npj | digital public health

Article

https://doi.org/10.1038/s44482-026-00049-z

Large language models rate human happiness higher than humans do

Check for updates

Le Fang[1] ✉, Tianyi Zhang[1], Lina Christin Brockmeier[1,2], Shiquan Zhang[1], Runqiu Fei[1], Ying Ma[1], Songyan Teng[1], Simon D'Alfonso[1], Peter Koval[1], Hong Jia[3], Jorge Goncalves[1] & Vassilis Kostakos[1]

Feelings are central to wellbeing, but measuring them remains difficult: surveys provide reliable data but are burdensome. LLMs introduce a provocative alternative: acting as proxies that articulate emotions and experiences on behalf of individuals. However, little work has systematically evaluated the validity of such an approach, and this is our contribution. Here, we use longitudinal smartphone data to personalise an LLM (Gemini-2.5-flash-lite) and approximate human responses to an affective survey. Using 17 weeks of multimodal sensing and questionnaires from 83 participants, we evaluate how well an LLM can predict affect across different time ranges, affect types, and individuals. Our findings reveal a systematic positivity bias: the LLM predicts people as happier than they actually report, while downplaying their negative feelings. This work highlights the importance of rigorous evaluation and affective-science grounding when using LLMs to infer human feelings from digital traces.

Humans automatically associate daily events with their feelings, such as inspired and ashamed[1,2]. These momentary feelings are central to mental wellbeing: they reflect environmental demands, guide behaviour, and motivate self-regulation[3,4]. However, measuring affects remains challenging. Surveys can provide reliable and detailed data, but they burden users and fail to support real-time feedback[5–9]. On the other hand, passive smartphone sensing offers continuous behavioural signals, but it is difficult to translate these signals into subjective experience[10]. Prior work has explored predicting affective states from such passively sensed data using machine learning with promising results, but these approaches rely on extensive, hand-crafted feature engineering[11,12].

Large language models (LLMs) provide a new possibility: by combining survey responses[13] with rich digital traces (e.g., smartphone-sensor data), LLMs could simulate or infer individuals' affects[14,15]. This capability has the potential for helping users reflect on their feelings without the need to manually answer affective questionnaires. Given LLMs' success in reasoning over natural language and contextual information, it is tempting to assume that sufficiently rich personal data might allow them to approximate how people feel in everyday life. However, little is known about how well LLMs can predict human affects[16], or whether such predictions are systematically biased[17].

In this paper, we present the first large-scale study using longitudinal smartphone data to personalise an LLM (Gemini-2.5-flash-lite) and evaluate its ability to simulate human responses to affective surveys (e.g., the Positive and Negative Affect Schedule; PANAS). We obtained a dataset spanning 17 weeks of passive sensing (including app usage, keyboard activity, and on-screen text) alongside weekly affect questionnaires as ground truth. We used this dataset in a controlled experimental design, systematically varying the volume and temporal span of data used to personalise the LLM and assessing predictive performance across 83 participants and 17 weeks.

Our results show that simply providing larger volumes or longer histories of smartphone data yields only marginal gains. Instead, predictive performance varies more strongly across affect types, time ranges, and participants. Most strikingly, we find that LLMs systematically predict participants as happier than they actually report: positive affects such as alert and determined are inflated, while negative affects such as nervous or upset are downplayed or missed. This systematic positivity bias highlights both the promise and the risk of LLM-based human simulation: although such systems can approximate affects, they may distort the lived reality of people's daily affective experiences by smoothing over distress.

This study identifies systematic biases in large language models' inference of human affect from digital traces, most notably a tendency to predict participants as happier than they report. This contribution is grounded in a systematic evaluation of predictive performance across affect types and varying volumes and temporal spans of personal data.

These findings caution against uncritical use of LLMs to infer human affect from digital traces: optimism bias may make systems appear more

[1]School of Computing and Information Systems, University of Melbourne, Melbourne, VIC, Australia. [2]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany. [3]School of Computer Science, University of Auckland, Auckland, New Zealand. ✉e-mail: le.fang2@unimelb.edu.au

reassuring than accurate, potentially overlooking distress. At the same time, they point to the need for methods that ground passive sensing data in affective science, so that models move beyond surface-level interpretation of digital traces and more faithfully capture the complexity of human emotional experience. By confronting this bias, future systems could more safely and effectively support reflection, self-awareness, and wellbeing.

Results

All results use the zero-shot prompting setup (see Supplementary Note 1: Prompts), with no in-context examples or task-specific fine-tuning. We evaluate whether a large language model can infer participants' affective states from longitudinal smartphone-derived diaries using a consistent prompting setup.

We evaluated up to 17 weekly I-PANAS-SF surveys per participant (N = 83) across seven Time Ranges. In total, we obtained 4949 affect prediction files. Across these evaluations, the model processed 63.53M input tokens over 5384 requests (API call error rate 0.09%). End-to-end processing took ~30 hours, at a total cost of USD $19.09, using the Google Gemini-2.5-flash-lite model.

Table 1 | Time-range–specific performance in predicting individual affects (number of prediction files in parentheses)

Time Range	Accuracy (%)	RMSE	Kendall's τ_b
Mode(1,302)	36.1[33.8, 38.4]	1.16[1.11, 1.23]	—
First Week(1,219)	39.7[37.3, 42.4]	1.08[1.03, 1.13]	0.48[0.43, 0.52]
Previous Week(1,219)	**49.1**[46.6, 51.4]	**0.92**[0.88, 0.97]	**0.57**[0.53, 0.61]
7d(1,096)	30.8 [28.5, 32.9]	1.43 [1.37, 1.49]	0.29 [0.24, 0.33]
3d(1,009)	29.6[27.5, 31.9]	1.46[1.39, 1.53]	0.27[0.22, 0.32]
24h(726)	28.8[26.2, 31.4]	1.47[1.39, 1.55]	0.26[0.21, 0.32]
12h(588)	30.2[27.3, 33.2]	1.47[1.38, 1.56]	0.26[0.19, 0.32]
6h(552)	29.0[25.8, 32.1]	1.48[1.40, 1.58]	0.25[0.19, 0.31]
3h(513)	27.8[24.5, 30.8]	1.51[1.43, 1.60]	0.24[0.18, 0.30]
1h(465)	28.3[25.2, 31.6]	1.55[1.45, 1.67]	0.22[0.14, 0.28]

Accuracy and RMSE are means of per-file metrics, where one file is a single prediction across the 10 affects for one participant, week, and time range; Kendall's τ_b pools predicted and ground-truth scores across all 10 affects within each time range or baseline. 95% cluster-bootstrap CIs (participants resampled with replacement, 1,000 iterations) are shown beneath each value. The Mode baseline predicts the most frequent Likert rating per affect (its τ_b is undefined, "—"); the First Week and Previous Week baselines use each participant's own earlier responses (counted from each participant's second week onward, as the first week has no earlier reference). **Bold**: best overall; underline : best time range excluding baselines.

The main parameter we manipulate in our analysis is the Time Range, which defines how much data we use to prime the LLM before prediction. In our analysis, the Time Range varies from 1 hour to 7 days. Due to the nature of the dataset, a longer Time Range does not necessarily mean more data. In one week, a participant may not use their phone a lot, while in another week, they can have a lot of concentrated usage in a period of 12 hours.

We first assess overall predictive performance. We then examine how the model's predictions differ from participants' self-reported affect when errors occur. Finally, we explore whether these patterns vary across affects and participants.

Our main findings are threefold: (1) extending the Time Range (i.e., providing more hourly windows per diary) did not consistently improve the LLM's predictions of participants' affects; (2) the LLM exhibits a positivity bias when predicting overall feeling, overestimating participants' happiness and suppressing their sadness; and (3) it predicts negative affects more reliably than positive affects, with substantial variability across affects and participants.

Overall predictive performance

We begin by evaluating the overall predictive performance of the LLM on affect inference tasks, using both individual affect prediction and overall feeling classification as outcome measures.

Table 1 summarises performance for predicting the ten individual affects in the I-PANAS-SF survey across all Time Ranges. Overall accuracy remains low, ranging between 27.8% and 30.8% across Time Ranges, with RMSE values exceeding 1.4. All Time-Range-based predictions underperform simple baselines that leverage participants' own historical reports, including the Previous Week baseline, which achieves substantially higher accuracy (49.1%) and lower error. The strength of these history-based baselines is expected: within-person affect is highly stable over time (emotional inertia), so a participant's recent self-reports are strongly predictive of their next one[18,19]. The ordinal metric leads to the same conclusion: pooled Kendall's τ_b for the LLM ranges from 0.22 (1h) to 0.29 (7d), well below the First Week (0.48) and Previous Week (0.57) baselines, indicating that the LLM's weak performance is not an artefact of exact-match scoring or of treating ordinal Likert ratings as continuous. Although τ_b increases slightly with longer Time Ranges, the CIs overlap substantially across Time Ranges. Together, these results indicate that the LLM struggles to reliably infer fine-grained affective states from behavioural diaries, even when provided with extended temporal context.

