The Limbic Layer: Transforming Large Language Models (LLMs) into Clinical Mental Health Experts

Max Rollwage
Limbic
London, UK
max@limbic.ai

Keno Juchems
Limbic
London, UK
keno@limbic.ai

Sashank Pisupati
Limbic
London, UK
sashank@limbic.ai

George Prichard
Limbic
London, UK
george@limbic.ai

Annamaria Balogh
Limbic
London, UK
annamaria@limbic.ai

Jessica McFadyen
Limbic
London, UK
jess@limbic.ai

Tobias U. Hauser
Limbic, London, UK
University College London, London, UK
Eberhard Karls University of Tübingen, Tübingen, Germany
tobias@limbic.ai

Ross Harper
Limbic
London, UK
ross@limbic.ai

Abstract

Large Language Models (LLMs) have emerged as powerful tools with potential applications across multiple sectors, including healthcare, where resource constraints make efficiency gains particularly valuable. However, the domain of mental healthcare presents distinct challenges, as its vulnerable patient population necessitates high standards of clinical performance and safety. To address this, we introduce the Limbic Layer — a novel system of machine learning (ML) models designed to augment LLMs with specialized clinical decision-making capabilities specific to a mental health setting. This study evaluated the impact of the the Limbic Layer on clinical performance and safety compared to a state-of-the-art, stand-alone LLM (OpenAI GPT-4). We investigated the impact from the perspective of service users as well as trained clinicians. In the first phase, users interacted with either a prompted LLM or an LLM powered by the Limbic Layer. We found that the Limbic Layer demonstrated superior performance in both therapeutic relationship building and clinical skills. In the second phase, clinicians blindly assessed the user conversations, rating the Limbic Layer as significantly more clinically accurate than the stand-alone LLM. This enhanced clinical accuracy was reflected in clinician preference, with 94% indicating they would prefer patients to be treated by the Limbic Layer compared to a stand-alone LLM. These findings suggest that the Limbic Layer significantly improves the clinical performance and safety profile of state-of-the-art LLMs, unlocking their implementation in real-world clinical settings.

1 Introduction

Large Language Models (LLMs) represent a significant advancement in artificial intelligence technology [7], with demonstrated applications across various domains including content creation and customer service [16]. However, their application in mental health treatment presents unique challenges due to the specific vulnerabilities of the patient population and the complexity of clinical decision-making required. These factors necessitate heightened standards for safety and clinical

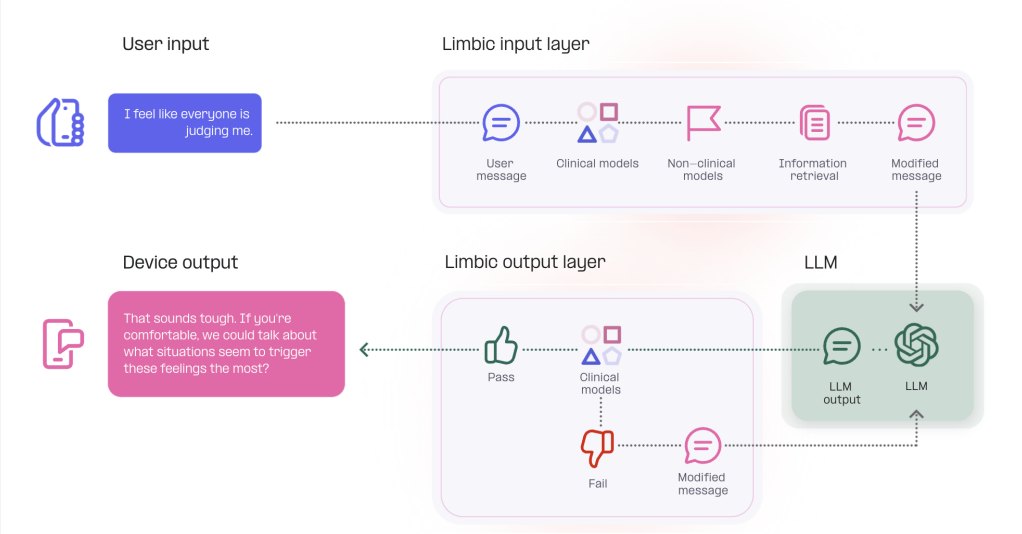

Figure 1: Illustration of Limbic Layer. The Limbic Layer sits as a mediating Layer between the user and the LLM. By processing and evaluating user input and guiding the LLM based on clinical insights, the Limbic Layer ensures the highest clinical accuracy. In turn, by assessing the LLM output, the Limbic Layer also ensures an optimal safety profile.

accuracy, creating barriers to LLM adoption in this field [6]. At the same time, current mental healthcare faces significant resource constraints, with labor-intensive therapeutic processes often resulting in extended wait times and suboptimal patient experiences [1, 10, 15]. This dichotomy creates an urgent need for innovative solutions to address the escalating mental health challenges faced by individuals worldwide [13].

