FlexiDialogue: Integrating Dialogue Trees for Mental Health with Large Language Models

João Fernandes[1,2][a], Ana Antunes[1,2][b], Joana Campos[1,2][c], João Dias[2,3,4][d] and Pedro A. Santos[1,2][e]

[1]*Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa, Portugal*

[2]*INESC-ID, Rua Alves Redol, 9, Lisboa, Portugal*

[3]*Faculty of Science and Technology, University of Algarve, Campus de Gambelas, Faro, Portugal*

[4]*CISCA, Campus de Gambelas, Faro, Portugal*

Keywords: Mental Health Virtual Assistants, Dialogue Systems, Large Language Models, Natural Language Understanding, Flexible Dialogue Trees, Mental Health Support, Multilingual Interaction, Conversational AI.

Abstract: The increasing prevalence of mental health issues among university students is exacerbated by limited access to support due to shortages of mental health professionals and the stigma associated with seeking help. Virtual mental health assistants can extend the reach of existing resources, but traditional systems reliant on scripted dialogues are constrained by inflexibility and limited adaptability to diverse user inputs. This paper introduces FlexiDialogue, a system that transforms rigid dialogue trees into instruction sets for large language models, facilitating dynamic, contextually appropriate, and multilingual interactions while maintaining the structure and quality of expert-validated dialogue flows. The system was evaluated in three phases: (1) determining how effectively large language models could map open-ended user responses to predefined dialogue tree options, allowing for more natural interaction without compromising control; (2) assessing the models' ability to paraphrase scripted dialogues to improve conversational fluidity while remaining grounded in the original tree; and (3) conducting an expert review to assess overall performance. Results demonstrated that FlexiDialogue enhanced the flexibility and coherence of interactions, with expert evaluations confirming its potential for mental health support.

1 INTRODUCTION

Concerns about university students' mental health have grown over the past decade (Schmerler et al., 2023). This group is at particular risk as many chronic mental health conditions emerge between the ages of 16 and 24 (McManus and Gunnell, 2020). When untreated, these conditions are linked to declining academic performance, harmful health behaviors, and rising rates of depression, anxiety, and suicidal ideation (Hutchesson et al., 2021). However, stigma and limited campus resources often deter students from seeking help (Pompeo-Fargnoli, 2022).

Socially Interactive Agents (SIAs) offer a promising solution to complement human mental health support resources, particularly where resources are limited or unavailable (Lazzarino et al., 2023). These digital entities can understand, respond to, and form emotional connections with users, which allows SIAs to show promise in mental health applications (Williams et al., 2023). SIAs expand support accessibility, particularly for those hesitant to seek human help, as users often feel less judged and more open to self-disclosure (Holthöwer and Doorn, 2022), crucial in mental health (Doan et al., 2020). They can also provide remote, on-demand access as conversational partners (S et al., 2023).

Traditionally, SIAs for mental health follow a symbolic approach, where developers meticulously define each behavior processed and generated by the

[a] https://orcid.org/0009-0009-8812-2419
[b] https://orcid.org/0009-0009-0512-3062
[c] https://orcid.org/0000-0002-0113-2211
[d] https://orcid.org/0000-0002-1653-1821
[e] https://orcid.org/0000-0002-1369-0085

Fernandes, J., Antunes, A., Campos, J., Dias, J. and Santos, P. A.
FlexiDialogue: Integrating Dialogue Trees for Mental Health with Large Language Models.
DOI: 10.5220/0013286900003938
In *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2025)*, pages 268-275
ISBN: 978-989-758-743-6; ISSN: 2184-4984

agent. In mental health contexts, this pre-scripted approach ensures that agents' decisions align with psychotherapy theory, minimizing risks to users and enhancing their well-being (Antunes et al., 2023b). A crucial technology in this framework is the dialogue tree, which structures interactions by guiding users through predefined conversational paths. Dialogue trees manage the flow of interactions, enabling creators to control conversations (Rose, 2014). Their reliability and feasibility contribute to their popularity in mental health settings (Teixeira et al., 2021). Despite their reliability, dialogue trees face challenges due to inflexibility, often resulting in mechanical interactions that fail to provide dynamic, contextually appropriate responses (Collins et al., 2016). Expanding them to accommodate diverse inputs can increase complexity and maintenance, hindering scalability across domains (Pinho, 2024).