We observe similar limitations when predicting participants' overall feeling, operationalised as affect balance[20,21] (pleasant vs. unpleasant). As shown in Table 2, accuracy and F1 scores appear high across Time

Table 2 | Time-range–specific performance for predicting overall feeling (pleasant vs. unpleasant), operationalised as affect balance (the sum of positive affect scores minus the sum of negative affect scores), with zero differences classified as pleasant

Time Range	Acc	Prec	Recall	F1	Spec	MCC
Mode	0.81[0.76, 0.86]	0.81[0.76, 0.86]	**1.00**[1.00, 1.00]	0.90[0.86, 0.92]	0.00[0.00, 0.00]	0.00[0.00, 0.00]
First Week	0.80[0.74, 0.85]	0.85[0.80, 0.89]	0.91[0.85, 0.96]	0.88[0.84, 0.91]	0.33[0.14, 0.51]	0.28[0.10, 0.44]
Previous Week	0.83[0.79, 0.86]	**0.89**[0.86, 0.92]	0.90[0.87, 0.92]	0.89[0.86, 0.92]	**0.55**[0.44, 0.63]	**0.44**[0.34, 0.53]
7d	0.83[0.78, 0.87]	0.83[0.78, 0.87]	1.00[0.99, 1.00]	0.90[0.87, 0.93]	0.08 [0.00, 0.23]	0.23 [-0.03, 0.42]
3d	0.83[0.78, 0.87]	0.83[0.78, 0.87]	1.00[0.99, 1.00]	0.90[0.88, 0.93]	0.04[0.00, 0.13]	0.16[-0.03, 0.31]
24h	0.82[0.77, 0.87]	0.82[0.77, 0.87]	1.00[0.99, 1.00]	0.90[0.87, 0.93]	0.04[0.00, 0.10]	0.14[-0.03, 0.27]
12h	0.82[0.75, 0.87]	0.82[0.75, 0.87]	1.00[0.99, 1.00]	0.90[0.86, 0.93]	0.02[0.00, 0.05]	0.09[-0.03, 0.19]
6h	0.83[0.77, 0.87]	0.83[0.78, 0.88]	1.00[0.99, 1.00]	0.91[0.87, 0.93]	0.00[0.00, 0.00]	-0.03[-0.04, 0.00]
3h	0.83[0.78, 0.88]	0.83 [0.78, 0.88]	1.00[1.00, 1.00]	0.91[0.87, 0.94]	0.00[0.00, 0.00]	0.00[0.00, 0.00]
1h	0.83[0.76, 0.89]	0.83[0.76, 0.89]	0.99[0.99, 1.00]	0.91[0.87, 0.94]	0.00[0.00, 0.00]	-0.03[-0.05, 0.00]

95% cluster-bootstrap CIs (participants resampled with replacement, 1,000 iterations) are shown beneath each value. The Mode baseline predicts, for each affect, the globally most frequent Likert value across all participants. The First Week and Previous Week baselines use participant-specific historical data. **Bold** indicates the best overall metric across all rows; underline denotes the best-performing time range (excluding all baselines). Ties are resolved using unrounded metric values.

Ranges; however, these metrics are driven by near-perfect recall and extremely low specificity, resulting in Matthews Correlation Coefficients close to zero or negative. This pattern suggests that the model's predictions are unevenly distributed across outcome classes, limiting its discriminative ability.

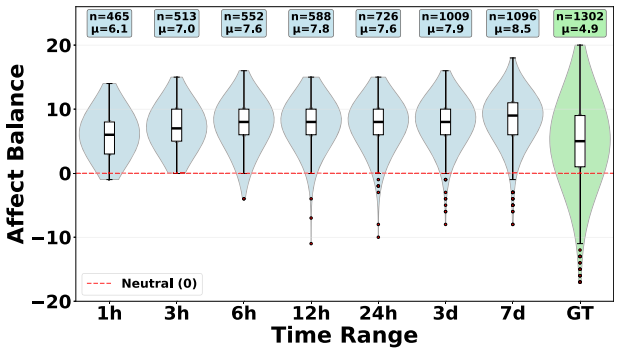


Fig. 1 | Distribution of affect balance (sum of positive minus negative affect scores) across 7 Time Ranges compared to the ground truth (GT; rightmost violin). The LLM's predictions consistently aggregate to affect balances above zero, indicating a systematic bias to interpret participants as *happy* (positive feeling). The ground truth distribution (green) shows a lower mean ($\mu = 4.9$), highlighting that the LLM systematically overestimates participants' positive feeling or underestimates their negative feeling compared to their self-reports.

Across Time Ranges, performance differences are modest, with no consistent improvement as the temporal span increases. In addition, prediction accuracy and RMSE show only weak associations with token usage (see Supplementary Figs. 1 and 2).

These aggregate performance metrics alone do not explain how or why prediction errors occur. To gain further insight, we next examine the structure of prediction errors by comparing predicted and self-reported affective states.

Positivity bias in affect predictions

To better understand how the model's predictions deviate from participants' self-reported affect, we analyse the direction and structure of prediction errors beyond aggregate accuracy metrics. Rather than asking whether predictions are correct, this analysis focuses on how predicted affective states differ from ground truth when discrepancies occur.

Figure 1 compares the distributions of predicted and ground-truth affect balance, defined as the sum of positive minus negative affect scores, across all seven Time Ranges. Across conditions, predicted affect balances are consistently shifted toward positive values relative to the ground truth. This pattern holds regardless of how much user history is included, indicating that prediction errors are not randomly distributed but instead skew toward pleasant affective states.

To further characterise this skew, we examine the joint distribution of predicted and ground-truth affect balance (Fig. 2). A pronounced asymmetry emerges: when participants reported a negative affect balance (unpleasant overall feeling), the model frequently predicted a positive affect balance (pleasant). In contrast, cases in which participants reported a

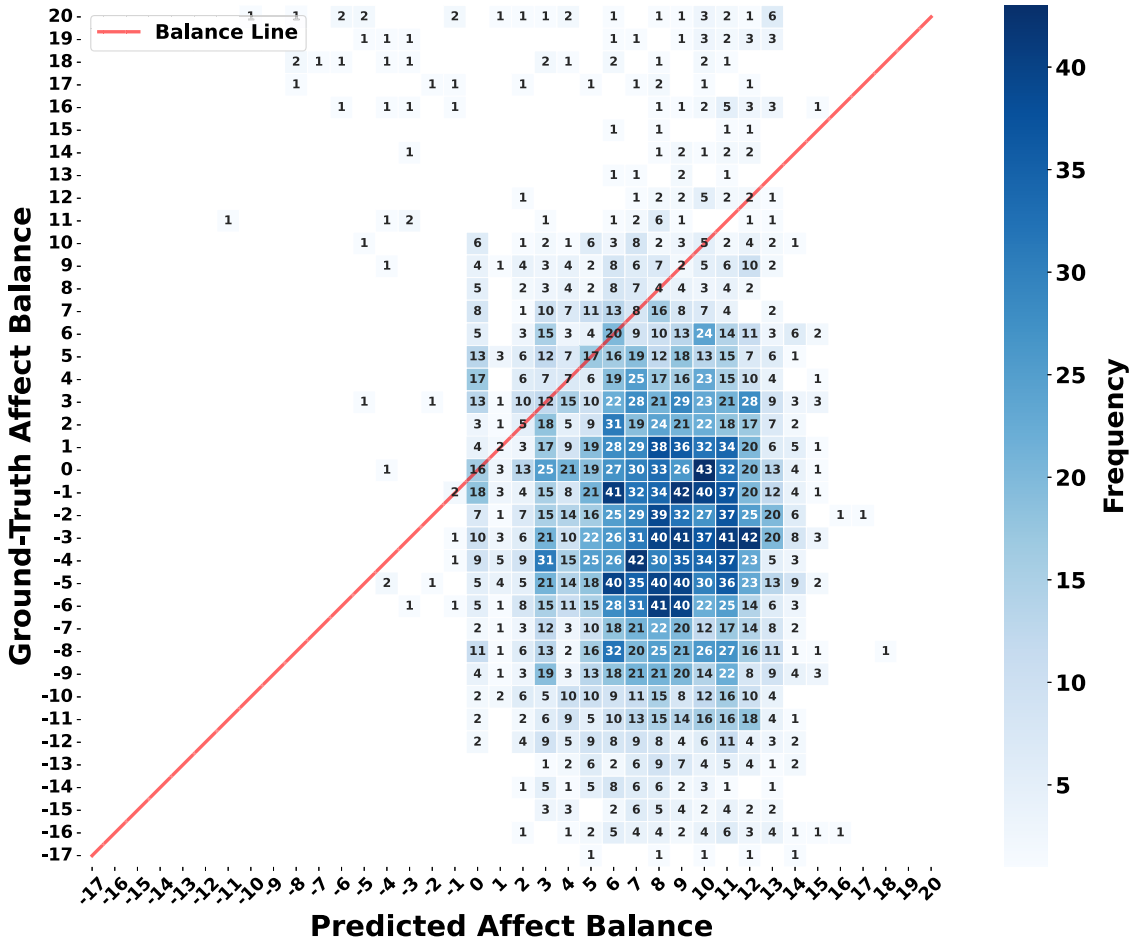
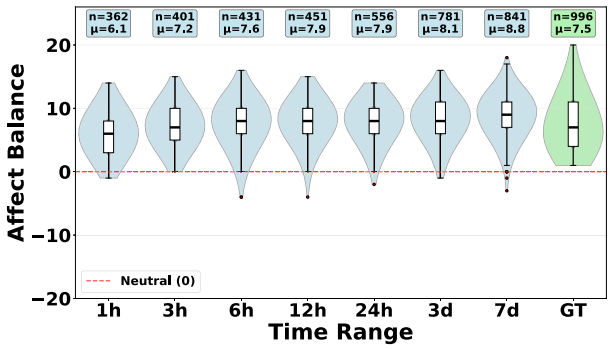
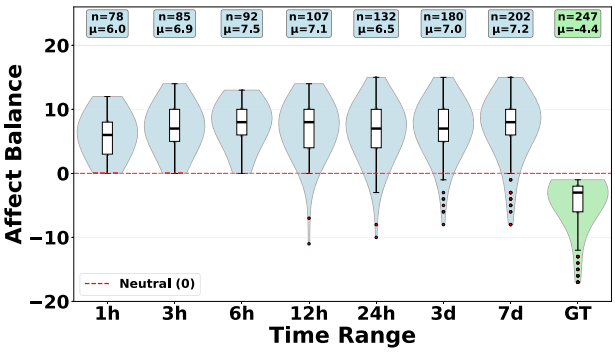