The advent of LLMs has created an opportunity for artificial intelligence (AI) to support this struggling healthcare sector [14]. One prospect is an AI companion, providing continuous AI-assisted support for individuals seeking mental health assistance [9]. However, significant concerns remain regarding the ethical implications and clinical safety of LLM deployment in therapeutic settings [11]. Clinical practitioners may be doubtful as to whether LLMs are ready for this application, due to critical safety implications, as well as uncertainty around clinical performance.

This study addresses these concerns by investigating LLM capabilities within a mental health context, specifically investigating whether specialized clinical augmentation can enhance their therapeutic utility. We present the Limbic Layer, a novel ensemble of ML models specifically developed and trained for specialized clinical functions (see Figure 1). These models have been trained on large, domain-specific, real-world data to move their clinical capacity beyond textbook knowledge. The system architecture positions the Limbic Layer as a mediator between LLM-generated responses and user interactions, implementing clinical decision-making protocols while maintaining natural conversation flow. Thus, the Limbic Layer effectively guides the behavior of LLMs and provides safety guardrails to deliver talking therapies in a reliable, effective, and safe manner.

To evaluate whether the Limbic Layer improves clinical performance of LLMs in a mental health setting, we conducted a two-part study with both users and trained clinicians. Phase 1 examined user interactions, recruiting individuals with previous mental health experience to engage with both a standard prompted LLM and an LLM equipped with the comprehensive clinical tools and logic provided by the Limbic Layer. Users rated the quality of their interaction on dimensions of satisfaction and usefulness. Phase 2 involved evaluation by practicing cognitive-behavioral therapy (CBT) clinicians, who assessed conversation transcripts from Phase 1 for clinical accuracy and safety.

We observed significantly enhanced performance metrics for LLMs augmented with the Limbic Layer across clinical usefulness, accuracy, and safety, as evaluated by both users and clinicians. These findings position the Limbic layer as a potential pathway for implementing LLM technology in clinical mental health settings while maintaining appropraite safety and efficacy standards.

2 Method

2.1 The compared models

2.1.1 Conversational feature

Here, we compared a stand-alone LLM with a system (the Limbic Layer) that combined LLMs with specialized clinical ML-models to guide and safeguard the behaviour of the LLM.

Both models were prompted to recreate a conversation with the user akin to a cognitive behavioral therapy session. This included four major components:

- Agenda setting: agreement on which problem or issue the user would like to discuss in their session.
- Exploration: guided questioning that aims to more deeply explore of the problem and the contributing factors to the issue.
- Treatment recommendation: the information from the exploration phase is used to decide on the most useful CBT exercise for the user's problem, which is then explained to the user to collaboratively decide which exercise to conduct.
- Intervention delivery: the user is guided through the steps of the chosen CBT exercise.

Broadly, this conversation structure follows the core structure of a CBT therapy session [4]. Therefore, the user experience should resemble the interaction with a therapist. This structural similarity allows us to evaluate the conversations with the LLM on similar dimensions as therapy sessions are evaluated during CBT therapist training.

2.1.2 Model design

For this study, we compared a prompted GPT-4 model with a system combining the Limbic Layer with GPT-4 (LLM + Limbic Layer). Both models were prompted to deliver the same form of a guided session, such that the system prompts were optimized to achieve the highest level of performance possible on this task. Importantly, we wished to ensure that both models were equated in every detail besides the input from the specialized clinical ML models, so that group difference could be clearly attributed to the effect of the Limbic Layer rather than to prompt engineering. Therefore, the system prompts were equated in both conditions. Moreover, the groups were also equated and blinded in terms of the user interface, instructions, and the scenarios that should be discussed by users.

The only difference between the two conditions was that the clinical decision-making was driven by specialized clinical ML models of the Limbic Layer in the experimental group, with the hypothesis that these specialized models would lead to superior clinical decision-making and therefore more clinically valuable and accurate therapeutic conversations.

We chose GPT-4 as our comparison model as at the time of the data collection it was the most performant model on multi-step conversation benchmarks in the LLM chatbot arena (https://lmsys.org/blog/2023-05-25-leaderboard/).