An alternative to the pre-scripted approach is using data-driven models to guide agent behaviors and decisions. Rather than manually defining each dialogue line, researchers can train neural models to dynamically process and generate dialogue. Large Language Models (LLMs) have become a popular tool for developing conversational agents. These extensive neural networks are trained on vast datasets of textual data, allowing them to handle various input types, such as text and speech (Bai et al., 2024; Bharathi Mohan et al., 2024), and infer emotions from user sentences (Zhu et al., 2024). These capabilities enable LLMs to generate fluent responses, enhancing the naturalness of human-agent interactions (Zhang et al., 2019). However, in high-stakes scenarios where control is critical, LLMs alone are insufficient. These models do not truly understand text; they recognize and replicate statistical patterns, which can lead factually incorrect or misleading responses (Bharathi Mohan et al., 2024). Additionally, LLMs can perpetuate harmful biases from their training data, resulting in ethically questionable outputs (Desai et al., 2023).

Motivated by the rigidity of pre-scripted systems and the unreliability of current data-driven approaches, we present FlexiDialogue, a mental health assistant combining dialogue trees' structure with LLMs' adaptability. By grounding LLM responses in validated dialogue trees, FlexiDialogue enables natural conversations while mitigating LLM hallucinations, ensuring high-quality, context-aware dialogues.

We compared three LLMs for this effect: LLaMA 3.1 (Dubey et al., 2024), GPT-3.5 (An et al., 2023) and GPT-4o mini (OpenAI, 2024)—across three evaluation phases: (1) mapping open-ended responses to predefined dialogue options for natural interaction; (2) paraphrasing scripted dialogues while maintaining structure; and (3) expert reviews to assess overall performance. Results showed that FlexiDialogue enhanced flexibility and coherence, with expert evaluations confirming its potential as a valuable mental health support tool.

2 RELATED WORK

2.1 Pre-Scripted Mental Health Agents

Studies indicate that SIAs for mental health often use pre-scripted dialogues for clear communication and essential information gathering. For example, Woebot (Siddals et al., 2024) uses Cognitive Behavioral Therapy principles to offer structured dialogues promoting emotional regulation and self-reflection. Similarly, Tess (Belser, 2023) provides accessible, cost-effective support through mental health professional-approved responses, operating across multiple platforms, including Facebook Messenger, mobile text, and voice-enabled services such as Alexa and Google Home, enhancing accessibility.

MHeVA is another pre-scripted SIA for early anxiety detection in students (Antunes et al., 2023b).

The system uses Theory of Mind to assess the rapport it establishes with the student, allowing it to adapt the conversation dynamically. If a sufficient level of trust is built, MHeVA progresses to more sensitive topics related to anxiety; otherwise, it focuses on further enhancing rapport by discussing non-sensitive subjects. The agent carefully manages the flow of conversation to avoid abrupt changes in topics, ensuring a smooth and comfortable interaction. At the end of the dialogue, MHeVA provides feedback on the student's anxiety status, offering insights into whether anxiety is present and, if so, how severe it might be. MHeVA aims to help students recognize potential anxiety issues early on, promoting mental health support in a discreet and accessible manner.

We leveraged MHeVA dialogue tree structure and full access to test and enhance the flexibility of its response generation and to interpret students' input by using LLMs. Further, inspired by (Belser, 2023), we aim to increase MHeVA accessibility by integrating the system into Whatsapp.

2.2 Large Language Models for Mental Health Agents

In the context of mental health support, LLMs have been applied to address challenges such as loneliness and suicide risk among students, with Rep-

lika (Maples et al., 2024) providing empathetic interactions to support student well-being and mental health needs. Furthermore, Psy-LLM (Lai et al., 2023) leverages AI-based LLMs to scale up psychological services globally, enhancing accessibility and providing tailored support for diverse populations. These approaches aim to enhance user experience by enabling flexible, context-aware communication. However, a common challenge is the lack of grounding mechanisms to ensure LLM generated responses align with clinical guidelines. Without grounding, generated content may diverge from recommended practices, risking reliability.

