NeuroSupport: EEG-Based Evaluation of Mental Health LLMs for Youth in Muslim-Majority Indonesia

Anonymous Author(s)
Faculty of Psychology
Universitas Gadjah Mada
email@example.com

Abstract

Indonesia, the world's largest Muslim-majority country, faces a severe shortage of accessible mental health support. While large language models (LLMs) offer a scalable path to AI-assisted supportive dialogue, current evaluation relies heavily on text-based heuristics or subjective self-reports, failing to capture whether generated responses actively engage the human brain's neural validation and emotional regulation circuits. To address this gap, we introduce *NeuroBench*, a dataset of electroencephalograms (EEG) signals from 60 Indonesian youth (ages 16–24) reading psychologist- and LLM-generated supportive responses to twelve culturally grounded distress vignettes in Bahasa Indonesia. For each response, we record three reward-relevant neural signals, yielding approximately 12,000 response–EEG pairs with full clinician annotations for safety and cultural appropriateness. In the proposed competition, participants are tasked to build models that predict the EEG signals that were actually recorded when responses were read by the recruited youths. Specifically, given the distress vignette, a candidate supportive response, and basic participant demographics, models must forecast three well-established neural signatures of therapeutic engagement: the feedback-related negativity, the late positive potential, and frontal alpha asymmetry, which index valuation, emotional salience, and approach motivation, respectively. The competition will be evaluated using mean squared error and Pearson correlation against the true EEG features on a held-out set of 20% of participants, with an additional bonus track for models that jointly predict the clinician-rated safety and cultural appropriateness scores. By anchoring evaluation in direct neural evidence rather than surface-level text similarity, NeuroBench aims to catalyze the development of LLMs that genuinely resonate with the brain's reward and emotion-regulation systems, pushing the frontier of neurally validated, culturally-informed mental health applications.

1 Introduction

1.1 The ML/AI challenge and its relevance

Large language models (LLMs) now demonstrate performance closely aligned with professional clinical perspectives across mental health support tasks, including depression detection, differential diagnosis, and empathic response generation [4, 5, 7]. However, current evaluation pipelines for mental health LLMs rely almost exclusively on text-based benchmarks, clinician agreement, or self-report preference measures. While useful, these metrics cannot determine whether AI-generated support engages the *neural mechanisms* underlying perceived social reward and emotional validation, processes central to therapeutic benefit [10, 8].

Preprint. Under review.

This creates a critical gap: we have no way of knowing whether a response that *sounds* helpful is actually experienced as rewarding at a neurobiological level. Without such a measure, we cannot reliably align LLMs with the deep affective states that matter most in supportive dialogue.

1.2 Why this problem is impactful

Indonesia offers a strategic context for addressing this gap. As the world's largest Muslim-majority country, it faces a severe mental health workforce shortage: fewer than 1,000 psychiatrists and 451 clinical psychologists for over 270 million citizens [12]. Scalable LLM-powered support tools could help bridge this gap, but their deployment requires more than surface-level evaluation, it requires confidence that AI-generated responses actually engage the reward circuitry that human support activates. A neural benchmark would provide that confidence, enabling policy and clinical deployment grounded in neurobiology, not just text metrics.

1.3 Why an EEG-grounded competition task is the right approach

We address this gap through *NeuroBench*, a competition benchmark built on a novel **paired** EEG–dialogue dataset. The same twelve distress vignettes are presented to $N = 60$ Indonesian youth alongside responses drawn from two sources: licensed clinical psychologists and state-of-the-art LLMs. Continuous 64-channel EEG is recorded for every response, capturing three reward-relevant signals: frontal alpha asymmetry, late positive potential, and reward positivity. This paired design produces a neurobiological reference standard against which any LLM can be directly compared.

The competition participants are then tasked to develop models that can predict the three reward-relevant neural signals (frontal alpha asymmetry, late positive potential, and reward positivity) evoked by a given response, using only the distress vignette, the response text, and basic participant demographics as input.