Fig. 2 | Ground-truth versus LLM-predicted affect balances (positive minus negative I-PANAS-SF scores). Cell values show frequency (darker colour indicating higher). The red diagonal line marks perfect agreement. A strong concentration of values *below the diagonal*, especially when ground-truth affect balances are negative, shows that when participants reported being *sad overall*, the LLM still predicted them as *happy*, revealing a systematic positivity bias.



(a) Cases where the **ground-truth affect balance is positive**.



(b) Cases where the **ground-truth affect balance is negative**.

Fig. 3 | Systematic positivity bias in predicted affect balance across 7 time ranges. Participants are interpreted as happier than they actually reported, limiting the model's ability to predict real negative affects. Violin plots compare the distribution of LLM-predicted affect balance (the sum of positive affect scores minus the sum of negative affect scores) at each time range against the ground-truth distribution (GT, rightmost violin in each panel). **a** The LLM predicts participants as even happier than they actually reported, exaggerating the strength of positive feelings across all time ranges. **b** The LLM still predicts participants as being happy, with distributions shifted toward positive values. Neutral cases (affect balance = 0) are excluded to maintain a clear separation between the two panels.

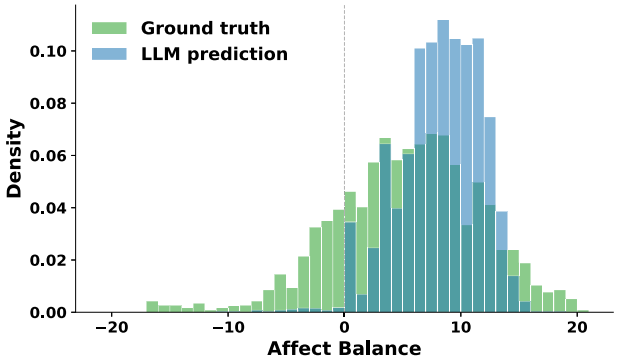


Fig. 4 | Predicted vs. ground-truth affect-balance distributions. The LLM-predicted distribution (blue; n = 4949 predictions) is shifted toward positive values relative to the ground-truth distribution (green; n = 1167 unique weekly surveys from 83 participants); mean affect balance is + 7.69 [95%CI: +7.30, +8.07] for predictions vs. +4.87 [+3.82,+5.90] for ground truth (shift = +2.84 [+ 1.77, +4.00]). The two distributions differ significantly (Kolmogorov--Smirnov D = 0.28, $p < 10^{-60}$; Earth Mover's Distance =3.28 affect-balance points). At the matched-pair level, the shift is asymmetric: in the subset where ground-truth affect balance is negative (n = 876 pairs), predicted balance still averages + 6.94 [+6.14, + 7.78] (vs. ground-truth mean −4.70 [−6.68, −3.14]); in the subset where ground-truth is positive (n = 3, 823 pairs), predicted and ground-truth means are nearly indistinguishable (+7.86 [+7.51, +8.22] vs. +7.35 [+6.62, +8.07]). 95% confidence intervals are from a cluster bootstrap resampling participants with replacement (1000 iterations). This pattern is consistent with the model defaulting to a prior centred near + 7 rather than tracking variation in input.

positive affect balance but the model predicted a negative balance are comparatively rare. This imbalance indicates that prediction errors disproportionately occur when participants experience unpleasant affect.

We next split predictions by whether the ground-truth affect balance was positive (pleasant) or negative (unpleasant). As shown in Fig. 3, predicted affect balances remain positively shifted in both cases. Notably, even when participants reported an unpleasant affect balance, the model often predicted a pleasant balance. Together, these results indicate a consistent tendency for the model's predictions to skew toward pleasant affective states. A direct overlay of the predicted and ground-truth affect-balance distributions is provided in Fig. 4, quantifying the shift with the Kolmogorov–Smirnov statistic and Earth Mover's Distance. See Supplementary Note 2: Case Study for the ground-truth and predicted affect scores for the three case studies (guided by Figure 2).

This tendency helps explain the aggregate performance observed in Table 2, where high accuracy and recall co-occur with extremely low specificity and low Matthews Correlation Coefficients. Rather than reflecting balanced discrimination between positive and negative affective states, these metrics are driven by the model's frequent prediction of pleasant affect.

Having identified a consistent skew in prediction errors, we next examine possible factors that might influence this behaviour. In particular, we assess whether the amount of user history available to the model, differences across participants and affects, or specific usage patterns shape how predictions diverge from self-reported affect. The following subsections examine these factors in turn.

Variation across affects

We examine how participants report individual affects in their weekly surveys and how predictive performance varies across affect types. See Supplementary Fig. 3 for the distributions of positive and negative affect items across participants and weeks. We test whether predictive performance differs between positive and negative affects and find that, across all seven Time Ranges, the LLM predicts negative affects more reliably than positive affects (Fig. 5).

After establishing this polarity-level difference, we examine heterogeneity at the level of individual affects. As shown in Table 3 (per-affect 95% CIs in Supplementary Tables 1 and 2), predictive performance varies substantially across affect items: some affects (e.g., *ashamed*) are predicted more accurately than others (e.g., *inspired*). Overall, variability in performance is more affect-driven than Time-Range-driven, consistent with the modest differences observed across Time Ranges in aggregate results. A complementary ordinal-metric view (Kendall's τ_b; per-affect CIs in Supplementary Table 3) shows the same positive–negative asymmetry: pooled across affects, the LLM achieves modest τ_b on negative affects (~0.13–0.19 across Time Ranges) but essentially zero τ_b on positive affects, while the historical baselines (First Week, Previous Week) achieve substantially higher τ_b on both. The performance of two adaptive approaches is discussed in the following subsection.

Variation across participants

We investigate if the prediction performance across Time Ranges varies by individual participants in our dataset. Heterogeneity of participant-level Time Range performance is reflected in an accuracy heatmap (Fig. 6). This graph answers the question: for each participant, and for each of the 17 weeks in the dataset, which Time Range gave the best prediction?

We observe substantial between-person heterogeneity in accuracy, with idiosyncratic participant-level patterns that persist across Time Ranges.

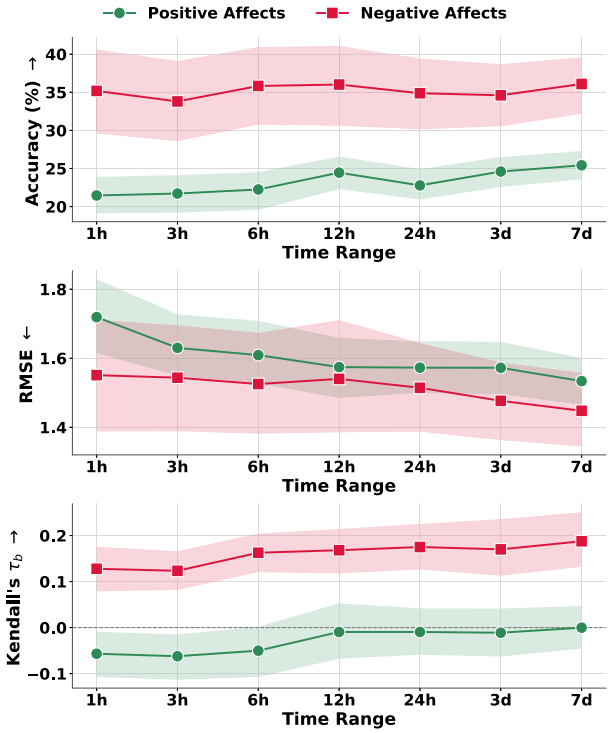


Fig. 5 | Predictive performance for positive (green circles) versus negative (red squares) affects across Time Ranges, by accuracy (top; higher is better), RMSE (middle; lower is better), and Kendall's τ_b (bottom; higher is better). Across all seven Time Ranges, the LLM predicts negative affects more reliably than positive affects; on the ordinal metric, τ_b for positive affects is essentially zero (CI band overlapping the dashed zero line) while negative affects retain a modest positive τ_b. Performance improves slightly at longer Time Ranges. Shaded bands are 95% cluster-bootstrap CIs (participants resampled with replacement, 1000 iterations).