To address this, prompt engineering has been used to provide grounding, ensuring outputs adhere to validated guidelines. For example, SouLLMate (Guo et al., 2024) and Wysa (Legaspi Jr et al., 2022) utilize prompt engineering to align responses with mental health support needs and therapeutic guidelines. Additionally, LLMs have been grounded in principles from agent-based theories, specifically the Belief-Desire-Intention and Ortony-Clore-Collins models (Antunes et al., 2023a).

Our approach similarly grounds LLM-driven dialogue within an expert-validated dialogue tree, ensuring conversations follow psychological frameworks and meet the needs of context-sensitive communication in mental health settings.

3 FLEXIBILIZING INTERVENTION GUIDES FOR MENTAL HEALTH

We present FlexiDialogue, a system that combines pre-scripted dialogue trees with LLMs to create agents capable of processing multilingual user inputs while offering flexible, non-rigid responses (Figure 1). We followed a three-step approach: (1) creation of an agent based on a dialogue tree, (2) integration of LLMs into the agent for grounded processing of user input and generation of contextually appropriate responses, and (3) implementation of the remote communication platform.

3.1 Creation of an Agent Based on Dialogue Tree

The system is based on a validated dialogue tree, using the pre-existing tree developed in prior work for MHeVA, a mental health support agent for university students (Antunes et al., 2023b). This tree was chosen for its alignment with our goal of supporting student mental health and its creation in collaboration with a mental health expert, ensuring its relevance and validity. In this implementation, the dialogue tree includes information about MHeVA's character, such as the dialogues it will deliver and tasks it needs to perform. FlexiDialogue incorporates dialogue trees created by the FAtiMA Toolkit, converting MHeVA into Flexi-MHeVA for enhanced interaction. The development of the tree involved expert collaboration to ensure its suitability for addressing anxiety-related topics with university students.

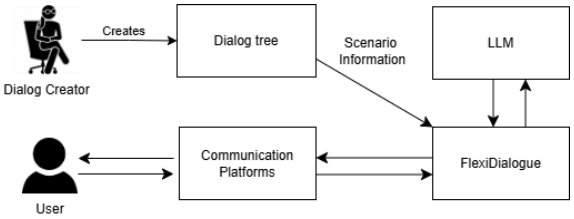

Figure 1: FlexiDialogue's overall architecture. The system followed a three-step approach: (1) creation of a validated dialogue tree by a dialogue creator, where the dialogue tree contains scenario information about the agent and defines how it will interact with users; (2) integration of LLMs into the agent for grounded processing of user input and the generation of contextually appropriate responses; and (3) implementation of a remote communication platform, enabling interaction with users and facilitating their access.

3.2 Integration of LLMs into the Agent

Flexi-MHeVA leverages LLMs for grounded user input processing, which is then mapped to an option within the dialogue tree. It also utilizes grounded response generation, where the selected dialogue option is used to guide the LLM in generating a new context-aware response while ensuring it remains appropriate. To achieve this, prompts were designed with a clear task description incorporating user inputs and dialogue tree options or phrases from the dialogue tree as inputs. This ensures that outputs remain consistent and efficient while aligning with the dialogue tree structure.

3.2.1 Grounded User Input Processing

Unlike the original MHeVA, which used multiple-choice options, Flexi-MHeVA enables free-text responses, using an LLM for natural language understanding (NLU) to interpret user input. To provide the most relevant response, we map the user's input to one of the predefined dialogue tree options, considering the prior conversation context.

When a user interacts with Flexi-MHeVA, the process begins with a question from the original MHeVA

dialogue tree, such as, “Have you ever had an anxiety attack?”. In the original MHeVA, users would select from a list of possible answers. In Flexi-MHeVA, users can provide open-ended responses. The system then uses an LLM with a ranking prompt to analyze both the agent’s question and the user’s response, selecting the most appropriate option from the original dialogue tree. To select the most appropriate option, the LLM uses a ranking mechanism to assess the similarity between the user’s input and the available options. Each option in the dialogue tree is ranked from most to least suitable, with Flexi-MHeVA selecting the top-ranked option to continue the dialogue. This ranking system keeps the conversation on track while enabling a more personalized interaction.

