Supplementary Information

Adaptive stress-aware recommendation of supportive engagement activities for digital mental health: a contextual bandit approach with domain-specific language models

Makaeva Ziinat, Muhammad Ahmed, Md Ariful Islam Mozumder, Sabeeha Jabeen, Hee-Cheol Kim

Supplementary Tables

Supplementary Table S1: Controlled emotion-encoder comparison with 95% confidence intervals (t-distribution across 5 seeds).

Model	Macro F1 [95% CI]	Weighted F1 [95% CI]
BERT-base	0.565 ± 0.012 [0.548, 0.582]	0.746 ± 0.009 [0.733, 0.759]
RoBERTa-base	0.568 ± 0.010 [0.555, 0.582]	0.756 ± 0.013 [0.739, 0.774]
MentalRoBERTa	0.571 ± 0.011 [0.556, 0.586]	0.755 ± 0.013 [0.737, 0.773]

Supplementary Table S2: Per-class emotion classification (MentalRoBERTa), pooled across 5 seeds on the GoEmotions test set ($5 \times 4{,}590 = 22{,}950$ predictions), reconstructed from the averaged confusion matrix. In the weighted-average row, precision and recall are computed from the pooled predictions, whereas F1 is the mean ± SD across seeds.

Class	Precision	Recall	F1	Support (pooled)
Anxiety	0.312	0.777	0.445	870
Burnout	0.377	0.633	0.473	1,910
Sadness	0.515	0.781	0.620	740
Overwhelm	0.383	0.645	0.481	1,625
Neutral	0.955	0.735	0.831	17,805
Weighted avg	0.828	0.723	0.755 ± 0.013	22,950
Macro avg	–	–	0.571 ± 0.011	–

Supplementary Table S3: Temperature scaling: mean fitted temperature and expected calibration error (ECE) before and after, 5 seeds. Predictions and F1 are unchanged by construction, which was verified for every seed and model. Values come from an independent rerun of the training protocol, in which the uncalibrated ECE was 0.083–0.109 (0.091–0.103 in the main run); the ECE ranking of encoders was not stable across runs.

Model	Mean T	ECE before	ECE after	Relative reduction
BERT-base	1.40	0.104	0.068	−34.5%
RoBERTa-base	1.43	0.109	0.062	−42.7%
MentalRoBERTa	1.35	0.083	0.052	−37.2%

Supplementary Table S4: LLM-as-judge evaluation (LLaMA-3.3-70B) on the stressed subset ($N = 80$; 1–5 scale; one-sample t-tests against 3.0). Over all 100 posts, including 20 neutral, the emotion-detection score was 3.87, against 2.81 for the K-means alternative.

Metric	Score	95% CI	p
Emotion-detection accuracy	4.34	[4.10, 4.57]	< 0.001
Activity appropriateness	4.14	[3.94, 4.34]	< 0.001

Activity score by emotion: anxiety 5.00, sadness 4.45, burnout 4.05, overwhelm 3.05. Supervised classifier vs. K-means: +1.06 emotion detection, +0.25 activity appropriateness.

Supplementary Table S5: End-to-end error propagation (shared LinUCB, 20 seeds, 30 users, 100 sessions per user). Paired t-tests against ground truth.

Emotion observed by the bandit	Mean reward	SD	paired p
Ground truth	0.360	0.056	–
Predicted (MentalRoBERTa confusion matrix)	0.096	0.044	< 0.001
Random label	−0.366	0.017	< 0.001

Supplementary Table S6: Stress-routing threshold sensitivity (MentalRoBERTa, pooled across 5 seeds, $N = 715$ per seed). Metrics refer to the stress class. The F1-maximising threshold (0.5) differed from the operating threshold (0.6) by 0.007 F1 ($p = 0.047$, $n = 5$).

Threshold	Precision	Recall	F1	% routed
0.3	0.770	0.898	0.829	60.2
0.4	0.782	0.883	0.830	58.3
0.5	0.795	0.871	0.831	56.6
0.6 (operating)	0.800	0.850	0.824	54.9
0.7	0.807	0.830	0.819	53.1
0.8	0.819	0.802	0.811	50.5

Supplementary Table S7: Characteristics of the external validation participants ($N = 70$).

Characteristic	Category	n (%)
Age group	18–24	30 (42.9)
	25–34	36 (51.4)
	35–44	3 (4.3)
	Not reported	1 (1.4)
Gender	Female	59 (84.3)
	Male	11 (15.7)
Current status	Student	55 (78.6)
	Employed	7 (10.0)
	Student and employed	5 (7.1)
	Other	2 (2.9)
	Prefer not to say	1 (1.4)
DASS-21 Stress (doubled score)	Normal (≤ 14)	27 (38.6)
	Mild or above (> 14)	43 (61.4)

Supplementary Table S8: Personalisation configurations (master protocol). "Init." gives the initialisation of the personalised and shared bandits (warm: cloned from the population prior; cold: started from scratch). CRN: common random numbers.