2.2 Part 1: User Interaction

2.2.1 Study design

For the initial phase of this study, which focused on user interaction, a total of 40 participants were recruited through the online recruitment platform Prolific. Participants were randomly allocated to test one of the conditions (i.e. LLM or LLM + Limbic Layer) in a between-subjects study design. Two participants were excluded from the study as there were problems with saving their interactions with the models, leading to a final sample of N=38 participants. Participants were screened during recruitment based on self-identification as individuals with prior (but not current) experience dealing with mental health conditions. Crucially, they also affirmed that they were not currently in urgent need of mental health support or at risk of harming themselves or others. Prior to their involvement in the study, all participants underwent a process of informed consent. Participants were compensated with a payment of £10 for their participation in the study.

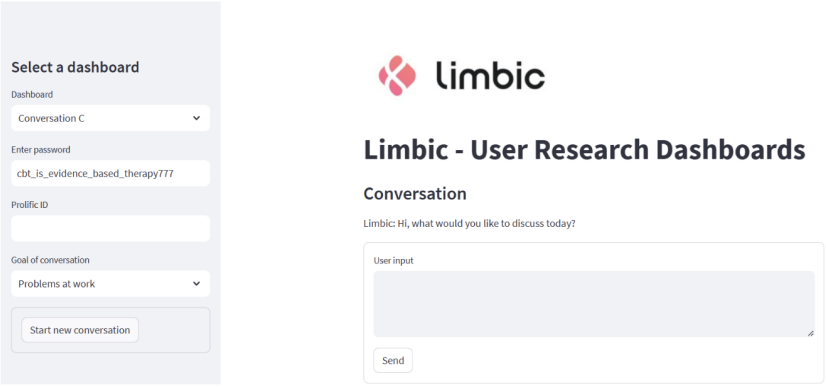

Figure 2: User Evaluation Interface for testing both GPT-4 + Limbic Layer and GPT-4

The users were instructed to discuss problems that they experienced in their everyday life. As such, the users were completely free to discuss whatever scenario they wished to discuss. To help the users generate ideas of what to discuss, five broad potential scenario classes were suggested in the user instructions: work, work-life balance, socializing, relationships, and motivation. Over the course of approximately one hour, participants interacted with the models through a user-friendly messaging interface (see Figure 2). This interface provided a seamless connection to the models, hosted on a dedicated website. Importantly, participants were given the freedom to converse openly within these scenarios, and they received no specific guidelines or instructions on what exactly to discuss with the models.

Participants were blind to the nature of the specific LLM they were engaging with. This approach ensured that their interactions were based purely on their responses to the content presented to them, rather than being influenced by any preconceived notions about the LLM.

2.2.2 Measures of clinical performance

To evaluate the efficacy of the interactions from a user perspective, we employed the Working Alliance Inventory (WAI-SR) which is used as a standard measure in human therapy to assess the quality of the therapeutic relationship between the client and the therapist [8]. This measure serves as a comprehensive metric for gauging the overall satisfaction with therapeutic treatment and the quality of the therapeutic interaction. Notably, therapeutic alliance has been consistently recognized as one of the strongest predictors of treatment outcomes [3], thus representing a critical prerequisite for treatment success. The WAI-SR encompasses the evaluation of the collaborative relationship between the therapist and the patient, encapsulating dimensions such as emotional bond, alignment regarding the prioritized treatment goals, and consensus on the tasks integral to achieving these objectives. These subscale dimensions are measured with items such as, "I believe my therapist likes me" for emotional bond, "My therapist and I agree on what is important for me to work on" for goal agreement, and "I feel that the things I do in therapy will help me to accomplish the changes that I want" for task alignment.

2.3 Part 2: Clinician Evaluation

2.3.1 Study design

In the subsequent phase of this study, a cohort of 17 experienced clinicians evaluated the conversation transcripts in terms of safety and clinical accuracy. All clinicians were actively practicing experts in CBT with an average experience of 10.5 years (range 2 to 26 years). This level of expertise ensured the assessments were informed by a rich clinical background when evaluating the clinical accuracy, effectiveness, and utility of the different models. Informed consent was obtained from each clinician before participation.