3.2.2 Grounded Response Generation

Once the best dialogue tree option is selected, the corresponding scripted response is retrieved. The LLM then paraphrases and adapts this response to better suit the user’s input, ensuring more natural and personalized interactions.

FlexiDialogue includes a language detection prompt to ensure responses are in the user’s preferred language. It starts by asking the student for their language preference and uses NLU to detect the language based on their response, even if the language is not explicitly stated but inferred from related terms. For example, by default, the system assumes English, but if the student mentions “Portuguese,” it defaults to European Portuguese.

To enhance conversation fluidity, FlexiDialogue employs a phrase generation prompt, generating alternative phrasings of predefined responses to making interactions more natural and less repetitive. Currently, the system limits variations in phrasing to maintain sensitivity to the student’s emotional state while allowing better control over the responses. Before delivering the generated response, the system checks if the language is English. If it is, the response is not translated, otherwise, it is translated into the user’s selected language using the translation prompt, allowing for seamless communication across different linguistic contexts.

3.3 Remote Accessibility

To enhance accessibility, FlexiDialogue leverages WhatsApp as a communication platform, enabling users to remotely engage with Flexi-MHeVA. WhatsApp, being a widely familiar application, simplifies user adoption, making it easier for students to utilize the system (Kaysi, 2023). The integration with Twilio Sandbox uses the WhatsApp Business API, ensuring end-to-end encryption of messages, which guarantees secure communication between users and the system. A Twilio account was created for students to contact the system. Additionally, the combination of Twilio, ngrok, and Flask allows for real-time message exchanges, providing a seamless and secure interaction experience (Miller et al., 2022; Reddy et al., 2022).

After contacting the system via WhatsApp, the user receives a confirmation message from the sandbox indicating that they are connected. To avoid manually entering the user’s phone number each time a connection is made, the user must send any message, even something as simple as ”hello”. Upon receiving this initial message after the connection, FlexiDialogue saves the user’s phone number and initiates the conversation between Flexi-MHeVA and the user by exchanging messages with this specific number. The conversation begins with Flexi-MHeVA explaining that it will be synchronous, with one message sent and responded to at a time. The first question asks about the user’s preferred language, and the conversation proceeds according to the dialogue tree, as shown in Figure 2.

4 EXPERIMENTAL SETUP

This work aimed to develop a flexible and adaptable dialogue system that transforms dialogue trees into dynamic instructions for LLMs. This involved integrating NLU capabilities for more natural and contextually appropriate interactions, as well as adding multilingual support and a communication platform via WhatsApp. The evaluation is structured in two parts. First, the performance of three models—LLaMA 3.1, referred to hereafter as LLaMA 3, GPT-3.5 turbo, and GPT-4o mini—is compared in tasks such as ranking and paraphrasing, assessing their ability to interpret user input and perform the required functions. Multiple iterations were conducted to refine and optimize the prompts for each task in each model. Second, an expert review of Flexi-MHeVA in a simulated multilingual WhatsApp environment, evaluating conversational coherence, sensitivity, and effectiveness as a mental health support tool.

4.1 Evaluation

Grounded Response Generation Evaluation: The dialogue tree’s original phrase was provided to the LLM, which was then prompted to generate a similar phrase. The goal was to evaluate if the generated phrase met specific criteria: it should maintain the core content, only introduce minor word variations,

Figure 2: Beginning, middle, and end of the conversation between user and Flexi-MHeVA.

and avoid introducing any incorrect information. We compared the original and generated phrases to ensure adherence to these rules, as accurate paraphrasing is essential in maintaining rapport and ensuring sensitive phrasing, especially for mental health support.

Grounded User Input Processing Evaluation: When Flexi-MHeVA posed a question, we sent a response to assess the LLM's ability to comprehend the input accurately and select the correct option, aiming for the one most similar in meaning to our response. This evaluation focused on the LLM's interpretative accuracy in identifying the intended answer, essential for enabling a coherent and contextually grounded interaction within the dialogue system.