Some participants have a single best Time Range while others have alternating best Time Ranges. For accuracy in general, some participants are consistently high-performing across 7 Time Ranges and 17 weeks (mostly green, e.g., P325), while others are consistently low (mostly yellow, e.g., P987). We also observe some participants for whom predictive accuracy alternates between short and long Time Ranges across the study. These could indicate instances where some event took place that perhaps weighed more on their feelings.

Since no Time Range is observed to consistently work well for different participants and affects, we examine if a dynamic method that selects the best Time Range per participant-affect could improve the overall performance on predicting the affect scores. We explored adaptive approaches that use the best Time Range in terms of accuracy and RMSE separately. Adaptive approaches (select the best Time Range for each participant-affect pair) have better overall predictive performance for positive affects but worse for negative affects (than using fixed Time Ranges), while achieving balanced performance across individual affects (see Supplementary Table 4 for details). Beyond these participant-level patterns, we next identify the specific participant–affect–Time-Range combinations where the model performs best and worst.

Best and worst performance

We observe that for some participants and some affects, the LLM achieved very high accuracy. For example, looking at overall accuracy for predicting "Hostile" across 7 Time Ranges, there are 5 participants (P263, P347, P524, P618, P739) whose accuracy was higher than 97.6%, while there are 5 participants (P082, P195, P314, P456, P587) whose accuracy was worse than 1.2%.

We observed substantial variability across combinations of participants, affects, and Time Ranges. While some combinations were predicted perfectly (100% accuracy), others failed completely (0% accuracy), even when evaluated over the same Time Ranges. This pattern suggests that the model can accurately infer affect when relevant behavioural signals are present, but is also prone to being misled by unrelated or ambiguous sensor data.

Across participants, overall accuracy varied substantially, ranging from approximately 55% for the highest-performing participants to around 12–14% for the lowest-performing participants when averaged across all affects and Time Ranges. Even among top-performing participants, accuracy remained well below ceiling, indicating that the model does not reliably predict all individual affects simultaneously.

Discussion

Recent work has explored using large language models (LLMs) to simulate human responses, sparking debate over whether such simulations can meaningfully approximate real people. Prior studies have done so by providing LLMs with, for example, short interview transcripts to prime them in simulating survey responses for 1,000 individuals[22]. Similarly, other work has supplied LLMs with transcripts of thought reflections and video daily diaries to assess personality traits[23].

Our study is the first to move beyond transcripts and systematically evaluate simulations of subjective affect grounded in longitudinal smartphone sensing data. By leveraging digital traces to prime LLMs, we examine their ability to predict weekly affective survey responses. Results reveal substantial variability across Time Ranges, affects, and individuals, showing that while LLMs can perform well under certain conditions, their accuracy in predicting subjective feelings is generally low and context-dependent. Most notably, we find a systematic positivity bias at the level of affect balance: the LLM tends to predict humans' overall affect as positive, indicating a more pleasant emotional state than humans report. Importantly, our results do not imply that all language models exhibit a similar over-positive bias, but they do provide evidence that LLMs can be biased when analysing human affect.

Several factors plausibly contribute to this systematic positivity bias. First, the pretraining corpora of modern LLMs—primarily large-scale web text—reflect a well-documented positivity bias in human language use across diverse domains and languages[24]. Second, post-training procedures such as Reinforcement Learning from Human Feedback (RLHF) reward helpful, agreeable, and non-distressing outputs[25], which may favour positively-valenced interpretations of ambiguous behavioural evidence. Because both mechanisms are shared across modern commercial LLMs, the positivity bias we observe is unlikely to be specific to Gemini-2.5-flash-lite and may instead reflect a broader property of off-the-shelf LLMs applied to affective inference.

Model configuration was refined through iterative trial-and-error. Based on a long-context benchmark[26], we selected Google's Gemini 2.5 series for its competitive performance in both context window length and accuracy at the time. We examined other commercial LLMs in preliminary tests, but found the Gemini 2.5 series handled our long sensor-diary inputs more reliably; several alternatives also lacked end-user controls to relax safety filters that incidentally blocked benign sensor-derived content, making them impractical for this study. Early runs used `Gemini-2.5-flash`, but we switched to `Gemini-2.5-flash-lite` after finding comparable performance in small-scale tests with five users, using the first three weeks of data. This change reduced costs by two-thirds ($0.10 vs. $0.30 per 1M input tokens) without compromising predictive quality. `Gemini-2.5-pro` was also tested, but despite similar predictive accuracy, its much higher costs ($1.25–$2.50 per 1M input and $10–$15 per 1M output) made it impractical for large-scale experiments.

Hyperparameters were tuned empirically. A low temperature (0.2–0.5) led to overly rigid predictions, while setting it to 1.0 better captured the variance across users, reflecting the creative rather than deterministic nature of the task. Fixed thinking budgets (e.g., 512 tokens) sometimes under- or

Table 3 | Affect-level performance across time ranges and baselines: accuracy (higher is better) and RMSE (lower is better), pooled within affect across all predictions for every participant and week.

Affect	Metric	Time Range							Baseline		
		1h	3h	6h	12h	24h	3d	7d	Mode	First Wk.	Previous Wk.
Active^{+}	Acc (%)	24.7	22.4	21.7	26.0	22.9	26.1	27.2	29.0	37.1	**46.8**
	RMSE	1.54	1.49	1.50	1.43	1.46	1.44	1.46	1.19	1.17	**1.04**
Alert^{+}	Acc (%)	28.2	27.3	26.3	26.5	26.9	27.3	28.3	30.6	34.4	**49.4**
	RMSE	1.36	1.34	1.35	1.31	1.29	1.33	1.38	1.13	1.32	**0.98**
Attentive^{+}	Acc (%)	26.7	29.8	28.8	32.8	30.0	30.0	30.2	35.2	42.5	**49.1**
	RMSE	1.38	1.31	1.32	1.28	1.25	1.27	1.24	1.06	1.08	**0.94**
Determined^{+}	Acc (%)	17.0	17.7	20.1	20.9	20.8	24.5	25.2	33.0	41.4	**49.3**
	RMSE	1.96	1.80	1.75	1.72	1.70	1.68	1.56	1.12	1.09	**0.98**
Inspired^{+}	Acc (%)	10.8	11.3	14.3	16.0	13.4	15.2	16.2	35.3	38.8	**44.5**
	RMSE	2.19	2.08	2.02	2.01	2.02	2.02	1.94	1.08	1.07	**1.01**
Afraid^{-}	Acc (%)	38.3	32.9	37.5	36.9	34.7	33.8	33.3	34.0	40.5	**49.6**
	RMSE	1.68	1.67	1.66	1.68	1.67	1.64	1.63	1.71	1.18	**1.10**
Ashamed^{-}	Acc (%)	44.1	42.1	43.1	44.0	43.0	43.0	44.3	43.9	43.9	**52.6**
	RMSE	1.40	1.41	1.41	1.44	1.42	1.38	1.33	1.40	1.13	**1.00**
Hostile^{-}	Acc (%)	48.8	50.5	49.6	49.8	48.8	49.9	50.5	51.8	48.9	**59.5**
	RMSE	1.19	1.19	1.22	1.20	1.23	1.20	1.17	1.17	0.95	**0.85**
Nervous^{-}	Acc (%)	20.4	21.2	23.9	23.0	22.6	22.0	25.4	27.3	31.3	**44.7**
	RMSE	1.81	1.80	1.70	1.76	1.68	1.64	1.60	1.22	1.32	**1.12**
Upset^{-}	Acc (%)	24.3	22.2	25.0	26.4	25.3	24.4	27.1	40.6	37.9	**45.2**
	RMSE	1.61	1.58	1.60	1.56	1.53	1.48	1.45	1.05	1.08	**1.02**
Positive^{+}	Acc (%)	21.5	21.7	22.2	24.5	22.8	24.6	25.4	32.6	38.8	**47.8**
	RMSE	1.72	1.63	1.61	1.57	1.57	1.57	1.53	1.11	1.15	**0.99**
Negative^{-}	Acc (%)	35.2	33.8	35.8	36.0	34.9	34.6	36.1	39.5	40.5	**50.3**
	RMSE	1.55	1.54	1.53	1.54	1.51	1.48	1.45	1.33	1.14	**1.02**

95% cluster-bootstrap CIs for every cell are provided in Supplementary Tables 1 and 2; the corresponding per-affect Kendall's τ_b (with CIs) is reported separately in Supplementary Table 3. The Mode baseline predicts the most frequent Likert rating per affect; the First Week and Previous Week baselines use each participant's own earlier responses. **Bold**: best overall; underline : best time range excluding baselines (ties resolved using unrounded values). Positive affects $^{+}$ (green); negative $^{-}$ (red).

over-allocated resources; adopting a dynamic budget (−1) allowed the model to adapt flexibly to the length of input diaries. Default settings for top-*p* and other sampling parameters were retained, as adjustments did not yield meaningful improvements.

