Building Emotional Intelligence into Digital Therapy AI Agents through Neurofeedback

Sam Nallaperuma-Herzberg[1], Rishabh Balse[1], Sonia Kozuth[1], Lilith Stenhouse[1], Anna Bevan[1], Tristan Bekinschtine[1], Pietro Lio[1]

[1]University of Cambridge, UK

Abstract

We present a novel emotionally intelligent agent framework for delivering cognitive behavioural therapy (CBT). The system aggregates text sentiment cues with neurofeedback, yielding a fine-grained perception of user state building empathy into the agent. A reinforcement learning (RL) planner maps this affective state to appropriate therapeutic acts, which are verbalised by a large language model (LLM). To enhance reliability, the LLM agent is augmented with a meta-cognitive control layer that continuously self-monitors and refines of its responses. In preliminary studies, the proposed system has demonstrated improved therapeutic efficacy over standard LLM-based agents, as measured by standard psychotherapy metrics. These results highlight the potential of combining neurofeedback, affective computing, RL decision making, and LLM generation to deliver clinically meaningful, scalable CBT paving the way for safe, personalised mental health support at population scale.

Introduction

Globally, more than one billion people live with a mental disorder. However, the mental health services remain chronically understaffed (World Health Organization 2025). Artificial intelligence (AI) driven digital mental health applications are emerging as an alternative solution. From early rule trees such as Eliza (Weizenbaum 1966) to modern LLM applications such as Woebot (and 2025), Wysa (Beatty et al. 2022) and Koko (Kaponis, Kaponis, and Maragoudakis 2023) have reached tens of millions of users worldwide (Swi 2025). However, there are limitations in these digital therapy applications that are necessary to be addressed in order to scale up their use worldwide and increase their effectiveness. Lack of embodied empathy is one of the limitations. Primarily, this arises from the absence of physiological measures and the lack of adaptation to hidden affect. Another less explored aspect is meta-cognitive self-monitoring. This may lead to LLMs outputting fluent but clinically unsafe advice. Moreover, missing theory of mind is another limitation where static templates cannot reason over a user's evolving beliefs.

Inspired by human psychology we introduce two key aspects of emotional intelligence : empathy and metacognition to AI agent. We integrate a text based sentiment analysis with EEG based stress monitoring to a perception module, a multimodal proximal policy optimisation (PPO) (Yu et al. 2022) planner, a fine-tuned Llama-3 model with CACTUS dataset (Lee et al. 2024), and a meta-cognitive filter. The agent senses sentiment and stress, chooses a CBT strategy, verbalises it, then self-critiques before replying. Evaluations with human participants with standard psychotherapy metric are conducted. The preliminary results suggest the potential for applicability in the real world in a context of digital mental health.

Background

Two key aspects of human emotional intelligence are empathy and meta-cognition. Empathy is the ability to understand the mental state of the others which help humans to adapt to others' state of mind. Meta-cognition considers monitoring and controlling of our own thoughts and behaviour (Fleming 2014).

Pure text sentiment misses activation and stress markers. State of the art EEG affect work reports θ, α and β waves corresponding to stress (Kamińska, Smółka, and Zwoliński 2021). Combining these with linguistic cues yields a richer affective state space.

In the context of digital therapy meta-cognition can correspond to filtering inaccurate or unsafe therapeutic responses. Prior digital therapy work relies on static toxicity filters. Meanwhile, critic guided approaches (Kim et al. 2023; Gou et al. 2024) show effectiveness in automatically flagging hallucinated or off-topic replies.

The Consultation and Relational Empathy (CARE) measure was developed by Mercer et al. (2004) (Mercer et al. 2004) to assess patients' perceptions of relational empathy in clinical consultations. It consists of 10 items, each rated on a 5-point Likert scale from "Poor" to "Excellent." The measure evaluates aspects such as making the patient feel at ease, really listening, and showing care and compassion. The CARE Measure has demonstrated high internal reliability (Cronbach's alpha = 0.92) and strong validity across diverse patient populations .

No prior digital therapy system unifies neurofeedback powered empathy, meta-cognitive self-repair and LLM fluency. The proposed Mindful-AI system fills this gap.

Hyperparameter	Value
dropout rate	0.279
TCN filters	16
GCN hidden units	64
MLP dense units	64

Table 1: Hyperparameters used for training the STReSS model, tuned with Optuna. (Akiba et al. 2019)

Proposed Agent Framework

The proposed framework consists of four agents: Perception, Planner, Generator and Meta-cognition. Perception includes two key aspects stress scoring and sentiment scoring.