To ensure the integrity of the evaluation process, we took rigorous steps to eliminate bias. Each clinician was provided with three conversation transcripts from Chatbot A (LLM + Limbic Layer) and

three from Chatbot B (LLM) to evaluate in a within-subjects design. Importantly, the clinicians were fully blinded about the underlying characteristics and architecture of these chatbots. This approach ensured that their assessment was grounded solely in the content and quality of the conversations. Employing a within-subjects design, the clinicians undertook a structured evaluation process, where they were presented with two distinct sets of conversations:

(i) Chatbot A - Chatbot B: In this sequence, clinicians reviewed conversations from the LLM + Limbic Layer (Chatbot A) and then from the stand-alone LLM (Chatbot B).

(ii) Chatbot B - Chatbot A: The sequence was reversed, wherein clinicians first evaluated conversations from the stand-alone LLM (Chatbot B) followed by those from the LLM + Limbic Layer(Chatbot A).

The order of presentation was counterbalanced across clinicians.

Clinicians were instructed that the transcripts were derived from interactions between users and AI chatbots in the context of a digital mental health conversation. They were then tasked to evaluate the clinical performance of each chatbot similarly to how they would evaluate the performance of a trainee CBT therapist.

To ensure comprehensive engagement and assessment, clinicians were instructed to select and provide screenshots of one conversation they favored the most and another they disliked the most. Clinicians devoted approximately one hour to the evaluation process. The conversations were around 2-4 pages ($\sim$800-1600 words) on average. To acknowledge their contribution, each clinician received compensation of £25 for their time and involvement.

2.3.2 Measures of clinical performance

After the detailed evaluation process for each of the chatbots, the clinicians were tasked to quantitatively rate the clinical performance of each chatbot using the the Cognitive Therapy Scale Revised (CTSR) questionnaire [5]. After having evaluated each chatbot independently clinicians were also asked for comparative preference ratings.

The CTSR is an assessment tool employed in the training and approval of CBT trainee therapists. This measure assesses the proficiency of therapists in delivering CBT. The scale is routinely used in the field of mental health as it systematically evaluates two pivotal dimensions: 1) adherence to the prescribed therapy method, and 2) the skill demonstrated by the therapist in implementing the therapeutic approach. The CTSR scale is divided into 4 different subscales:

(i) Assessment of Therapeutic Relation: A crucial facet of the CTSR is the evaluation of the therapeutic relation, focusing on the manner and way of interactions by the therapist. This dimension is scored from 0 to 6, where a score of 0 implies that the therapist's behavior may lead to patient disengagement, fostering a sense of distrust or hostility. On the opposite end, a score of 6 indicates a therapist's exceptional interpersonal effectiveness, even in the face of challenging scenarios.

(ii) Structure and Time Management: The CTSR further evaluates the therapist's adeptness in maintaining session structure and managing time effectively. A score of 0 on this dimension would reflect poor time management that can lead to aimless or excessively rigid sessions. Conversely, a score of 6 denotes excellent time management, indicative of a therapist's capacity to effectively manage session structure, even when confronted with difficulties.

(iii) Conceptual Integration: The assessment of conceptual integration gauges the therapist's ability to develop and integrate appropriate conceptualizations within the therapy framework. A score of 0 indicates a lack of an appropriate conceptualization, while a score of 6 signifies exemplary development and integration of conceptualizations, even in challenging scenarios.

(iv) Therapeutic Change: Lastly, the CTSR evaluates the therapist's application of cognitive and behavioral methods to facilitate therapeutic change. A score of 0 implies a failure or misuse of therapeutic methods. A score of 6 signifies an exceptional range and application of therapeutic methods, even when confronted with challenges.

Finally, we were interested in evaluating how potential changes on these dimensions would translate into preference ratings when directly comparing the LLMs holistically. Hence, as part of the clinician evaluation phase clinicians were asked to directly compare the LLMs and express their

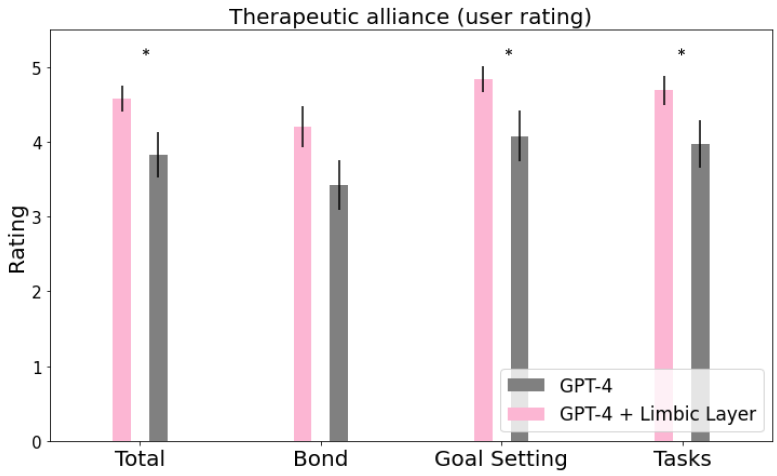

Figure 3: User Evaluation of Therapeutic Alliance (WAI-SR): GPT-4 + Limbic Layer vs. GPT-4. Group average scores and standard errors are presented. *p<.05