Expert Evaluation: After completing the prompt tests and ensuring that Flexi-MHeVA was fully operational, Flexi-MHeVA was tested by a student support expert involved in developing the original MHeVA dialogue tree. The expert simulated student scenarios with varying anxiety levels, interacting with the system in Portuguese, English, and Spanish via WhatsApp. Post-interaction, the expert completed a questionnaire and participated in an interview to assess system performance, multilingual support, and conversational coherence.

4.2 Grounded Response Generation Results

In paraphrasing, LLaMA 3 struggled to produce complete sentences and made errors, such as stating "at least 6 hours" instead of "more than 6 hours". GPT-3.5 also had shortcomings, as it failed to avoid using certain words, like "doom," which could negatively impact student sensitivity. On the other hand, GPT-4o mini successfully generated similar phrases while avoiding overly strong words, addressing the issue that GPT-3.5 encountered with terms like "doom".

4.3 Grounded User Input Processing Results

The subsequent Table 1 compares the ranking performance of these three models, providing insight into their accuracy and effectiveness in selecting the correct option. The term "Incomplete" indicates that the model selected the correct option as its response but did not include all available choices, resulting in one or more missing options. Conversely, "Incorrect" signifies that the model failed to provide the correct answer. Across 77 tests, LLaMA 3 achieved 53.2% correct responses, 33.8% partially correct, and 13% incorrect. GPT-3.5 achieved 87.01% correct responses, 3.9% incomplete, and 9.09% incorrect. GPT-4o mini demonstrated 100% accuracy, with no incomplete or incorrect responses. Considering only fully correct responses, LLaMA 3's accuracy increases to 87%, and GPT-3.5's to 90.91%. These results highlight high performance levels for all models, with GPT-4o mini being the most accurate.

4.4 Expert Results

The expert filled out a questionnaire, with the feedback provided in Tables 2 e 3, and an interview was conducted to gather further insights. The results from the expert's evaluation indicated that Flexi-MHeVA demonstrated empathy, reduced the stigma associated with seeking help for mental health issues, and delivered natural, coherent dialogue. The questions were clear, the diagnostic assessments were accurate, and the system effectively supported interactions in English, Spanish, and Portuguese. However, despite these strengths, the evaluation also identified key areas for improvement to enhance Flexi-MHeVA's ef-

Table 1: Results of Ranking applied at LLaMA 3 and GPT-3.5 and GPT 4o-mini out of 77 questions and answers.

LLM	Correct	Correct (%)	Incomplete	Incomplete (%)	Incorrect	Incorrect (%)
LLaMA 3	41	53.2%	26	33.8%	10	13%
GPT-3.5	67	87.0%	3	3.9%	7	9.1%
GPT-4o mini	77	100%	0	0%	0	0%

fectiveness as a mental health support tool. Its ability to establish rapport and maintain a natural conversation flow may depend on individual preferences, such as whether students prefer voice calls, avatars, or text-based communication, all of which influence engagement and user experience. In general, Flexi-MHeVA did not affect sensitivity. While Flexi-MHeVA generally did not negatively impact sensitivity, there were instances where the choice of language could have been more considerate. For example, during a conversation, Flexi-MHeVA used the phrase:

> "I understand. Let me pose another question to you, Rodrigo. Have you ever experienced a sense of fear as if something terrible is about to occur?"

Although Flexi-MHeVA's role is solely to detect anxiety, words like 'terrible' can impact the sensitivity of the response, particularly for someone who may already be experiencing heightened anxiety. Another issue arose when the system generated this sentence, where a synonym for 'nervous' was replaced with 'anxiety', as if assuming the student had anxiety, which could lead to confusion or misinterpretation by the students, for example:

> "Lately, you've been having trouble falling asleep. Is something on your mind causing you anxiety?"