To address missing predictions from insufficient user data or transient API errors, our modular pipeline can be extended to incorporate context-aware strategies, such as leveraging prior ground truth (e.g., a user's historical means) or similarity-based estimates (e.g., nearest-neighbour users).

Finally, we used zero-shot prompting (no in-context examples and no task-specific fine-tuning). This avoids anchoring on a small, potentially biased exemplar set and reduces spurious inferences from illustrative examples, while keeping evaluation conditions identical across users and Time Ranges for comparability. This design choice differs from our earlier work[15], which examined few-shot performance with a smaller participant pool; here, the goal is to characterise the LLM's intrinsic priors at scale, holding the prompting setup constant across all participants and Time Ranges. Because the task-specific instruction is a swap-in module, it can be replaced by a few-shot template that prepends *k* labelled exemplars without altering any other component of the pipeline. This enables clean ablation studies—e.g., varying *k*, exemplar selection policy (random, affect-balanced, user-personalized), and provenance (global vs. user-specific)—to isolate the impact of exemplars under otherwise fixed prompts, schemas, and hyperparameters. Overall, this configuration balances accuracy, comparability, and cost for large-scale prediction.

Our results show that increasing token usage, whether through longer prompts (i.e., more hourly windows of sensor-derived diary entries), extended reasoning traces (more thoughts), or more elaborate outputs, did not lead to meaningful improvements in prediction performance. This finding is consistent with prior studies on reasoning[27], contextual information usage[28], and prediction performance[29], which collectively show that although it may be intuitively expected that additional context, reasoning, or evidence would enhance outcomes, the observed relationships with prediction performance were consistently weak.

A likely explanation is that the model does not necessarily distinguish between informative and redundant input. Adding more hourly windows in a sensor diary may introduce repetitive or noisy contextual details rather than genuinely informative signals about a person's affective state. Similarly, longer reasoning chains may reflect verbosity rather than deeper inference, and more elaborate outputs can inflate token counts without providing additional useful evidence. These results suggest that model performance is not simply a function of token volume, but instead depends on the informativeness of the content[30,31].

Another factor relates to the model's use of a dynamic "thinking budget," where the number of reasoning tokens is pre-determined based on the assumed complexity of the prompt (e.g., input length or task difficulty). Such allocation may overestimate or underestimate the true demands of the task, causing the model to either overthink or rush into an answer. In other words, the model may lack the ability to conclude reasoning at an appropriate point, leading to inefficiencies or premature responses[32].

Reliable prediction of the effects is a prerequisite for building LLM-based systems that can faithfully simulate human behaviours. In this context, our findings highlight a key limitation: LLMs may not effectively prioritize the most salient affective information when provided with large streams of sensor-derived data. Future work should explore alternative

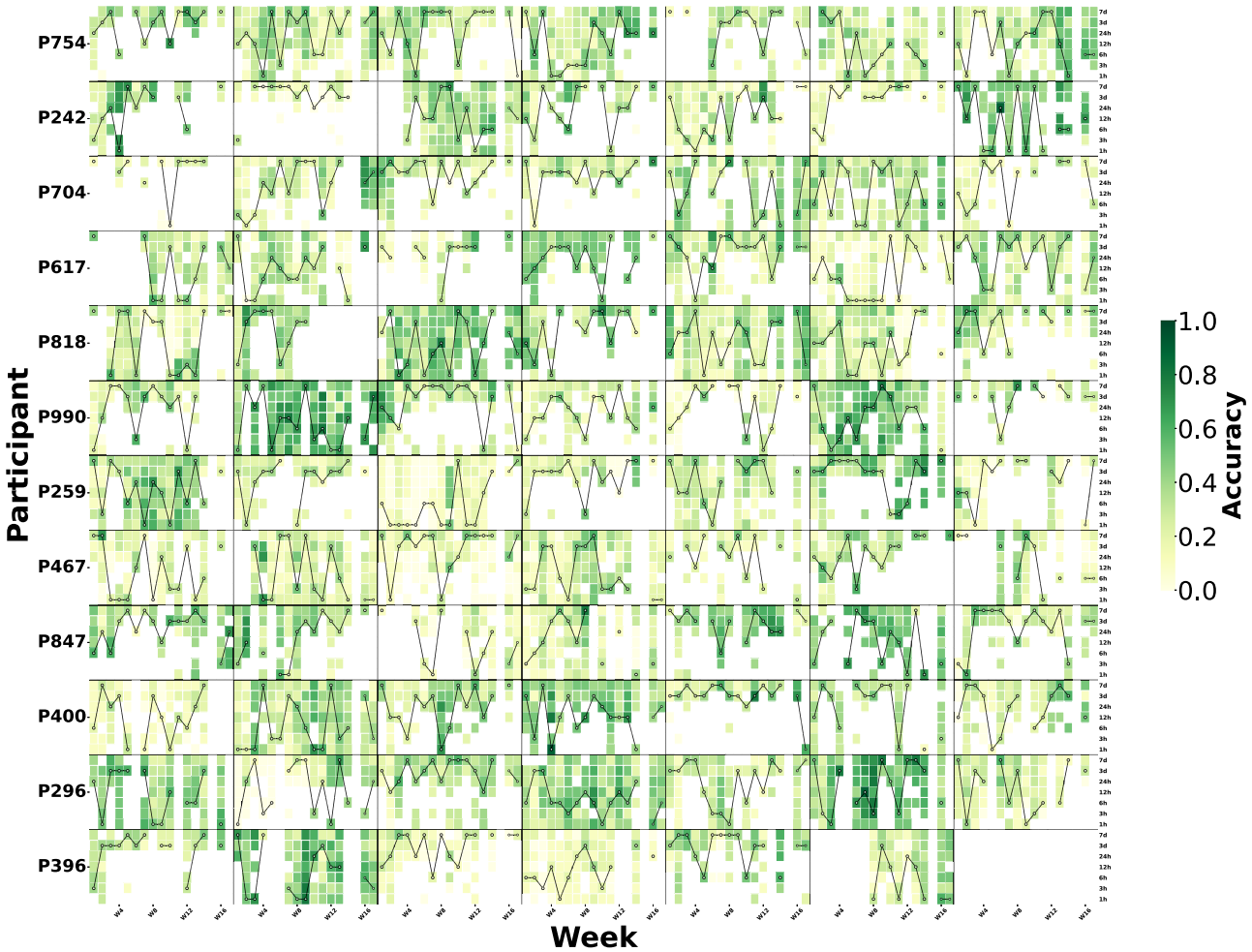


Fig. 6 | Participant-week accuracy by Time Range heatmap. This heatmap visualizes prediction accuracy across 83 participants, 17 weeks, and 7 Time Ranges. Each participant occupies a grid block whose 7 rows are the Time Ranges (from 1 hour up to 7 days before a weekly survey, corresponding to the number of hourly windows included in a diary) and whose 17 columns are the study weeks. Participant blocks are tiled seven per row across the figure (the *x*-axis therefore shows the 17 weeks repeated for each participant in a row). The accuracy colormap spans 0--100% (yellow-green), with best-performing ranges marked by white circles and connected by black lines.

temporal and structural representations, for example, reasoning day by day instead of using only fixed 1-hour windows, or first analysing one sensor (or one category of sensors) before integrating reasoning across multiple categories. Another direction is to enrich the diary with derived behavioural features or pre-computed embeddings as additional in-context signals, combining the interpretability of narrative diaries with the predictive signal that feature-based methods extract from raw sensor streams. Such structured approaches could help the model focus on the most informative signals at each stage, leading to better prediction performance through the relevance and organization of data rather than simply increasing its quantity.

Our analyses reveal a systematic positive bias: LLMs tend to portray people as feeling happier than they report. Across case studies (see Supplementary Note 2: Cast Study for details), positive states such as Alert, Attentive, or Determined were frequently inflated, while negative states such as Nervous, Hostile, or Upset were consistently downplayed or missed. This positivity bias arises when diaries lack overt cues of distress, leading the model to interpret routine notifications or communications as neutral or positive. In addition, this bias occurs when salient but ambiguous events, such as phone calls, are reframed as evidence of engagement rather than potential sources of conflict or anxiety.

This bias contrasts with principles of affective science, which emphasize a pervasive negativity bias in human judgement and feeling[33,34]. This universal tendency to overweight negative information is often framed in evolutionary terms: humans evolved to be more strongly motivated to identify and avoid threats than to seek out opportunities because of the asymmetrically high cost of failing to avoid threat (e.g., death or permanent incapacitation) versus missing an opportunity (e.g., foregoing resources or a chance to reproduce). Relatedly, Frijda's law of hedonic asymmetry[35] and research on emotion duration[36] both show that negative experiences linger and accumulate, whereas positive experiences are often more fleeting and situational. Yet our LLM-based simulations reversed this asymmetry: negative states often disappeared unless directly signalled, while positive states were inferred even from minimal cues. In other words, the model projects a smoother, more upbeat emotional trajectory than the feelings people actually reported.