holistic preferences across four critical dimensions: general preference, clinical accuracy, perceived harmfulness (i.e. safety), and preference for treating patients in clinical practice.

(i) General Preference: Clinicians were asked to provide insight into their general preference for the stand-alone LLM versus the LLM + Limbic Layer. This aspect aimed to ascertain which model resonated more with the clinicians from a holistic standpoint, considering their overall performance and alignment with clinical principles.

(ii) Clinical Accuracy: Another vital facet is evaluating the clinical accuracy exhibited by the two LLM variations. Clinicians assessed which model captured and conveyed therapeutic insights more accurately.

(iii) Harmfulness Perception: Clinicians were also asked to express their perceptions of potential harm associated with interactions with the two LLM models, thus representing the critical dimension of safety.

(iv) Preference for treating patients: Clinicians were asked which model they would prefer to use in real-world clinical treatment of their patients.

3 Results

3.1 Part 1: User Evaluation

The Working Alliance Inventory (WAI-SR) was employed to discern the quality of the collaborative relationship between users and the LLMs, particularly focusing on three distinct dimensions: emotional bond, goal agreement, and task alignment.

(i) Overall Therapeutic Alliance: A significant increase of the overall therapeutic alliance in interactions involving the Limbic Layer was observed (t=-2.13, p=.020, one-sided; see Figure 3). Users engaging with the LLM+Limbic Layer reported a notably enhanced sense of alliance, which speaks to the Limbic Layer's proficiency in fostering a clinically meaningful and impactful interaction.

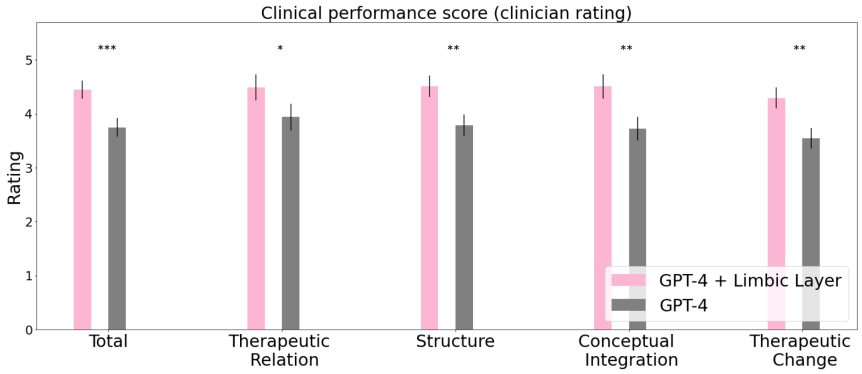

Figure 4: Clinician Evaluation of Clinical Performance (CTSR score): GPT-4 + Limbic Layer vs. GPT-4. Group average scores and standard errors are presented. * p<.05, ** p<.01, *** p<.001

It is noteworthy that the therapeutic alliance achieved through the Limbic Layer was in the range of ratings typically given for human clinicians [12].

(ii) Emotional Bond: No significant differences were observed on the emotional bond subscale of the WAI-SR ($p > .05$).

(iii) Goal Agreement: In contrast, the goal agreement subscale showed higher scores for the LLM + Limbic Layer condition (t=-2.0, p=.027, one-sided), underscoring the Limbic Layer's potential for enhancing the alignment between users and the LLM on the identified therapy goals.

(iv) Task Alignment: Similarly, the task alignment subscale exhibited a significant difference (t=-1.91, p=.032, one-sided) between the two groups. Users engaging with the LLM + Limbic Layer expressed heightened satisfaction with the perceived effectiveness of the therapeutic tasks recommended by the system.

In summary, the user evaluation revealed clinical benefits of the Limbic Layer, leading to higher ratings of therapeutic alliance that were specifically related to the subscales of goal agreement and task alignment.

3.2 Part 2: Clinician Evaluation

3.2.1 CTSR Results

To evaluate the clinical performance of the two different LLM conditions, the CTSR scale was used to assess clinical performance (see Figure 4).