A broader range of response options is necessary to better address the diversity in how students express emotions. Additionally, providing clearer instructions on how students should respond would enhance the interaction, indicating that a different format of response is expected. These adjustments are essential because users often communicate in more elaborate language than Flexi-MHeVA anticipates. Without a broader array of responses, Flexi-MHeVA risks delivering generic replies that may not adequately address specific user concerns, leading to frustration and decreased engagement. Flexi-MHeVA currently lacks clear communication to students that it is intended primarily for triage purposes rather than for addressing serious mental health cases. Moreover, detecting strong language in the student's input is crucial.

At the end of the diagnostic process, if a student shows signs of anxiety, offering support service contacts or yo send an email to the appropriate resources would ensure they are directed to professional help.

Integrating WhatsApp is beneficial as it is familiar to students, allows quick responses, and reduces barriers to seeking mental health support, especially for those hesitant to reach out in person. However, Flexi-MHeVA's responses could be more human-like. Since interactions are synchronous, where users must wait for a response before continuing, improving message synchronization and making interactions more natural is essential. Real-life conversations are often asynchronous, with free-flowing exchanges. Enhancing Flexi-MHeVA to better mimic this would improve its effectiveness in supporting students.

Table 2: Response Scale Options from the Questionnaire completed by the Expert about the interaction with Flexi-MHeVA.

	Strongly Disagree	Disagree	Partially disagree	Neutral	Partially agree	Agree	Totally agree
The agent showed empathy during the conversation.						x	
The agent reduces the stigma of seeking help for mental health issues.						x	
The conversation flowed naturally and coherently.						x	
The questions were clear and easy to understand.						x	
The diagnosis or assessment made by the agent was correct and appropriate.						x	
The use of WhatsApp has made communication more comfortable and efficient.							x
The messages were appropriate and didn't affect sensitivity.						x	
Interaction with the agent would be useful for students.							x

5 DISCUSSION

For Flexi-MHeVA to function effectively, clear and precise instructions are essential. When prompts are explicit and straightforward, the LLM generally follows guidelines accurately. This contrasts with a potential "wizard" mode scenario, where a human intermediary reviews the agent's responses before they reach the user. Since Flexi-MHeVA relies on autonomous response generation, prompt design is critical to maintain functionality without real-time supervision. Effective prompt engineering requires iterative testing and refinement to align with each model's strengths and limitations.

Although GPT-4o mini showed superior performance in tests, further evaluation across diverse topics is necessary to confirm its robustness. The same applies to GPT-3.5 and LLaMA 3, whose performance may improve with refined prompts. Overall, these insights underscore the need for prompt designs that are explicit, well-structured, and concise to optimize performance across varied interactions.

Flexi-MHeVA demonstrates strengths in maintaining empathy, rapport, and a natural dialogue flow—all crucial for reducing stigma around seeking mental health support. The agent performed effectively in delivering clear, accessible communication across English, Spanish, and Portuguese, though certain limitations in phrasing emerged. In some cases, Flexi-MHeVA used terms like "terrible" or made assumptions about user's anxiety levels without sufficient contextual sensitivity, which can risk undermining the system's empathy. Although GPT-4o mini generally avoided terms like "doom," it occasionally

Table 3: Open-ended Responses from the Expert in the Questionnaire.

Question	Answer
List any problems you encountered during the interaction.	The range of answers should be wider. There should be more detailed instructions on how to respond. There should be prompts for certain words and directions to support offices, helplines or other important support.
Did the agent manage to create a safe and welcoming environment during the interaction? What could be done to improve this relationship?	Yes, the dialog tree.
Do you think the student would like the interaction with the agent? Do you think the student would use it again?	Yes.
Do you think the use of WhatsApp facilitates access to this type of mental health resource?	Yes.
Do you think the conversation would be suitable on WhatsApp or would it be better on another platform?	I find the WhatsApp platform suitable.

produced phrases that might unsettle individuals with anxiety, underscoring the need for expert review to refine phrasing and ensure emotional safety.