Row	Configuration	Sessions/user	Seeds	Init. (pers./shared)	CRN	Status
1	Original, 30 users	100	5	warm / warm	no	Null result
2	Corrected, 30 users	100	5 + 20	cold / warm	yes	Confirmatory
3	Sanity check, 30 users	100	5	cold / cold	yes	$\Delta = +0.219$, $p < 0.001$

Supplementary Notes

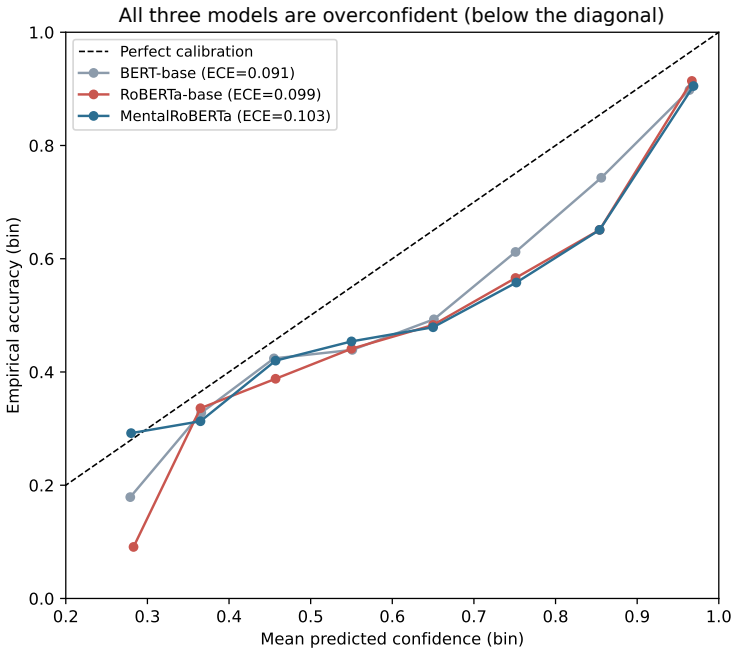
Supplementary Note 1: emotion-encoding alternatives. *Zero-shot LLaMA-3.2-3B.* The model had been QLoRA-tuned on MentalChat16K for conversational mental-health support but was never trained on our labels or taxonomy. On a small set of qualitative test cases it reached about 70% stress accuracy and 58% emotion accuracy and often produced malformed structured output. These estimates come from an informal evaluation set, not from the multi-seed benchmark protocol, so they are indicative only and do not constitute a controlled ablation. *K-means clustering.* Five clusters were formed over nine LIWC features plus the stress probability. A random forest recovered the cluster assignments with 86% leave-one-out accuracy, which checks only cluster separability. Keyword-based emotion validation reached 28%, indicating that the clusters encode stress intensity rather than emotion type; LIWC features are also unavailable at inference time for free text.

Supplementary Note 2: reward generator. The generator draws the click first and sets playtime to zero when there is no click (Methods). Assertions across a grid of random draws verified that a missing click always yields $R = -0.7$, that the early-exit and sustained-play playtime ranges never overlap, and that $R \in \{-0.7\} \cup [0.3, 1.0]$. All reported results use this generator.

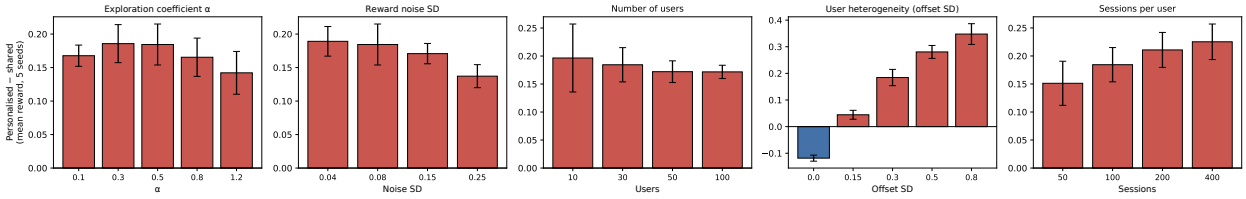
Supplementary Note 3: rule-based policy. The rule-based policy does not observe reward, update parameters, use per-user information or history, or explore; it always selects the first-listed activity for the detected emotion. For four of five emotions this coincides with the population-optimal activity; only for sadness does it differ (music therapy instead of positive memory). This makes the rule-based policy a strong baseline; a weaker fixed activity would have inflated LinUCB's relative advantage.

Supplementary Note 4: why both personalisation configurations are reported. The original configuration tests whether an architecture with a shared prior shows measurable personalisation; the corrected configuration tests whether LinUCB can learn persistent per-user heterogeneity when not constrained by a dominant shared prior. They differ in two respects, the initialisation of the personalised bandit and the use of common random numbers; the sanity check (Supplementary Table S8, row 3) holds initialisation fixed but does not remove the second difference. Random and population-best are lower in the personalisation experiments than in the population-level experiment (analytically 0.204 and 0.324 vs. 0.233 and 0.426) because user offsets are added to $\mu_{e,a}$ before clipping to $[0, 1]$, so that for activities with $\mu_{e,a}$ near 1, upward offsets are truncated while downward ones are not.

Supplementary Figures



Supplementary Figure S1: Reliability diagram pooled across 5 seeds. All three encoders lie below the diagonal in the 0.5–0.9 confidence range, indicating overconfidence.



Supplementary Figure S2: One-at-a-time sensitivity analysis of the personalisation gain (personalised minus shared mean reward, 5 seeds) around the corrected configuration (baseline: $\alpha = 0.5$, noise SD 0.08, 30 users, offset SD 0.30, 100 sessions per user). Red bars, significant positive gain; blue, significant negative gain. The gain is negative only when users are identical (offset SD 0).