(i) Overall Clinical Competence: The incorporation of the Limbic Layer had a significant impact on the overall clinical competence of the LLM, as evidenced by a significant improvement on the total CTSR score (t=4.03, p<.001). This outcome indicates that the Limbic Layer leads to overall improved clinical decision-making. Interestingly, the average score of the LLM + Limbic Layer was 4.45, on a scale that ranges between 0 and 6. This is a high score even for human therapists, with therapist average scores ranging around this level [2].

(ii) Therapeutic Relation: In the assessment of the therapeutic relation, the benefit of the Limbic Layer was once again evident with significantly higher therapeutic relations than the stand-alone LLM (t=2.17, p=.045). This indicates that the Limbic Layer leads to more efficient relationship building with the user, as judged from an expert clinician's perspective.

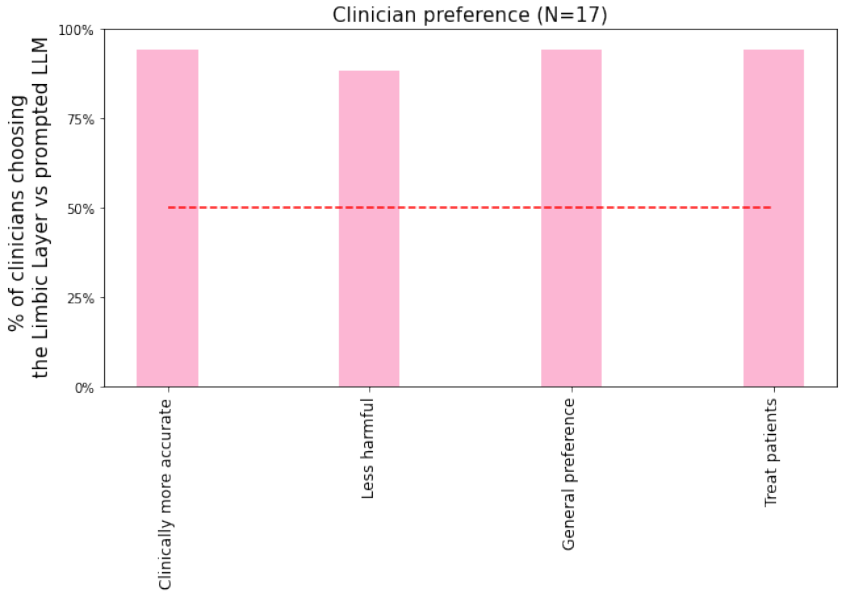

Figure 5: Clinician preference ratings: percentage of clinicians Preferring Limbic Layer + GPT-4 compared to vanilla GPT-4 across multiple clinical dimensions.

(iii) Structure and Conceptual Integration: The LLM + Limbic Layer exhibited significant improvements in both the structural (t=3.55, p=.0027) and conceptual integration (t=3.38, p=.0038) dimensions of the CTSR. This further accentuates the substantial impact of the Limbic Layer in refining the structuring of therapeutic sessions and the identification of relevant clinical characteristics contributing to the experienced problems.

(iv) Therapeutic Change: In the evaluation of the therapeutic change dimension, the Limbic Layer's influence was once again evident, with a clear improvement on the stand-alone LLM (t=3.78, p=.0016). This result further solidifies the Limbic Layer's role in facilitating therapeutic change, demonstrating this new technology's ability to enable more accurate decision-making and selection of relevant CBT exercises.

Collectively, the results of the CTSR evaluation showcase the influence of the Limbic Layer on various dimensions of clinical decision-making capabilities (see Figure 4). This alignment between clinician evaluation and user perception strengthens the proposition that the Limbic Layer not only enriches the LLM interactions but also effectively bridges the gap between technology and clinical excellence in mental health support.

3.2.2 Preference ratings

The CTSR scores showed that the Limbic Layer significantly improves LLM performance across multiple clinical dimensions. In a final step, we assessed whether these clinical improvements translated into clinicians having a preference for using the different models in real-world clinical practice with patients. We asked clinicians to choose between the stand-alone LLM and the LLM + Limbic Layer with respect to their preference for general performance, clinically accuracy, safety and preference for treating patients (see Figure 5).

(i) Clinical Accuracy: The superiority of the LLM + Limbic Layer was consistently evident in terms of clinical accuracy. Almost all (16 out of 17) clinicians deemed the Limbic Layer-enhanced model as clinically more accurate than the LLM alone.