The broader implication is that LLMs process diaries at the surface level, without abstracting events into psychological constructs such as appraisal, arousal, or valence[37]. Without such grounding, the model defaults toward optimism, systematically misrepresenting users' affective lives. Future systems could mitigate this bias by enriching passive sensing data with affective-science features. For example, repeated late-night notifications might be coded as disrupted rest (arousal)[38,39], sustained app switching as reduced perceived control and increased stress[40,41], and silence in communication streams as social withdrawal[42,43]. Such mappings would allow the LLM to move beyond

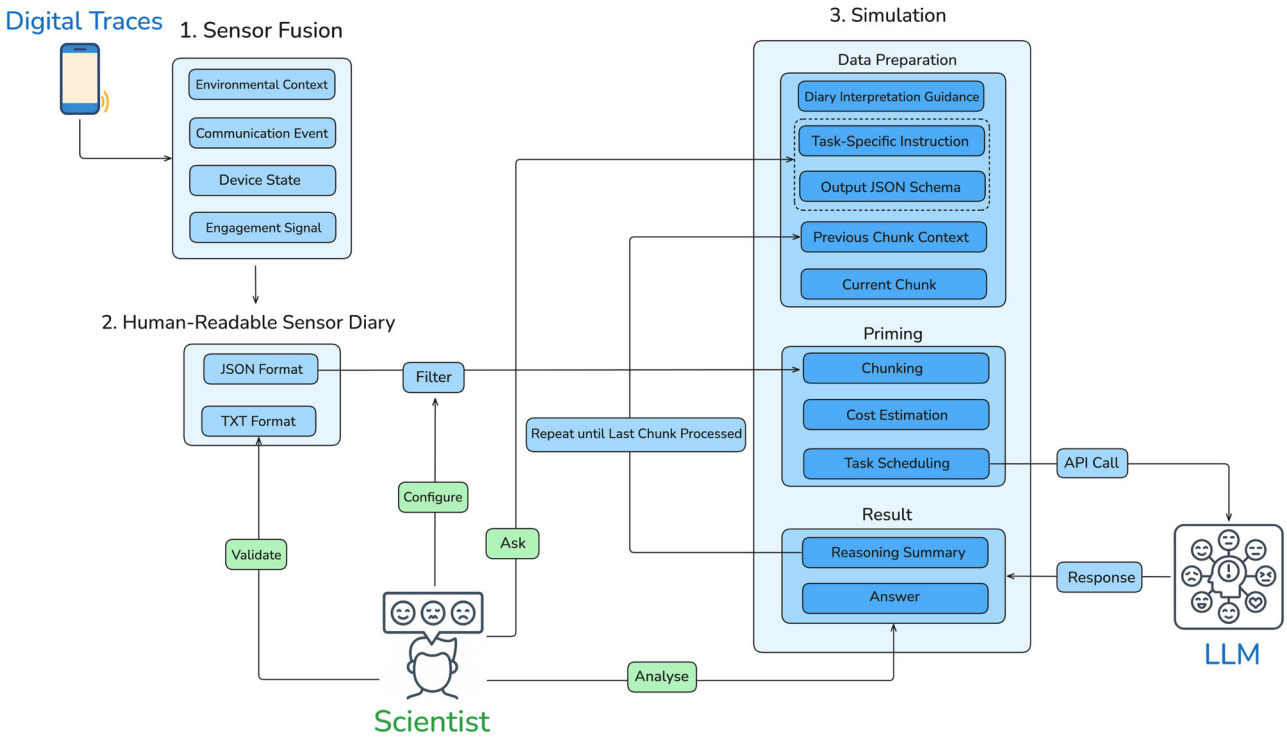


Fig. 7 | An affective simulation pipeline. Smartphone traces are fused into sensor diaries, which prime an LLM to generate simulated participants' responses with reasoning summaries. Blue boxes represent processes within the pipeline, while green boxes indicate actions performed by the Scientist (e.g., validate, configure, analyse).

surface-level interpretation of device events, reduce bias, and more faithfully reproduce self-reported affects, ultimately improving its ability to support user reflection and wellbeing.

A key limitation of this study is the uneven availability and reliability of passive sensing data. Sensor coverage varied across devices, operating systems, and user behaviours, and participants could complete surveys on devices not registered for data collection, leading to imbalances in data availability across Time Ranges (e.g., fewer 1-hour than 7-day diaries). As a result, shorter Time Ranges often contained insufficient contextual information to support reliable affect prediction. More broadly, sparsity and heterogeneity of passive digital traces are intrinsic to real-world data collection and place fundamental constraints on how accurately LLMs can infer subjective affect when affective cues are weak, ambiguous, or absent. Finally, our primary findings derive from a single model (Gemini-2.5-flash-lite) under a zero-shot setup, and from a limited sample ($N = 83$) from one university cohort; whether the positivity bias generalises to other populations, or persists under few-shot prompting, remains to be established. Future work could also examine whether prediction quality and the positivity bias vary across participant demographics (e.g., age, gender, university level, and language background), which our sample size precludes analysing reliably.

In conclusion, we use longitudinal smartphone data to evaluate how accurately large language models can predict affective survey responses. This study provides the first systematic evaluation of how reliably an LLM can approximate users' self-reported emotional states. Our quantitative analyses reveal three central findings. First, predictive performance is heterogeneous across affect types: negative affects are predicted more reliably than positive affects, and variability is more affect-driven than data-driven. Second, aggregating predictions into an overall affect balance (positive minus negative) shows a systematic positivity bias: the LLM consistently interprets people as feeling happier than their self-reports indicate, often predicting net positivity when users report as being predominantly sad. Third, prediction accuracy is highly user-specific, with no single Time Range consistently optimal across all users, suggesting that adaptive or personalised approaches are needed.

To complement these quantitative results, our case studies (see Supplementary Note 2: Case Study) illustrate how this systematic positivity bias emerges in practice, with routine device events reframed as neutral or positive and potentially distressing events interpreted as engagement or focus. Together, these findings highlight both the promise and the limitations of LLM-based human affect inference. While such models can approximate survey responses under certain conditions, they systematically overestimate happiness and underrepresent sadness. Future work should therefore focus on grounding passive sensing data in affective-science constructs to move beyond surface-level interpretations, reduce bias, and build systems that more faithfully predict users' affects, ultimately improving their ability to support reflection and wellbeing.

Methods

Our study was approved by the Office of Research Ethics and Integrity, The University of Melbourne, and all methods were performed in accordance with the relevant guidelines and regulations (ethics protocol number 25051). Informed consent was obtained from all participants in the study.

We developed an affective simulation pipeline that leverages pre-existing participant-generated traces to answer questions on behalf of these individuals (Fig. 7). A dataset is used to generate "diary" entries (human-readable sensor narratives) on an hourly basis to summarise what we know about the participant (e.g., behaviour, actions, context). Our pipeline uses this data to "prime" an LLM, enabling it to simulate asking the participant a question at a specific point in time. The LLM responds in the manner it predicts the participant would, while also generating a justification and explanation for its answer.

Data collection via digital phenotyping

In our pipeline, we used a dataset of smartphone activity from our previous study under the same ethics approval[44]. The dataset was collected over 17 weeks from 151 participants, providing more than 242.9 million records. The AWARE software[45] was installed on participants' Android devices to passively record sensing data and administer in-situ questionnaires; iOS devices were excluded because the platform's restrictions on background

sensing substantially limit AWARE's sensor coverage. In this study, we chose the International Positive and Negative Affect Schedule - Short Form (I-PANAS-SF)[46] as our prediction task. We selected I-PANAS-SF because it was the most frequently administered questionnaire in the dataset, ensuring maximal data availability and statistical power. Beyond practicality, I-PANAS-SF is a validated, widely used measure of affect that captures both positive and negative states with minimal participant burden, making it well suited to repeated smartphone-based data collection. At the end of each week, the participants completed the I-PANAS-SF to assess their affect during the preceding week. The I-PANAS-SF questionnaire includes five positive affect items (active, determined, attentive, inspired, alert) and five negative affect items (upset, hostile, ashamed, nervous, afraid). Participants rated each item on a 5-point Likert scale (1 = never, 5 = always) to indicate how frequently they had experienced each affect over the past week. The demographic characteristics of the 83 participants in our analysis sample are summarised in Table 4.

Passive sensing data was collected from a range of smartphone sensors, including hardware sensors (Bluetooth, GPS location, and network connectivity) and software sensors (application usage, battery level, call and message logs, screen status, keyboard input, and on-screen text). To address privacy concerns, the on-screen text was collected only from specific sources: Twitter, Facebook, Reddit, Pinterest, Tumblr, Chrome, Google, and Firefox, as well as applications categorised under News/Magazines and Books/Reference. At the time of the study, the participants had the option to further disable logging for certain apps, or to completely disable logging for a certain period of time. Furthermore, passwords or similarly sensitive information were masked at the time of recording.