Figure 1: Conventional conversational systems (left) versus the proposed system with multi modal sensing adapted conversational capabilities (right).

Stress Perception

ST-GNN Brain Model for Stress Detection The proposed model architecture for modelling EEG stress signals integrates dilated temporal convolutions with graph convolutions to extract multi-scale temporal patterns and aggregate inter-channel dependencies, taking advantage of the inherent spatial-temporal structure of EEG data. Algorithm 1 further explains each step of the model's design. The model is implemented with PyTorch. Table 1 describes the optimised hyperparameters used for training the model.

Derivation of Stress Score Stress score at time t, $Stress_t$ is calculated as:

Let $x_c(t)$ be the raw EEG voltage from channel $c \in \{1, \ldots, C\}$ sampled at 128 Hz. Every $\Delta = 7$s we form a sliding window of length $T = 5$s with $O = 2$s overlap, so that at time $t_i = i\Delta$s we define the windowed data matrix

$$X^{(i)} \;=\; \big[x_c(t_i - T + 1{:}\,t_i)\big]_{c=1}^{C} \;\in\; R^{C\times(T\cdot f_s)}.$$

We convert to Volts via

$$V^{(i)} \;=\; 10^{-6}\,X^{(i)},$$

The stress evaluation model gives a nonlinear mapping

$$s^{(i)} \;=\; f\big(V^{(i)}\big).$$

The proprietary Spatial-Temporal Residual Stress Neural Network (STReSS) model $f(\cdot)$ maps the windowed multi-channel EEG segment $V^{(i)} \in R^{C\times T_s}$) to a real-valued scalar prediction

Algorithm 1: STReSS

Residual TCN
$H \leftarrow X$
for each dilated convolutional block **do**
  $R \leftarrow H$
  $H \leftarrow \mathrm{Conv1D}\big(H\big)$
  $H \leftarrow \mathrm{Norm}(H)$; $H \leftarrow \mathrm{Act}(H)$; $H \leftarrow \mathrm{Dropout}(H)$
  $H \leftarrow H + R$
end for
$H \leftarrow \mathrm{AdaptiveAvgPool1D}(H)$
Residual GCN
for each sample b in batch **do**
  $(E, W) \leftarrow \mathrm{toSparse}\big(A[b]\big)$
  $h \leftarrow H[b]$
  $R \leftarrow h$
  $h \leftarrow \mathrm{GCNConv}\big(h, E, W\big)$
  $h \leftarrow \mathrm{Act}(h)$; $h \leftarrow \mathrm{Dropout}(h)$
  if $\dim(R) \neq \dim(h)$ **then**
    $R \leftarrow \mathrm{LinearProj}(R)$
  end if
  $h \leftarrow h + R$
  store h
end for
$H \leftarrow \mathrm{stack}(\ldots)$
Readout & Classification
$g \leftarrow \mathrm{Readout}(H)$
$g \leftarrow \mathrm{Dropout}(g)$
$\hat{y} \leftarrow \mathrm{MLP}(g)$
return $\hat{y}$

Thus, at each t_i we obtain a continuous stress estimate $\bar{s}^{(i)}$, which can be stored or streamed for real-time monitoring.

Text Sentiment Perception

BERT model (Devlin et al. 2019) trained with goEmotions dataset (Demszky et al. 2020) is employed for emotion classification. For simplicity, multidimensional emotion outputs are aggregated into a single sentiment score. Class probabilities P of emotion categories are taken as individual emotion scores. Accordingly, the Sentiment score $Sent_t$ is defined as:

$$Sent_t = \frac{1}{2}(P_{optimism} + P_{joy} - P_{sad} - P_{angry} + 1). \quad (1)$$

Deep Reinforcement Learning based Planner

Planner consists of a deep reinforcement learning (RL) approach (Sutton and Barto 2018). In RL agent's goal is to learn an optimal policy Π^*, a functional mapping from the observed current state $S_t \in \mathcal{S}$ of the environment to the optimal action A^* ($\Pi^* : S \rightarrow A^*$). A numerical reward $R_{t+1} \in \mathcal{R} \subset R$ is provided for each action $A_t \in \mathcal{A}$. The agent aims to maximize the cumulative future reward G_t for timestep t,

defined as follows:

$$G_t = \sum_{\tau=0}^{T-t-1} \kappa^\tau \, R_{t+\tau+1} \,, \tag{2}$$

where T denotes the total number of timesteps and $\kappa \in [0, 1)$ denotes the discount factor (Sutton and Barto 2018).