(ii) Harmfulness Perception: The assessment of potential harm during interactions showed that the majority (15 out of 17 clinicians) found the Limbic Layer to be less harmful than the prompted LLM, signifying the enhanced safety derived by the Limbic Layer's additional safety layers.

(iii) Clinical Adoption: An overwhelmingly positive sentiment towards the Limbic Layer was further evidenced by 16 out of 17 clinicians expressing their inclination to employ this model for treating

Table 1: Summary results of User and Clinician Evaluation

Clinical Dimension	GPT-4	GPT-4 + Limbic Layer	Percentage Increase	p-value
Therapeutic Change	3.56	4.29	21%	$p = .0016$
Conceptual Integration	3.75	4.54	21%	$p = .0027$
Clinical Structure	3.85	4.54	17%	$p = .0038$
Therapeutic Relation	3.98	4.56	15%	$p = .045$
Clinician Preference	5.89%	94.11%	1600%	$p < .0001$
Therapeutic Alliance	3.94	4.58	16%	$p = .02$

their patients compared to stand-alone LLMs. This not only indicates their trust in the Limbic Layer's clinical utility, but also signifies its potential as an effective tool in their therapeutic toolkit.

(iv) General Preference and Conclusion: The collective preferences of the clinicians overwhelmingly shifted towards the Limbic Layer, with 16 out of 17 clinicians exhibited a strong preference for this model across multiple clinical dimensions.

To conduct statistical tests on these preference questions, we combined all ratings across these preference domains, yielding a highly significant preference for the Limbic Layer compared to the prompted LLM across all dimensions of clinical preference (χ^2 = 47.1, $p < 0.0001$).

4 Discussion

The results of the clinician evaluation (see Table 1) indicate a significant enhancement in various clinical dimensions through the incorporation of the Limbic Layer compared to a stand-alone LLM. Therapeutic Change, Conceptual Integration, Clinical Structure, and Therapeutic Relation all exhibited substantial improvements ranging from 15% to 21% for the LLM + Limbic Layer.

Remarkably, clinician preference leaned overwhelmingly towards the LLM + Limbic Layer combination, with a 94% preference over the stand-alone LLM. This substantial increase of 1600% in preference underscores the transformative influence of the Limbic Layer.

Moreover, the Therapeutic Alliance rated by the users demonstrated a significant 16% increase for the LLM + Limbic Layer condition, further highlighting the positive impact of the Limbic Layer on fostering an effective therapeutic relationship with users. Overall, the comprehensive results underscore the potential of the Limbic Layer in elevating the clinical aspects of LLMs in mental health settings, enhancing clinical performance to expert levels and thus enabling usage of this technology in real-world settings.

Interestingly, the strongest results were observed on clinical dimensions related to clinical decision-making (e.g. task selection) and safety, rather than aspects of the emotional bond between the user and the model. These results are expected as ratings on these subscales are mainly driven by empathetic responses and conversational capabilities which would be driven by the LLMs conversational capabilities, a capability in which stand-alone LLMs already excel and which were equated across groups. Therefore, the benefits of the Limbic Layer seem to be selective to the clinical components of therapeutic conversations related to clinical decision-making and safety.

In the landscape of rapidly advancing technological innovations, the integration of LLMs holds promise for revolutionizing various facets of our world. In this study, we evaluated the performance of LLMs in a mental health setting. The study introduced the concept of the Limbic Layer – a pioneering set of ML models which provide clinical guidance and accurate decision-making capabilities to LLMs for the context of mental healthcare. The Limbic Layer was proven to provide transformative enhancement in the clinical accuracy, usefulness, and overall quality of LLM interactions. These effects were shown in higher therapeutic alliance ratings from users, higher clinical accuracy ratings through clinicians, and a clear preference of clinicians for the real-world usage of the LLM + Limbic Layer compared to a stand-alone LLM.

The augmentation of LLMs with the Limbic Layer unequivocally led to interactions that were not only more clinically accurate but also fostered a higher therapeutic alliance, as highlighted by user endorsements. Clinicians, who provide a benchmark for clinical excellence, echoed this sentiment

with higher clinical skill ratings. Thus, the magnitude of the results were noteworthy and striking. The Limbic Layer achieved clinical performance scores in range of experts [2]. This result does not only exemplify that the Limbic Layer can boost clinical performance of standard LLMs, but importantly that this boost is significant enough to achieve clinical performance of practical relevance. This is a critical finding, indicating that the usage of the Limbic Layer can unlock the utilization of LLMs in real-world clinical settings.

