seven mental health diseases” (p. 285). Another review (Le Glaz A et al., 2021) of 58 articles aims to characterize the use of ML/NLP (Natural Language Processing) techniques for mental health and to consider their potential for use in clinical practice. While ML/NLP models “may be considered a new paradigm in medical research,” “these processes tend to confirm clinical hypotheses” rather than introduce new ideas and results. The authors did identify “unexplored” areas from which ML/NLP techniques may offer uniquely beneficial information, “ie, patients’ daily habits that are usually inaccessible to care providers.” Despite this perspective, the article stresses that these techniques are “tools to support clinical practice,” following the theme that existing practices should be augmented, not automated.

Iyortsuun et al. (2023) cite challenges using AI technologies to diagnose mental illness, including that “[o]nly four search databases (Google Scholar, PubMed, Scopus, and Web of Science) were used to collect data” and that “the focus was limited to seven mental health diseases” in their analysis. There are persisting concerns beyond this, too: “LLMs often exhibit demographic biases, producing less empathetic responses when interacting with underrepresented groups” (Raza et al., 2024). Additionally, “LLMs trained on large datasets...may inadvertently absorb and amplify existing societal biases” (Raza et al., 2024), and studies document “racial bias in AI-mediated psychiatric diagnosis and treatment” (Nature Digital Medicine, 2025). Additionally, “ChatGPT 3.5 has demonstrated unsafe triage rates, misclassify urgent mental health crises, and potentially delay critical care” (Mazumdar et al., 2024). These findings “underscore the need for rigorous evidence on diagnostic accuracy, safety guardrails, and circumstances where hybrid ‘LLM + clinician-in-the-loop’ workflows outperform either component alone” (Kim & Wang, 2025).

Generative chatbots have evolved from rule-based scripts at their simplest (e.g., Woebot) into dynamic modern agents. While Therabot, Dartmouth’s generative AI-powered therapy chatbot, “achieved a 51% reduction in depressive symptoms in a randomized controlled trial” (Jacobson, Heinz, & Winters, 2025), Stanford researchers conversely found “persistent diagnostic stigma across multiple LLM families” (Moore et al., 2025). Recent work emphasizes the “digital therapeutic alliance” in which “AI-human collaborative systems can achieve empathy ratings higher than human-only interactions,” with “conversations co-authored by AI...rated as more empathic and supportive” (Malouin-Lachance et al., 2025). However, these advances remain complex, “challenging personalization, ethical concerns, and long-term impact” in providing safe and precise mental health solutions (Malouin-Lachance et al., 2025).

This project addresses the gaps in existing solutions by building a retrieval-augmented LLM wrapper that leverages the power of AI in executing established clinical practice: The Olive chatbot dynamically assesses patient self-reports, selects validated screening inventories for mental health diagnosis, and generates structured outputs for review by a human clinical provider.

# **4. Methodology**

The first major choice in designing the Olive chatbot application was for it to run locally. Olive is programmed in Python and interfaces with LLMs via Ollama, the models of which can easily be chosen and changed with the MODEL\_NAME string on line 4 of Olive/llm/local\_llm.py. The app also leverages a Retrieval-Augmented Generation (RAG) pipeline, allowing it to base each user interaction on the most relevant diagnostic criteria and clinical evidence data available. Data-driven prompting improves the safety and accuracy in Olive’s performance, reducing LLM hallucinations and aligning its recommendations with provided screening tools.

**Context (Datasets + DSM-5)**

The RAG index relies on context from four sources:

* DAIC-WOZ Database Transcriptions, a collection of transcripts from 189 conversations between a human patient and a virtual assistant chosen to contextualize depression, anxiety, and PTSD,
* Kaggle OCD Patient Dataset: Demographics & Clinical Data, a comprehensive collection of information pertaining to 1500 individuals diagnosed with Obsessive-Compulsive Disorder (OCD),
* PTSD Repository Study Characteristics Dataset, including basic information about PTSD study design, clinical setting, whether the study looked at PTSD symptom clusters, whether there were any subgroup analyses, type of diagnostic measure used (e.g., diagnostic interview, self-report), and provider credentials, and
* Excerpts from DSM-5, a work detailing authoritative diagnostic criteria for mental health disorders consistent with clinical standards.

The excerpts from DSM-5 describe 10 disorders/groups, which were chosen to provide disorders with a range of complexities and rarities to assess LLM performance. They are listed below along with the associated inventories for measuring disorder presence/severity provided to be administered by Olive:

All patients take the PHQ-4, a 4-question assessment signaling the severity of their symptoms.