Despite these strengths, identified challenges suggest areas for improvement. Future directions include refining prompt phrasing to ensure clarity, empathy, and appropriateness in mental health contexts. Additionally, integrating message synchronization could improve conversation flow, making it more natural. Testing Flexi-MHeVA with university students could provide valuable insights into rapport and empathy, potentially exploring new interfaces like Virtual Reality or Augmented Reality to enhance user engagement. Further, adding a feature to connect users to professional support, such as automated email referrals, could bridge the gap between virtual and real-world resources. Additionally, creating specialized prompts tailored to different types of conversations (e.g., sensitive health topics versus casual dialogue) could improve the agent's ability to respond appropriately and sensitively. Expanding the dialogue tree to support open-ended responses and offering users clear instructions, such as when a "yes" or "no" response is appropriate, could further enhance the interaction flow. Finally, implementing a mechanism to detect urgent language in user inputs could identify users needing immediate assistance and direct them to appropriate resources. With these enhancements, Flexi-MHeVA could become a scalable, accessible resource, complementing traditional mental health services, especially in remote or resource-limited settings where immediate help may not be available.

6 CONCLUSION

This work introduced Flexi-MHeVA, a flexible mental health assistant that uses LLMs to transform static dialogue trees into dynamic, engaging conversations. Expert evaluations highlighted its potential to foster rapport, support multilingual interactions, and reduce the stigma associated with seeking mental health support. Integration through WhatsApp ensures accessibility and convenience for remote users. The evaluation also emphasized the critical role of prompt engineering and model capabilities in performance, with GPT-4o mini excelling in ranking tasks and generating contextually appropriate responses. While further testing and refinement are necessary, these promising results position Flexi-MHeVA as a valuable tool for accessible and effective mental health support.

ACKNOWLEDGEMENTS

We express our gratitude to Dr. Carla Boura for sharing her knowledge and contributing to the evaluation process. This study received Portuguese national funds from FCT - Foundation for Science and Technology through the PhD grant 2021/06419/BD, projects UIDB/50021/2020, SLICE PTDC/CCI-COM/30787/2017, IDP/04326/2020, and Project CRAI C628696807-00454142 (IAPMEI/PRR).

REFERENCES

An, J., Ding, W., and Lin, C. (2023). Chatgpt. *tackle the growing carbon footprint of generative AI*, 615:586.

Antunes, A., Campos, J., Guimarães, M., Dias, J. a., and Santos, P. A. (2023a). Prompting for socially intelligent agents with chatgpt. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, IVA '23, New York, NY, USA. Association for Computing Machinery.

Antunes, A., Guimarães, M., Santos, P. A., Dias, J., Boura, C., and Campos, J. (2023b). Mheva: Mental health virtual assistant for high education students. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–4.

Bai, Y., Chen, J., Chen, J., Chen, W., Chen, Z., Ding, C., Dong, L., Dong, Q., Du, Y., Gao, K., et al. (2024). Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Belser, C. A. (2023). *Comparison of natural language processing models for depression detection in chatbot dialogues*. PhD thesis, Massachusetts Institute of Technology.

Bharathi Mohan, G., Prasanna Kumar, R., Vishal Krishh, P., Keerthinathan, A., Lavanya, G., Meghana, M. K. U., Sulthana, S., and Doss, S. (2024). An analysis of large

language models: their impact and potential applications. *Knowledge and Information Systems*, pages 1–24.

Collins, J., Hisrt, W., Tang, W., Luu, C., Smith, P., Watson, A., and Sahandi, R. (2016). Edtree: Emotional dialogue trees for game based training. In *E-Learning and Games: 10th International Conference, Edutainment 2016, Hangzhou, China, April 14-16, 2016, Revised Selected Papers 10*, pages 77–84. Springer.

Desai, B., Patil, K., Patil, A., and Mehta, I. (2023). Large language models: A comprehensive exploration of modern ai's potential and pitfalls. *Journal of Innovative Technologies*, 6(1).

Doan, N., Patte, K. A., Ferro, M. A., and Leatherdale, S. T. (2020). Reluctancy towards help-seeking for mental health concerns at secondary school among students in the compass study. *International Journal of Environmental Research and Public Health*, 17.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, Q., Tang, J., Sun, W., Tang, H., Shang, Y., and Wang, W. (2024). Soullmate: An adaptive llm-driven system for advanced mental health support and assessment, based on a systematic application survey. *arXiv preprint arXiv:2410.11859*.