State Space The state of the planner agent at time t is represented by a vector consisting the perception parameters stress a_t and sentiment b_t as described above.

$$S_t = \langle Stress_t, Sent_t \rangle. \tag{3}$$

Action Space The agent's discrete action space is encoded as a 1-dimensional vector with 3 possible choices, with each dimension capturing one of the three affective perspectives: positive, neutral and negative.

Reward Our reward aggregates empirically validated markers described above: the difference in sentiment (Δa_t), and stress (Δb_t). Accordingly, we formalise the per-turn reward as:

$$R_t = -\Delta Stress_t + \Delta Sent_t. \tag{4}$$

Proximal Policy Optimisation We employ proximal policy optimisation (PPO) algorithm (Schulman et al. 2017) to find the best policy that maximises the cumulative future reward.

Implementation The RL planner is implemented in PyTorch using Stable-Baselines3 library (Raffin et al. 2021). PPO with a two-layer MLP policy (128 units each). Training hyperparameters include a batch size of 256, Adam optmiser with learning-rate 3×10^{-4} and a discount factor $\gamma = 0.99$. We warm-start the policy with 10-epoch behavioural cloning on 5000 transitions, then fine-tune for 200 PPO epochs.

Generator LLM

Our generator model is based on Llama-3.1-8B and fine-tuned using GaLore (Zhao et al. 2024) on the CACTUS dataset (Lee et al. 2024). The CACTUS dataset is a multi-turn dialogue corpus that emulates real-world interactions between a counselor and a client using Cognitive Behavioral Therapy (CBT). Each data instance contains structured annotations, including the client's *thought*, *patterns*, *cbt_technique*, *cbt_plan*, *attitude*, and the actual *dialogue*. From this resource, we curated a subset of 7,000 dialogues corresponding to the top seven CBT techniques (1,000 examples per technique) and divided them into 6,500 training and 500 test samples. Each dialogue consists of alternating turns between a *client* and a *counselor*. For fine-tuning, we reformatted the data into the Llama-style chat format, where the client turns were assigned to the `user` role and counselor turns to the `assistant` role. Additionally, we included a system prompt indicating the client's attitude (positive, neutral, or negative) as follows:

```
System: "The client has the
following attitude: <attitude>."
```

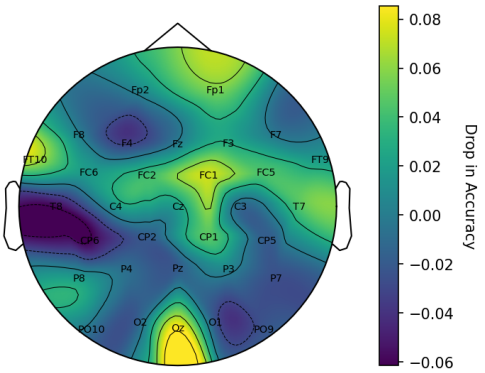
This contextual information allowed the model to tailor its responses according to the client's emotional tone. Fine-tuning was performed using the GaLore optimizer with the following hyperparameters: rank = 32, learning rate = 1e-5, and batch size = 32. The final model was trained on 6.5k dialogues labelled with CBT strategies.

Meta-cognitive Moderator

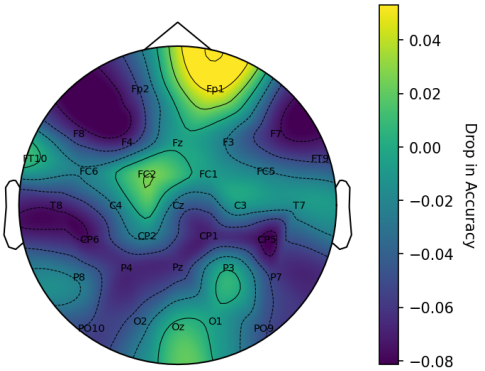
The proposed meta-cognitive moderation mechanism consists a critic framework to constraint the generator's responses. For each turn of conversation, a critic moderation is applied to the generator output. For each turn of conversation, the critic agent is supplied with the toxicity score and conversation history and instructed to provide suggestions to reduce toxicity using its tooling. The therapist attempts to reduce toxicity using this advice. The output is accepted via ensemble once the toxicity falls below a specified threshold and the critic approves the message. A maximum of three iterations are permitted before triggering a safe-harbour message and human therapist intervention. Toxicity is assessed by a RoBERTa classifier (Liu et al. 2019) trained on the Jigsaw dataset (Ian Kivlichan and Culliton. 2020). In order to ground the therapy process, a RAG system is employed to provide clinically relevant context to the critic from approved sources.