1. Depression
   1. PHQ-9, Depression
   2. BDI-II, Severity of Depression
   3. QIDS-SR16, Depression
2. Anxiety & Panic Disorders
   1. GAD-7, Generalized Anxiety Disorder Screener
   2. PDSS-SR, Panic Disorder
   3. BAI, General anxiety
   4. SPIN, Social anxiety
3. PTSD & Trauma
   1. PCL-5, PTSD symptoms
   2. IES-R, Trauma Response
   3. ACE-q, Childhood trauma
4. Substance Use
   1. AUDIT, alcohol use
   2. DAST-10, drug abuse
   3. ASSIST, alcohol, smoking and more
5. ADHD
   1. ASRS, ADHD
6. OCD
   1. OCI-R, OCD
7. ASD
   1. AQ-10, Autism
8. Bipolar
   1. MDQ, bipolar disorder
9. Psychotic Disorders
   1. PQ-B, psychosis risk
   2. DES-II, dissociation
10. Personality Pathology
    1. PID-5-BF, personality pathology

The initial intention was to provide data on all 10 groups, but datasets were only found for depression, anxiety, PTSD, and OCD. This challenge will be discussed further later in this work.

After converting each dataset and DSM-5 excerpt into .jsonl files with the original text broken into manageable chunks, the all-MiniLM-L6-v2 embedding model tokenized each text chunk as part of an index containing 600 total chunks. This approach allows for a quick yet accurate retrieval of elements from the most relevant datasets and DSM-5 diagnostic criteria when queried. In each user session, the patient’s self-reported symptoms and their PHQ-4 screening results are combined into one query which is used to return the six most semantically similar text chunks. These retrieved chunks are limited to a total of 1,800 characters to prevent model overload while preserving precision. The number of characters and chunks retrieved, as well as the number of chunks in the index, are all variable; the values reported here were chosen given system and model limitation.

**User Interaction**

Each user interaction begins with Olive greeting the patient before asking for their identifying information and a self-report of their symptoms and concerns. Note that every test case used in this report is entirely fictitious and not based on any real person.A screenshot of a computer error message

AI-generated content may be incorrect.After the patient delivers their self-report, Olive administers PHQ-4.  
A screenshot of a computer error

AI-generated content may be incorrect.

Based on the patient’s self-report and PHQ-4 scores, Olive dynamically chooses relevant inventories from the files available in Olive/inventories/ and administers them to the patient.

A black screen with white text

AI-generated content may be incorrect.

Below is a screenshot of the prompt given to the LLM to choose inventories to be administered:

A computer screen with text on it

AI-generated content may be incorrect.

After the patient completes each inventory, Olive thanks them for their participation and generates a file with a name of the format “PatientInitials\_1995\_07\_20\_LlmUsed.csv” to emulate how a clinician might identify a patient without jeopardizing confidentiality. The generated file, pictured below, contains the patient's name, date of birth, self-report, and the LLM’s diagnostic impression based on Olive’s interaction with the patient, followed by the patient’s scores for each inventory completed.

A screenshot of a white sheet with black text

AI-generated content may be incorrect.

Below is a screenshot of the prompt given to the LLM to generate a diagnostic impression:

A computer screen with text

AI-generated content may be incorrect.

In a real use case, a mental health clinician could allow a new patient to have a conversation with Olive to collect self-reported information and inventory results before reviewing the results of the conversation in the output file, following from the many benefits discussed in the literature.

**Experimental Design**

In an attempt

# **Works Cited**

**Bradford A, Meyer AND, Khan S, et al. (2023).** Diagnostic error in mental health: a review. *BMJ Quality & Safety 2024,* 33:663-672. https://qualitysafety.bmj.com/content/33/10/663.citation-tools

**Gallup News. (2025).** Daily loneliness afflicts one in five in U.S. Retrieved March 2025, from <https://news.gallup.com/poll/651881/daily-loneliness-afflicts-one-five.aspx>

**Harvard Gazette. (2021, February 18).** Young adults hardest hit by loneliness during pandemic. Retrieved March 2025, from <https://news.harvard.edu/gazette/story/2021/02/young-adults-teens-loneliness-mental-health-coronavirus-covid-pandemic/>

**Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T., & Stephenson, D. (2015).** Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, 10(2), 227–237. <https://doi.org/10.1177/1745691614568352>

**Hostinger. (2025).** Large language model market growth forecast 2025. Hostinger Analytics Report 2025. *Hostinger*. Retrieved March 2025, from <https://www.hostinger.com/tutorials/large-language-model-market-growth>

**Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., & Pant, S. (2023).** A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*, 11(3), 285. <https://doi.org/10.3390/healthcare11030285>

**Jacobson, N., Heinz, A., & Winters, K. (2025).** Efficacy of a generative AI therapy chatbot: Randomized controlled trial. *NEJM AI*, *2*(3), e11234. <https://doi.org/10.1056/NEJMAI11234>

**Katharine A Smith, et al. (2023).** - Digital mental health: challenges and next steps. *BMJ Mental Health 2023,* 26:e300670. https://mentalhealth.bmj.com/content/26/1/e300670?twclid=2-4ielk6k2msm1gex1xvwv9uccv