2 Dataset, Reward Model, & Ethical Sourcing

Source: The primary dataset consists of: (1) twelve standardized emotional vignettes in Bahasa Indonesia, describing ecologically common distressing experiences for Indonesian university youth (academic failure, family conflict, social rejection, loneliness, future uncertainty) that would preserve emotional authenticity and ecological validity during neural reward processing; (2) matched supportive responses for each vignette generated by both the multi-agent LLM system and two licensed clinical psychologists; and (3) continuous 64-channel EEG recordings from $N = 60$ participants alongside trial-level ratings based on the visual analog scale (VAS). All participant data will be collected at Universitas Gadjah Mada (UGM) under full institutional ethics approval.

Representation: Participants will be recruited across faculties to represent diverse regional, linguistic, and socioeconomic backgrounds within the Indonesian university population. Vignettes will be piloted on 10 participants and reviewed by two licensed Indonesian clinical psychologists to confirm ecological validity and cultural appropriateness across contexts.

Bias Mitigation: Vignettes will be matched for valence (negative), arousal level, and word count, and reviewed to avoid content that presupposes specific religious, ethnic, or socioeconomic circumstances. Stimuli will additionally be reviewed for cultural appropriateness across Indonesia's religiously and ethnolinguistically diverse communities, including sensitivity to varying norms around emotional expression and help-seeking. Psychologist responses will be randomized across participants to prevent psychologist-specific confounds. All stimulus materials will be reviewed for gender-neutral language.

Privacy: No personally identifiable information will be stored alongside EEG or response data. Participant IDs will be randomly generated and stored separately from all physiological and behavioral data. Participants will provide written informed consent in Bahasa Indonesia covering all data uses. Data will be stored on secured servers.

Quality Assurance:

- **Human Evaluation.** Both vignette appropriateness and psychologist response quality will be independently rated by two licensed clinical psychologists on dimensions of empathic accuracy, clinical appropriateness, and cultural sensitivity. Inter-rater agreement will be quantified using Cohen's κ (target $\kappa \geq 0.70$). Trial-level RewP and LPP scoring will be verified by two independent EEG analysts blind to condition.
- **Conflict Resolution.** Rating disagreements will be resolved through adjudication discussion between raters and a senior researcher, with final decisions documented.

3 Evaluation

Data splits. The dataset comprises approximately 12,000 response–EEG pairs collected from 60 participants. To ensure strict participant-level independence, we split the data into training (48 participants, ≈9,600 pairs) and test (12 participants, ≈2,400 pairs) sets. The test set remains *completely unseen* during the competition; only the training set features and response texts are released, while withheld test labels are used for leaderboard evaluation. Both LLM-generated and clinician-written responses are mixed in the test set, and the source of each response (LLM or clinician) is not provided, reflecting a realistic blind evaluation scenario. Participants may use any cross-validation strategy on the training set, but the final rankings are determined solely on the fixed test split.

Primary track metrics: Given a distress vignette, a candidate supportive response, and basic participant demographics, models must predict the three reward-relevant neural signals: frontal alpha asymmetry (FAA), late positive potential (LPP), and reward positivity (RewP).

Accuracy of the predicted signals will be measured using mean squared error (MSE) and Pearson's correlation. The final leaderboard ranks submissions by a combined score that normalizes and aggregates both metrics.

Bonus mark metrics: Participants may submit models that additionally predict the measured (1) safety, and (2) cultural appropriateness, as rated by clinicians. These two auxiliary targets are predicted jointly with the EEG signals.

Evaluation mirrors the primary track, but the combined score now includes the average MSE and Pearson correlation for the two clinical scores, weighted equally with the neural signals

Submission and winner determination: Each team may submit predictions for the entire test set (EEG signals only for the primary track, or all five outputs for the bonus track) up to a fixed number of times per week.

A public leaderboard displays scores on a random 30% subset of the test pairs to allow continuous feedback; the final ranking is based on the private, held-out 70% remainder.

The winner of each track is the team with the highest private score at the competition deadline.

To encourage scientific insight, top-performing teams will be invited to present their methods, and all submissions must include a brief technical report describing the model architecture, training procedure, and any use of external data.

4 Hosting platform

We will host the competition on Kaggle, which provides a mature infrastructure for multi-output regression tasks with custom evaluation metrics, an active global community of machine-learning practitioners, and built-in public/private leaderboard management.