Table 4 | Study population: demographic characteristics of the 83 participants in the analysis sample

Characteristic	Value
Participants	83
Age (years)	24.2 ± 5.3 (range 18–52)
Gender	
Female	45 (54.2%)
Male	35 (42.2%)
Other	3 (3.6%)
University level	
Undergraduate	33 (39.8%)
Postgraduate	50 (60.2%)
Primary language	
English	45 (54.2%)
Other	38 (45.8%)

Values are *n* (%) unless otherwise noted.

Data handling

Our analysis pipeline ensures that all data transmissions to the cloud are encrypted, never retained for model training, and used solely for inference. Predictions are generated using Google's Gemini-2.5-flash-lite via the paid API service, under which prompts and responses are not used to improve Google products and may be only cached temporarily for compliance and policy checks[47].

Our pipeline incorporates multiple safety measures through a two-stage filtering process: first at the prompt level, and then at the model level. This layered approach is designed to minimise the risk of unsafe or inappropriate outputs. In the context of our study, prompts are restricted to questions grounded in sensor data and to predicting affects (e.g., I-PANAS-SF ratings), avoiding open-ended generation that could risk harmful or misleading content.

Sensor data to human-readable diaries

To convert raw smartphone sensor readings into human-readable diaries, we developed a narrative synthesis module extending the AWARE Narrator framework[48]. The module preprocesses inputs from 12 sensors, grouped into four categories (Table 5), and organises them into structured diary entries.

Sensor events are aggregated into a time window (default: 1 hour, aligning with the shortest 1-hour Time Range), to capture meaningful temporal associations. Location clusters are enriched with semantic place labels through reverse geocoding. Wi-Fi and Bluetooth scans are integrated using a configurable time gate (default: 5 minutes, aligning with previous guidelines[49]) to account for variation in the detection results. Our pipeline enables selective configuration of which sensors are included in the diary generation process, allows us to specify key temporal parameters such as the definition of night-time (default: 10 p.m. to 6 a.m., used to help identify the home), and select which phone applications to be included or excluded. The resulting signals are integrated into coherent diary entries that describe environmental context, communication events, device state, and user engagement.

Table 5 | Key statistics of one hourly window in a diary

Window ID | Year-Month-Date (Day of the Week) | Start - End Time		
Category	Sensor	Key Statistics
Environmental Context	WiFi	– Connection sequence: [WiFi1 → WiFi2] – Networks used – Nearby networks (5-min scans)
	Bluetooth	– Nearby devices (5-min scans)
	Location	– Location sequence and duration – Location details (areas, landmarks, address, distance to home)
Communication Event	Notification	– Notification count – App count – Notification logs breakdown (app: [full text])
	Calls	–Call count – People involved count – Status breakdown (made, received, missed) – Sequence (time, status, person ID)
	Message	– Message count – People involved count – Status breakdown (sent, received) – Sequence (time, status, person ID)
Device State	Battery	– Status sequence (charging/discharging, from X% to Y%)
	Installation	– Installed apps – Updated apps – Removed apps
Engagement Signal	Screen	– Event breakdown (on/off, locked/unlocked) – Screen activations count – Average time between activations – Session timelines (start-end, duration)
	Application	–App usage duration – App sequence by session – App switches – Usage breakdown (app: [time, percentage, revisit])
	Keyboard	– Typing session count – Keyboard event count – Typing duration – Typing speed (chars/min, words/min) – Typing logs breakdown (app: [potential typos, typing text])
	Screen Text	– Apps tracked count – Screen text logs breakdown (app: [full text])

The module generates diaries in two complementary formats: (1) text files for human inspection, and (2) structured JSON files for downstream analysis with LLMs. This dual representation provides both transparency for evaluation and scalability for prediction tasks. To establish our settings for this study, we iterated our design with the data of only 10 participants (e.g., optimising the system configuration, the diary structure, and the prompt template), drawing on location clustering functions offered by RAPIDS[50–52] and adapting narrative generation strategies from AWARE Narrator[48].

We retained 83 of 151 participants in the dataset for this study after applying a selection threshold of at least 9 of the 17 weeks, which corresponds to the elbow point in the retention curve where participant loss accelerates beyond this level. There are 45 females, 35 males, and 3 participants who self-identified as non-binary or another gender identity. Most are young adults, with 76 participants aged 18 to 30, five aged 31 to 35, and two aged 35 or older. Primary language is split between English (45) and non-English (38) speakers.

Predicting affects

Our analysis pipeline is configurable, enabling us to frame different questions for diary analysis. It is organised around three modular components (see Supplementary Note 1: Prompts): (1) an overall prompt template that provides general guidance for interpreting sensor diaries, (2) a task-specific instruction that specifies the self-reflection question to be asked, and (3) a JSON schema that constrains the model's output to a structured format. While the design is general, in this study we focus specifically on predicting participants' affect.

The task instruction is configured to predict weekly I-PANAS-SF ratings using zero-shot prompting (no in-context examples and no task-specific fine-tuning). To handle long diary files, we adopted a token-aware chunking strategy with an upper bound of 524k tokens, ensuring consistent processing across participants. Chunks were processed in chronological sequence, with reasoning carried forward across chunks to maintain temporal continuity. Each chunk was analysed using Google's Gemini-2.5-flash-lite, prompted to assign Likert scores from 1 to 5 for ten affects (positive: Active, Alert, Attentive, Determined, Inspired; negative: Afraid, Ashamed, Hostile, Nervous, Upset). Outputs adhered to the JSON schema, pairing each score with a reasoning string grounded in diary evidence.

Robustness was ensured through rate limiting to respect API quotas and an error-retry mechanism with exponential backoff, supporting reliable large-scale prediction of emotional states from passive smartphone diaries. Model hyperparameters were configured with a temperature of 1.0, allowing adaptation to variance across participants, and a dynamic thinking budget of -1.

We evaluated the affective predictions of the LLM using three metrics: accuracy, root mean squared error (RMSE), and Kendall's τ_b. Accuracy captures exact agreement with the reported Likert value (1–5), RMSE captures the magnitude of prediction errors, and τ_b—a rank-correlation coefficient appropriate for ordinal data with ties—captures the degree to which predictions preserve the ordering of participants' reported scores. Aggregate τ_b values pool predicted and ground-truth scores across the ten affects, whereas per-affect values pool within each affect. Performance was additionally assessed on affect balance, which distinguishes between overall positive and negative feelings, calculated as the sum of positive affects minus the sum of negative affects[20,21]. Unless noted otherwise, all reported metrics carry 95% percentile cluster-bootstrap confidence intervals, computed by resampling participants with replacement over 1000 iterations to respect the repeated-measures structure of the data.

Data availability

Due to the sensitive and personally identifiable nature of smartphone data, the dataset cannot be made publicly available. Researchers seeking access should contact the authors and may be required to obtain ethical approval and enter into a data-use agreement.

Code availability

The code used to generate and evaluate the LLM's prediction will be made available upon reasonable request and/or at the time of publication, subject to ethical approval and data privacy constraints.

Received: 23 March 2026; Accepted: 31 August 2026;
Published online: 27 October 2026

References

1. Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C. & Kardes, F. R. On the automatic activation of attitudes. *J. Personal. Soc. Psychol.* **50**, 229–238 https://doi.org/10.1037/0022-3514.50.2.229 (1986).
2. Fazio, R. H. On the automatic activation of associated evaluations: An overview. *Cogn. Emot.* **15**, 115–141 https://doi.org/10.1080/02699930125908 (2001).
3. Kuppens, P. & Verduyn, P. Looking at emotion regulation through the window of emotion dynamics. *Psychol. Inq.* **26**, 72–79 https://doi.org/10.1080/1047840x.2015.960505 (2015).
4. Nesse, R. M. & Ellsworth, P. C. Evolution, emotions, and emotional disorders. *Am. Psychol.* **64**, 129–139 https://doi.org/10.1037/a0013503 (2009).
5. Eisele, G. et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment* **29**, 136–151 https://doi.org/10.1177/1073191120957102 (2020).
6. Rintala, A., Wampers, M., Myin-Germeys, I. & Viechtbauer, W. Response compliance and predictors thereof in studies using the experience sampling method. *Psychol. Assess.* **31**, 226–235 https://doi.org/10.1037/pas0000662 (2019).
7. Koval, P., Kalokerinos, E. K., Verduyn, P. & Greiff, S. Introduction to the special issue: Capturing the dynamics of emotion and emotion regulation in daily life with ambulatory assessment. *Eur. J. Psychol. Assess.* **36**, 433–436 https://doi.org/10.1027/1015-5759/a000599 (2020).
8. Kaiser, C. & Oswald, A. J. The scientific value of numerical measures of human feelings. *Proc. Natl. Acad. Sci.***119**, https://doi.org/10.1073/pnas.2210412119 (2022).
9. Yan, T. & Williams, D. Response burden–review and conceptual framework. *J. Stat.* **38**, 939–961 https://doi.org/10.2478/jos-2022-0041 (2022).
10. Cornet, V. P. & Holden, R. J. Systematic review of smartphone-based passive sensing for health and wellbeing. *J. Biomed. Inform.* **77**, 120–132 https://doi.org/10.1016/j.jbi.2017.12.008 (2018).
11. Sükei, E., Norbury, A., Perez-Rodriguez, M. M., Olmos, P. M. & Artés, A. Predicting emotional states using behavioral markers derived from passively sensed data: Data-driven machine learning approach. *JMIR Mhealth Uhealth* **9**, e24465 https://doi.org/10.2196/24465 (2021).
12. Zhang, J., Yin, Z., Chen, P. & Nichele, S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf. Fusion* **59**, 103–126 https://doi.org/10.1016/j.inffus.2020.01.011 (2020).
13. Kitadai, A., Ogawa, K. & Nishino, N. Examining the feasibility of large language models as survey respondents. *2024 IEEE International Conference on Big Data (BigData)* 3858–3864, https://doi.org/10.1109/BigData62323.2024.10825497 2024).
14. Serapio-García, G. et al. A psychometric framework for evaluating and shaping personality traits in large language models. *Nat. Mach. Intell.* **7**, 1954–1968 https://doi.org/10.1038/s42256-025-01115-6 (2025).
15. Zhang, T., Teng, S., Jia, H. & D'Alfonso, S. Leveraging LLMs to predict affective states via smartphone sensor features. *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing* https://doi.org/10.1145/3675094.3678420 (2024).
16. Tak, A. N. et al. Mechanistic interpretability of emotion inference in large language models. *Findings of the Association for Computational*