Holthöwer, J. and Doorn, J. (2022). Robots do not judge: service robots can alleviate embarrassment in service encounters. *Journal of the Academy of Marketing Science*, 51:1–18.

Hutchesson, M. J., Duncan, M. J., Oftedal, S., Ashton, L. M., Oldmeadow, C., Kay-Lambkin, F., and Whatnall, M. C. (2021). Latent class analysis of multiple health risk behaviors among australian university students and associations with psychological distress. *Nutrients*, 13(2):425.

Kaysi, F. (2023). Mobile instant messaging application habits among university students. *Interactive Learning Environments*, 31(5):3211–3229.

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., and Wang, Z. (2023). Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Lazzarino, A. I., Salkind, J. A., Amati, F., Robinson, T., Gnani, S., Nicholls, D., and Hargreaves, D. S. (2023). Inequalities in mental health service utilisation by children and young people: a population survey using linked electronic health records from northwest london, uk. *Journal of Epidemiology and Community Health*.

Legaspi Jr, C. M., Pacana, T. R., Loja, K., Sing, C., and Ong, E. (2022). User perception of wysa as a mental well-being support tool during the covid-19 pandemic. In *Proceedings of the Asian HCI Symposium 2022*, pages 52–57.

Maples, B., Cerit, M., Vishwanath, A., and Pea, R. (2024). Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj mental health research*, 3(1):4.

McManus, S. and Gunnell, D. (2020). Trends in mental health, non-suicidal self-harm and suicide attempts in 16–24-year old students and non-students in england, 2000–2014. *Social Psychiatry and Psychiatric Epidemiology*, 55(1):125–128.

Miller, H. N., Voils, C. I., Cronin, K. A., Jeanes, E., Hawley, J., Porter, L. S., Adler, R. R., Sharp, W., Pabich, S., Gavin, K. L., et al. (2022). A method to deliver automated and tailored intervention content: 24-month clinical trial. *JMIR Formative Research*, 6(9):e38262.

OpenAI (2024). *GPT-4o mini: advancing cost-efficient intelligence*. Accessed: 2024-09-02.

Pinho, B. d. S. (2024). Planejamento não-determinístico para o gerenciamento do agente de diálogo plantão coronavírus.

Pompeo-Fargnoli, A. (2022). Mental health stigma among college students: misperceptions of perceived and personal stigmas. *Journal of American college health*, 70(4):1030–1039.

Reddy, V. N., Reddy, S. M., Vamshi, A. Y., Reddy, K. N., Dhanunjay, B., and Gopal, S. V. (2022). Whatsapp chatbot for career guidance. *International Research Journal of Engineering and Technology (IRJET)*, 9(10):2395–0072.

Rose, C. M. (2014). *Realistic dialogue engine for video games*. The University of Western Ontario (Canada).

S, P., Balakrishnan, N., R, K. T., B, A. J., and S, D. (2023). Design and development of ai-powered healthcare whatsapp chatbot. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pages 1–6.

Schmerler, J., Solon, L., Harris, A. B., Best, M., and Laporte, D. (2023). Publication trends in research on mental health and mental illness in orthopaedic surgery. *JBJS Reviews*, 11.

Siddals, S., Coxon, A., and Torous, J. (2024). " it just happened to be the perfect thing": Real-life experiences of generative ai chatbots for mental health.

Teixeira, M. S., Maran, V., and Dragoni, M. (2021). Towards semantic-awareness for information management and planning in health dialogues. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 372–377. IEEE.

Williams, A. J., Freed, M., Theofanopoulou, N., Daudén Roquet, C., Klasnja, P., Gross, J., Schleider, J., and Slovak, P. (2023). Feasibility, perceived impact, and acceptability of a socially assistive robot to support emotion regulation with highly anxious university students: mixed methods open trial. *JMIR Mental Health*, 10:e46826.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhu, Q., Chong, L., Yang, M., and Luo, J. (2024). Reading users' minds from what they say: An investigation into llm-based empathic mental inference.