The dataset is well within Kaggle's storage limits, and the platform has a strong track record of hosting medically oriented benchmarks under clear data usage agreements.

To promote reproducibility and discourage overfitting to the public leaderboard, we will run the final evaluation phase as a code competition, requiring participants to submit inference code that runs on the held-out test set within a fixed compute budget.

EEG data sharing will comply with Indonesian Ministerial Regulation (Permenkes No. 11/2017) on health data. Participant consent will explicitly cover anonymized data sharing on open-access platforms. No personal identifable data will be included in released files.

5 Potential Positive Impact

Fairness and Inclusivity. This study directly addresses equity in mental health care access. By testing whether AI tools are neurobiologically viable for populations with minimal access to professional services, it contributes evidence-based guidance for deploying AI equitably in low- and middle-income countries (LMICs), rather than reserving AI mental health innovation for high-income contexts.

Anticipated Societal Benefits. Indonesia's national digital health strategy (Strategi Nasional Kesehatan Digital Indonesia 2021–2024) explicitly commits to expanding digital mental health services. Findings will directly inform whether AI mental health tools can serve as neurobiologically acceptable first-line support, shaping national policy on AI-assisted care. Positive findings would support integration of LLM tools into the *Puskesmas* network, extending reach to rural and underserved populations.

Although evaluated in Indonesia, the proposed framework is intended to support future replication across other multilingual and Muslim-majority contexts where access to mental health professionals remains limited.

Scientific and Community Impact. The study produces: (1) the first neuroscientific comparison of AI versus human mental health support; (2) a culturally adapted EEG paradigm for Indonesian youth, a population absent from existing neuropsychological normative databases; (3) an open, FAIR-compliant dataset for the global research community; and (4) a validated, replicable methodology for LMIC-contextualized AI mental health research. This advances responsible AI development by grounding deployment decisions in neural evidence rather than behavioral proxies alone. More broadly, the framework provides a template for integrating neurophysiological signals into evaluation pipelines for socially interactive AI systems.

6 Proposed Study Timeline

The dataset, codebase, and materials will be finalized prior to data collection. The following milestones commit to delivering a complete dataset and technical report within 12 months of project initiation.

Month	Milestone	Details
1–2	Ethics & Stimuli	Submit ethics application (UGM Research Ethics Committee); develop and pilot 12 vignettes in Bahasa Indonesia; recruit 2 licensed psychologists; conduct inter-rater agreement pilot.
3	Setup & Recruitment	Configure 64-channel BioSemi EEG system; train research assistants; begin participant recruitment; pre-register study on OSF.
4–7	Data Collection	Collect EEG data ($N = 60$; 2 sessions per participant); administer PHQ-9, BIS/BAS, PANAS, AI-Experience questionnaires; provide post-session debriefing and mental health resources.
8–9	Pre-processing & Analysis	EEG artifact rejection (ICA); epoch extraction; RewP and LPP scoring; run primary LMMs and secondary analyses in R; compute effect sizes and moderation tests.
10–12	Dissemination	Manuscript preparation; deposit open dataset on OSF (BIDS format) for future research purposes; participants of winning entries are invited to code release on GitHub. We will submit to target journals, prepare policy brief for Indonesian Ministry of Health, etc.

Table 1: Twelve-month project timeline.

7 Plan for Publication

- **Primary Paper.** A full empirical report presenting the EEG paradigm, RewP/LPP results, self-report findings, and moderation analyses will be submitted to a high-impact peer-reviewed journal in clinical neuroscience or digital mental health (target: *NPJ Digital Medicine*, *Biological Psychology*, or *JMIR Mental Health*).
- **Methods Paper.** A shorter methods report documenting the culturally adapted Island Getaway paradigm for Indonesian samples and the associated Bahasa Indonesia vignette stimuli will be submitted as an open-access methods resource (target: *Behavior Research Methods*).
- **Policy Brief.** A non-technical summary of findings will be prepared for submission to Indonesia's Ministry of Health and relevant digital health policy bodies, contextualizing the results within the national digital health strategy.
- **Open Science.** Pre-registration, anonymized EEG dataset (BIDS), stimulus materials, inter-rater agreement data, and all analysis code will be made publicly available on OSF and GitHub prior to or concurrent with manuscript submission, in accordance with open-science best practices and FAIR data principles.