Linguistics: ACL 2025 13090–13120, https://doi.org/10.18653/v1/2025.findings-acl.679 (2025).
17. Salecha, A. et al. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus* **3**, pgae533 https://doi.org/10.1093/pnasnexus/pgae533 (2024).
18. Houben, M., Van Den Noortgate, W. & Kuppens, P. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychol. Bull.* **141**, 901–930 https://doi.org/10.1037/a0038822 (2015).
19. Trull, T. J., Lane, S. P., Koval, P. & Ebner-Priemer, U. W. Affective dynamics in psychopathology. *Emot. Rev.* **7**, 355–361 https://doi.org/10.1177/1754073915590617 (2015).
20. Diener, E. et al. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Soc. Indic. Res.* **97**, 143–156 https://doi.org/10.1007/s11205-009-9493-y (2009).
21. Pinquart, M. Age differences in perceived positive affect, negative affect, and affect balance in middle and old age. *J. Happiness Stud.* **2**, 375–405 https://doi.org/10.1023/a:1013938001116 (2001).
22. Park, J. S. et al. Generative agent simulations of 1000 people (2024).
23. Wright, A. G. C. et al. Assessing personality using zero-shot generative AI scoring of brief open-ended text. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-025-02389-x (2026).
24. Dodds, P. S. et al. Human language reveals a universal positivity bias. *Proc. Natl. Acad. Sci.* **112**, 2389–2394 https://doi.org/10.1073/pnas.1411678112 (2015).
25. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback (2022).
26. Burnham, T. & Adamczewski, B. LLMs now accept longer inputs, and the best models can use them more effectively. Epoch AI (2025).
27. Levy, M., Jacoby, A. & Goldberg, Y. Same task, more tokens: The impact of input length on the reasoning performance of large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 15339–15353, https://doi.org/10.18653/v1/2024.acl-long.818 (2024).
28. Liu, N. F. et al. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist*. **12**, 157–173, https://doi.org/10.1162/tacl_a_00638 (2024).
29. Sypherd, C., Petrov, S., George, S. & Belle, V. Incorporating token usage into prompting strategy evaluation. *arXiv preprint arXiv:2505.14880* https://doi.org/10.48550/arXiv.2505.14880 (2025).
30. Shi, F. et al. Large language models can be easily distracted by irrelevant context. *International Conference on Machine Learning*, 31210–31227, https://doi.org/10.48550/arXiv.2302.00093 (2023).
31. Zhang, Y., Das, S. S. S. & Zhang, R. Verbosity ≠veracity: Demystify verbosity compensation behavior of large language models, https://doi.org/10.48550/ARXIV.2411.07858 (2024).
32. Lanham, T. et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* https://doi.org/10.48550/arXiv.2307.13702 (2023).
33. Rozin, P. & Royzman, E. B. Negativity bias, negativity dominance, and contagion. *Personal. Soc. Psychol. Rev.* **5**, 296–320 https://doi.org/10.1207/s15327957pspr0504_2 (2001).
34. Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D. Bad is stronger than good. *Rev. Gen. Psychol.* **5**, 323–370 https://doi.org/10.1037/1089-2680.5.4.323 (2001).
35. Häuselmann, A., Sears, A. M., Zard, L. & Fosch-Villaronga, E. EU law and emotion data. *CoRR*, https://doi.org/10.48550/ARXIV.2309.10776 (2023).
36. Verduyn, P. & Lavrijsen, S. Which emotions last longest and why: The role of event importance and rumination. *Motiv. Emot.* **39**, 119–127 https://doi.org/10.1007/s11031-014-9445-y (2014).
37. Lin, Z. From prompts to constructs: A dual-validity framework for large language model research in psychology. *Annu. Rev. Psychol.* **78**, 2.1–2.26 https://doi.org/10.1146/annurev-psych-100925-034807 (2027).
38. Cajochen, C. et al. Evening exposure to a light-emitting diodes (LED)-backlit computer screen affects circadian physiology and cognitive performance. *J. Appl. Physiol.* **110**, 1432–1438 https://doi.org/10.1152/japplphysiol.00165.2011 (2011).
39. Levenson, J. C., Shensa, A., Sidani, J. E., Colditz, J. B. & Primack, B. A. The association between social media use and sleep disturbance among young adults. *Preve. Med.* **85**, 36–41 https://doi.org/10.1016/j.ypmed.2016.01.001 (2016).
40. Altmann, E. M. & Trafton, J. G. Timecourse of recovery from task interruption: Data and a model. *Psychon. Bull. Rev.* **14**, 1079–1084 https://doi.org/10.3758/bf03193094 (2007).
41. Mark, G., Wang, Y. & Niiya, M. Stress and multitasking in everyday college life: an empirical study of online activity. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 41-50, https://doi.org/10.1145/2556288.2557361 (2014).
42. Degerman, D. Silence, depression, and bodily doubt: toward a phenomenology of silence in psychopathology. *Philos. Psychol.* **38**, 126–149 https://doi.org/10.1080/09515089.2024.2354447 (2024).
43. Bortolan, A. Filling the silence: Self-experience and communication in social anxiety. *Erkenntnis*https://doi.org/10.1007/s10670-025-00986-w (2025).
44. Zhang, T., Russell, M., Kostakos, V. & D'Alfonso, S. Studentsense: Primary outcomes of a smartphone digital phenotyping study of university students in Australia. *Comput. Hum. Behav. Rep.* **19**, 100769 https://doi.org/10.1016/j.chbr.2025.100769 (2025).
45. van Berkel, N., D'Alfonso, S., Kurnia Susanto, R., Ferreira, D. & Kostakos, V. Aware-light: a smartphone tool for experience sampling and digital phenotyping. *Personal. Ubiquit. Comput.* **27**, 435–445 https://doi.org/10.1007/s00779-022-01697-7 (2022).
46. Thompson, E. R. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *J. Cross-Cultural Psychol.* **38**, 227–242 https://doi.org/10.1177/0022022106297301 (2007).
47. Google. Gemini API terms of service: Data use for paid services. Google for Developers (2025).
48. Zhang, T., Kojima, M. & D'Alfonso, S. Aware narrator and the utilization of large language models to extract behavioral insights from smartphone sensing data, https://doi.org/10.48550/ARXIV.2411.04691 (2024).
49. O'Neill, E. et al. *Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape*, 315–332 (Springer Berlin Heidelberg, 2006).
50. Vega, J. et al. Reproducible analysis pipeline for data streams: Open-source software to process data collected with mobile devices. *Front. Digit. Health* **3**, https://doi.org/10.3389/fdgth.2021.769823 (2021).
51. Canzian, L. & Musolesi, M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* 1293-1304, https://doi.org/10.1145/2750858.2805845 (2015).
52. Doryab, A., Chikarsel, P., Liu, X. & Dey, A. K. Extraction of behavioral features from smartphone and wearable data. *arXiv preprint arXiv:1812.10394* (2018).

Acknowledgements

This work is partly funded by the AI Centre for the Empowerment of Human Learning, The Research Council of Norway (357493).

Author contributions

L.F. and V.K. contributed to conceptualization and methodology. T.Z., L.F., and S.D. contributed to data curation. L.F. conducted the investigation, formal analysis, and visualization. V.K., H.J., and J.G. provided supervision. L.F., L.C.B., T.Z., S.Z., R.F., Y.M., and S.T. wrote the original draft. L.F., V.K., P.K., H.J., J.G., and S.D. reviewed and edited the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44482-026-00049-z.

Correspondence and requests for materials should be addressed to Le Fang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026