Evaluation

The experiments are run on a Nvidia A100 40GB GPU. The hyperparameter settings for each module in the pipeline are as described in the respective sections. STReSS achieves 0.81 $\pm$ 0.04 accuracy on SAM-40 dataset with personal baseline signal subtraction and 0.74 $\pm$ 0.04 without. The architecture enables interpretable readouts, resulting in insights on the importance of specific EEG channels in recognising a stressed state. The ablation studies presented in Figure 2 describe the importance of certain channels in stress detection. Channel ablation affect on model's accuracy, trained on raw EEG in the SAM-40 dataset. For raw EEG based model shown in Figure 2a, channels Oz, FC1, Fp1, T7 and FT10 are among the most meaningful ones for driving the model's ability to distinct the stressed from a non-stressed state, while T8, CP6, F4 likely contain noisy data which hurt performance. The model's accuracy can be elevated by removing some EEG channels from the training data. On the other hand, for cleaned EEG model as shown in Figure 2b, the Fp1 channel is strongly defined as the most important for effectively recognising a stressed state of the brain, while many channels (F7, CP5, CP6, T8) could be omitted in the training set to improve the model's performance. The ablation heatmap derived from training the model on a clean variant of SAM-40 resembles that obtained from training the model on raw data, but the importance of individual channels are generally less profound as the model no longer tends to data artifacts and relies on larger clusters of EEG channels to make decisions.



(a) raw EEG based model



(b) preprocessed EEG based model

Figure 2: The importance of each EEG channel in the 10-20 system, calculated by ablating each individual channel from the training data. The ablation studies were carried for two variants of the SAM-40 dataset. Light colours correspond to a positive drop in accuracy, which indicate a channel's significance in driving a distinction between a stressed and a non-stressed state. Dark colours represent a negative drop in accuracy, which suggests the signals coming from that channel merely confuse the model without providing meaningful information.

Human Study Design

A human validation study was conducted, inspired by the standard Turing Test, in which participants interacted with two systems without knowing which one they were communicating with. Participants engaged in conversations with two digital therapy systems and subsequently reported their experiences. The evaluated systems were: A – Mindful-AI, and B – the baseline system.

Each participant interacted with both systems for 10 minutes each, without being informed which system they were communicating with. After completing both sessions, participants completed a survey describing their experiences. To minimise order effects, the sequence of interactions (i.e., which system was experienced first) was randomised uniformly across participants. Participants represent a gender balanced healthy student cohort volunteered to use therapy for exam stress related issues. Exclusion criteria was individuals with diagnosed mental or physical health conditions. Study was conducted under the organisational ethics approval.

Figure 3: Study set-up showing a participant conversing with the system while wearing an EEG headset.

Evaluation Metric

Table 2: Adapted CARE measure items used in this study (Mercer et al. 2004).

Item	Item Text
1	Making you feel at ease (welcoming you; helping you talk about your concerns without feeling judged; starting the conversation gently)
2	Letting you tell your story (letting you explain things in your own words; listening carefully to your situation without interruption)
3	Really listening (paying close attention to what you say; showing interest in your thoughts and how they affect you emotionally)
4	Being interested in you as a whole person (acknowledging different areas of your life, not just the problem you mentioned)
5	Fully understanding your concerns (showing they grasp both what you're saying and what you're feeling about it)
6	Showing care and compassion (being supportive without trying to "fix" you or tell you what to think; validating your experience)
7	Being positive (helping you feel more hopeful or motivated without forcing reassurance)
8	Explaining things clearly (guiding you to understand your patterns or beliefs in a way that made sense to you)
9	Helping you take control (supporting you to come up with your own solutions; encouraging you to take small steps forward)
10	Making a plan of action with you (helping you think through next steps and making it feel doable)

The CARE measure was used to assess participants' perceptions of relational empathy in consultations. The set of items presented to participants in this study are shown in Table 2.

Client–Server System Setup

Both the baseline and Mindful-AI systems were implemented using a client server architecture with socket

based communication. The server hosted the processing pipeline, which included perception, planning, generation, and metacognitive critique agents, and was executed on an NVIDIA A100 (40 GB) GPU.