Acknowledgements

Not included in the anonymized submission. Will be added in the final version upon acceptance.

References

[1] Brohan, E., Chowdhary, N., Dua, T., Barbui, C., Thornicroft, G., & Kestel, D. (2024). The WHO Mental Health Gap Action Programme for mental, neurological, and substance use conditions. *Lancet Psychiatry*, *11*(2), 155–158.

[2] Cardamone, N., Olfson, M., Schmutte, T., Kulkarni, P. V., Liu, T., Cullen, S., Williams, N. J., & Marcus, S. (2024). Classifying unstructured text in electronic health records for mental health prediction models. *JMIR Medical Informatics*, *13*, e65454.

[3] Clayson, P., Carbine, K. A., & Larson, M. (2019). A registered report of error-related negativity and reward positivity as biomarkers of depression. *International Journal of Psychophysiology*.

[4] Guo, Z., Lai, A., Thygesen, J., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, *11*, e57400.

[5] Jin, Y., Liu, J., Li, P., Wang, B., Yan, Y., Zhang, H., Ni, C., Wang, J., Li, Y., Bu, Y., & Wang, Y. (2025). The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, *27*, e69284.

[6] Kujawa, A., Proudfit, G. H., & Klein, D. N. (2017). Neural response to social reward in daughters of depressed mothers. *Social Cognitive and Affective Neuroscience*, *12*(7), 1120–1130.

[7] Levkovich, I. (2025). Evaluating diagnostic accuracy and treatment efficacy in mental health: A comparative analysis of large language model tools and mental health professionals. *European Journal of Investigation in Health, Psychology and Education*, *15*, e15010009.

[8] Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*, 449–459.

[9] Rappaport, B. I., Pagliaccio, D., Pine, D. S., Klein, D. N., & Jarcho, J. M. (2019). Peer victimization and dysfunctional reward processing: ERP and behavioral responses to social and monetary rewards. *Frontiers in Behavioral Neuroscience*, *13*, 120.

[10] Rogers, C. R. (1951). *Client-centered therapy*. Houghton Mifflin.

[11] Song, I., Pendse, S. R., Kumar, N., & de Choudhury, M. (2024). The typing cure: Experiences with large language model chatbots for mental health support. *Proceedings of the ACM on Human-Computer Interaction*, *9*, 1–29.

[12] The Conversation. (2019, October 31). 260 million people and less than 1,000 psychiatrists: Indonesia's mental health worker shortage. `https://theconversation.com/260-million-people-and-less-than-1000-psychiatrists-indonesias-mental-health-worker-shortag`

A Analysis plan

Primary Metric: RewP amplitude (mean μV, 250–350 ms post-response onset) at electrode Cz, quantified via linear mixed-effects models (LMMs) with C, the Condition (AI vs. Psychologist) as the fixed effect and participant as a random effect. Effect size reported as Cohen's d. In the proposed framework, RewP amplitude functions as a biologically grounded alignment-sensitive metric reflecting neural responsiveness to supportive AI interactions.

Secondary Metrics:

- LPP amplitude (mean μV, 400–800 ms) across centroparietal sites (Cz, CPz, Pz), analyzed with the same LMM structure.
- Self-reported perceived helpfulness and understanding (0–100 VAS), compared via paired-samples t-tests (or Wilcoxon signed-rank where normality is violated).
- Pearson correlation between RewP amplitude difference scores and VAS difference scores across conditions.

Exploratory Metrics: Moderation of the RewP condition difference by PHQ-9 depression severity and prior AI chatbot exposure, tested via Condition $\times$ Moderator interaction terms in LMMs with Bonferroni correction ($\alpha = 0.025$).

Justifications: The RewP is the most well-validated electrocortical index of reward responsiveness and has been directly linked to social reward processing and depression vulnerability across multiple studies [6, 9, 3]. LMMs are appropriate for within-subject ERP data with missing trials due to artifact rejection. VAS measures provide ecological face validity for subjective therapeutic experience. The combination of neural and subjective metrics enables convergent validity testing.