**Kim, B. H., & Wang, C. (2025).** Large language models for interpretable mental-health diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, pp. 39–48). <https://doi.org/10.1609/aaai.v39i1.28906>

**Le Glaz A et al. (2021).** Machine Learning and Natural Language Processing in Mental Health: Systematic Review.*J Med Internet Res 2021,* 23(5):e15708. https://www.jmir.org/2021/5/e15708

**Linardon, J., et al. (2023).** Effectiveness of digital mental health tools to reduce depressive symptoms in low- and middle-income countries: A systematic review and meta-analysis. *JMIR Mental Health*, 10(1), e43066. <https://doi.org/10.2196/43066>

**Martin-Key N, Schei T, Barker E, Spadaro B, Funnell E, Benacek J, Tomasik J, Bahn S (2021).** The Current State and Diagnostic Accuracy of Digital Mental Health Assessment Tools for Psychiatric Disorders: Protocol for a Systematic Review and Meta-analysis. *JMIR Res Protoc 2021*. https://www.researchprotocols.org/2021/1/e25382

**Malouin-Lachance, A., et al. (2025).** Does the digital therapeutic alliance exist? An integrative review. *JMIR Mental Health*, 12(1), e69294. <https://doi.org/10.2196/69294>

Malouin-Lachance, A., et al. (2025). Does the digital therapeutic alliance exist? An integrative review. *PubMed*, 39924298.

**Malouin-Lachance, A., et al. (2025).** Exploring user characteristics, motives, and expectations in conversational AI. *Frontiers in Digital Health*, 7, 1576135. <https://doi.org/10.3389/fdgth.2025.1576135>

**Mazumdar, P., et al. (2024).** Evaluating GPT-4 for suicide-risk triage: A simulation study. *JMIR Mental Health*, 11, e47563. <https://doi.org/10.2196/47563>

**Moore, J., et al. (2025).** Stigma in AI therapy chatbots: An experimental analysis. *Journal of Artificial Intelligence Research*, 73, 111–132. <https://doi.org/10.1613/jair.1.14567>

**Nature Digital Medicine. (2025).** Racial bias in AI-mediated psychiatric diagnosis and treatment. *Nature Digital Medicine*, 2, s41746-025-01746-4. <https://doi.org/10.1038/s41746-025-01746-4>

**Nguyen, L., & Soffer, S. (2024).** Experiences of generative AI chatbots for mental health: Potential, limitations, and ethical considerations. *Nature Mental Health*, 1(10), 87–101. <https://doi.org/10.1038/s44159-024-00087-x>

**Omar, M., Soffer, S., & Charney, A. W. (2024).** Applications of large language models in psychiatry: A systematic review. *Frontiers in Psychiatry*, 15, 1422807. <https://doi.org/10.3389/fpsyt.2024.1422807>

**Prescott, M. R., et al. (2022).** Real-world effectiveness of digital mental health services during the COVID-19 pandemic: A systematic review. *PLoS ONE*, 17(8), e0272162. <https://doi.org/10.1371/journal.pone.0272162>

**Raza, S., et al. (2024).** Exploring biases of large language models in the field of mental health. *PMC Articles*, PMC11949086. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11949086/>

**ScienceDirect. (2024).** Enhancing mental health with artificial intelligence: Current trends. *Journal of Digital Health*, S2949916X24000525. <https://doi.org/10.1016/j.jdh.2024.100010>

**Springs Analytics. (2025).** Large Language Model Statistics and Numbers 2025. Springs Analytics Market Report 2025.

**The Jed Foundation. (2024).** Mental health and suicide statistics. Retrieved March 2025, from [https://jedfoundation.org](https://jedfoundation.org/)

**U.S. Surgeon General. (2023).** Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General’s Advisory on the Healing Effects of Social Connection and Community. U.S. Department of Health and Human Services.

**Various sources on LLM bias and safety from search results.** *(Please check and specify exact references for academic rigor)*

**Vernon, C., Holt-Lunstad, J., & Valtorta, N. (2022).** Social isolation trends in the United States, 2003–2020. *Social Science & Medicine*, 314, 115489. <https://doi.org/10.1016/j.socscimed.2022.115489>

**World Health Organization. (2025).** Suicide fact sheet. Retrieved March 2025, from <https://www.who.int/news-room/fact-sheets/detail/suicide>

**Yellow Bus ABA. (2024).** Loneliness statistics. Retrieved March 2025, from <https://www.yellowbusaba.com/post/loneliness-statistics>

**Zhu, J., et al. (2023).** Global burden and trends of depressive and anxiety disorders: A systematic analysis of the Global Burden of Disease Study 2019. *The Lancet Public Health*, 8(8), e1221–e1231. <https://doi.org/10.1016/S2468-2667(23)00145-8>