The reliable and significant improvements on all clinical dimensions (20% improvement on clinical dimensions through the Limbic Layer) translated to a striking preference of clinicians for the Limbic Layer for practical clinical application. An overwhelming 94% of clinicians preferred the LLM + Limbic Layer over a stand-alone LLM. To ensure the highest quality of care, clinicians are the gatekeepers for adoption of novel technology in clinical settings. Therefore, their strong preference towards the Limbic Layer suggests a highly increased likelihood of adoption in practice through the utilization of the Limbic Layer.

In essence, this study illuminates the transformative potential of LLMs in a clinical mental health context when combined with appropriate clinical and safety guidance (i.e. the Limbic Layer), paving a way towards a future where technology can achieve clinical excellence. These findings suggest that the combination of LLMs and the Limbic Layer might unlock the full potential of LLMs for mental healthcare.

5 Competing Interests Statement

MR, KL, SP, GP, AB, JM and RH are employed by Limbic Limited and hold shares in the company. TUH is working as a paid consultant for Limbic Limited and holds shares in the company.

References

[1] Rosie Adams, Tony Ryan, and Emily Wood. 2021. Understanding the factors that affect retention within the mental health nursing workforce: a systematic review and thematic synthesis. *International Journal of Mental Health Nursing* 30, 6 (2021), 1476–1497.

[2] Sven Alfonsson, Georgios Karvelas, Johanna Linde, and Maria Beckman. 2022. A new short version of the Cognitive Therapy Scale Revised (CTSR-4): preliminary psychometric evaluation. *BMC psychology* 10, 1 (2022), 1–7.

[3] C. T. Alldredge, G. M. Burlingame, C. Yang, and J. Rosendahl. 2021. Alliance in group therapy: A meta-analysis. *Group Dynamics: Theory, Research, and Practice* 25, 1 (2021), 13–28. `https://doi.org/10.1037/gdn0000135`

[4] Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

[5] Ivy-Marie Blackburn, Ian A James, Derek L Milne, Chris Baker, Sally Standart, Anne Garland, and F Katharina Reichelt. 2001. The revised cognitive therapy scale (CTS-R): psychometric properties. *Behavioural and cognitive psychotherapy* 29, 4 (2001), 431–446.

[6] Johana Cabrera, M. Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical Dilemmas, Mental Health, Artificial Intelligence, and LLM-Based Chatbots. In *Bioinformatics and Biomedical Engineering*, Ignacio Rojas, Olga Valenzuela, Fernando Rojas Ruiz, Luis Javier Herrera, and Francisco Ortuño (Eds.). Springer Nature Switzerland, Cham, 313–326.

[7] Aarohi Srivastava et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615 [cs.CL]

[8] Robert L Hatcher and J Arthur Gillaspy. 2006. Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy research* 16, 1 (2006), 12–25.

[9] Nikolaos Koutsouleris, Tobias U Hauser, Vasilisa Skvortsova, and Munmun De Choudhury. 2022. From promise to practice: towards the realisation of AI-informed mental health care. *The Lancet Digital Health* 4, 11 (2022), e829–e840.

[10] Patrick Larsson, Russell Lloyd, Emily Taberham, and Maggie Rosairo. 2022. An observational study on IAPT waiting times before, during and after the COVID-19 pandemic using descriptive time-series data. *Mental Health Review Journal* ahead-of-print (2022).

[11] Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023. Ethics of large language models in medicine and medical research. *The Lancet Digital Health* 5, 6 (2023), e333–e335.

[12] Thomas Munder, Fabian Wilmers, Rainer Leonhart, Hans Wolfgang Linster, and Jürgen Barth. 2010. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 17, 3 (2010), 231–239.

[13] Brittany N Rudd and Rinad S Beidas. 2020. Digital mental health: the answer to the global mental health crisis? *JMIR Mental Health* 7, 6 (2020), e18472.

[14] Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, Vol. 11. MDPI, 887.

[15] Johannes Thome, Jocelyn Deloyer, Andrew N Coogan, Deborah Bailey-Rodriguez, Odete AB da Cruz E Silva, Frank Faltraco, Cathleen Grima, Snaebjorn Omar Gudjonsson, Cecile Hanon, Martin Hollỳ, et al. 2021. The impact of the early phase of the COVID-19 pandemic on mental-health services in Europe. *The World Journal of Biological Psychiatry* 22, 7 (2021), 516–525.

[16] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv:2304.13712 [cs.CL]