On the client side, participants used an Emotiv Insight EEG headset connected via the Emotiv Launcher application on a laptop or desktop computer (see Figure 3). Raw EEG data were extracted using the Emotiv SDK and streamed to the server through a socket connection. Text-based conversations between participants and the systems were transmitted through the same communication channel.

Results

Participants' assessments reported significantly higher values for the CARE metric for the proposed system compared with the baseline model across most aspects. Mindful-AI achieved a substantially higher total CARE score per participant (median 42.0 [IQR 2.75]) than the baseline (median 30.0 [IQR 2.75]). Furthermore, several participants commented that the proposed system appeared "more human-like", "felt more attentive", and was "engaging while helping to solve problems". In contrast, the baseline model was described with comments such as "solution-focused", "quite verbose", "always agreeing", and occasionally "too optimistic", which participants felt made it "less human-like" and "less helpful". Given the small sample and ordinal scales, these findings should be interpreted as *preliminary but consistent* with the hypothesis that neurofeedback, strategy-aware generation, and meta-cognitive critic improve perceived empathy and therapeutic stance.

Conclusion and Further Work

Mindful-AI unites neurofeedback, RL planning and metacognitive self-repair. The preliminary results of the simulation suggest the effectiveness of the proposed approach compared to a non perception-based therapy agent. Moreover, the pilot study with human participants suggests the effectiveness of the proposed system over baseline therapy agents in terms of empathy, helpfulness and building trust. Future work will focus on extending this proof of concept into a clinically deployable system. This includes extending the perception pipeline to consider other modalities of physiological data, including heart rate variability and galvanic skin response, to further improve the perception model and consider multiple emotion categories. The meta-cognitive moderator will be enhanced further with annotations from human clinical experts. The pilot study will be extended to large cohorts of human participants, as well as having a human therapy as an additional baseline/ground truth to evaluate the model performance with higher statistical significance.

References

2025. Wysa Assure: The First Insurance-Specific Mental Wellbeing App.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2623–2631. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

and, Y. Q. 2025. Pilot Quasi-Experimental Research on the Effectiveness of the Woebot AI Chatbot for Reducing Mild Depression Symptoms among Athletes. *International Journal of Human–Computer Interaction*, 41(1): 452–459.

Beatty, C.; Malik, T.; Meheli, S.; and Sinha, C. 2022. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Frontiers in Digital Health*, Volume 4 - 2022.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Fleming, C. D., Stephen M. ;Frith. 2014. *The cognitive neuroscience of metacognition*. Springer Nature.

Gou, Z.; Shao, Z.; Gong, Y.; yelong shen; Yang, Y.; Duan, N.; and Chen, W. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *The Twelfth International Conference on Learning Representations*.

Ian Kivlichan, J. E. L. V. M. G., Jeffrey Sorensen; and Culliton., P. 2020. Jigsaw Multilingual Toxic Comment Classification.

Kamińska, D.; Smółka, K.; and Zwoliński, G. 2021. Detection of Mental Stress through EEG Signal in Virtual Reality Environment. *Electronics*, 10(22).

Kaponis, A.; Kaponis, A. A.; and Maragoudakis, M. 2023. Case study analysis of medical and pharmaceutical chatbots in digital marketing and proposal to create a reliable chatbot with summary extraction based on users' keywords. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, 357–363. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700699.

Kim, M.; Lee, H.; Yoo, K. M.; Park, J.; Lee, H.; and Jung, K. 2023. Critic-Guided Decoding for Controlled Text Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4598–4612. Toronto, Canada.

Lee, S.; Kim, S.; Kim, M.; Kang, D.; Yang, D.; Kim, H.; Kang, M.; Jung, D.; Kim, M. H.; Lee, S.; Chung, K.-M.;

Yu, Y.; Lee, D.; and Yeo, J. 2024. Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14245–14274. Miami, Florida, USA: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Mercer, S. W.; Maxwell, M.; Heaney, D.; and Watt, G. C. M. 2004. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, 21(6): 699–705.

Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; and Dormann, N. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *J. Mach. Learn. Res.*, 22(268): 1–8.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1): 36–45.

World Health Organization. 2025. WHO Special Initiative for Mental Health.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35: 24611–24624.

Zhao, J.; Zhang, Z.; Chen, B.; Wang, Z.; Anandkumar, A.; and Tian, Y. 2024. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. arXiv:2403.03507.