Evaluating LLM-Based Emotional Forecasting in Soloist: A 1-Year Longitudinal Case Study:

75% of the time, our forecasts are within 10 points of your actual reality.

ABSTRACT

This study evaluated the effectiveness of large language models (LLMs) for predicting emotional states using longitudinal mood tracking data from a single subject over 372 consecutive days. We employed OpenAI's GPT-4o-mini to generate 3-day forward emotional forecasts based on daily mood tracking data from the Daylio app (April 16, 2020 - April 16, 2021). The system processed 376 predictions using synthetically generated daily summaries combined with 4-day historical windows as input context. The LLM-based forecasting system demonstrated modest but statistically significant predictive capability (r = 0.369, $p < 0.05$) with a mean absolute error of 15.0 points on a 100-point scale. Overall, 52.2% of predictions fell within ±10 points of actual scores. Performance varied significantly by emotional state range, with excellent states (91-100) showing 9.6-point average error compared to 18.3-point average error for low states (30-50). The model excelled during stable periods but struggled with emotional volatility and sudden mood shifts. Future developments should focus on dynamic confidence modeling and enhanced temporal pattern recognition to improve prediction accuracy during periods of emotional volatility.

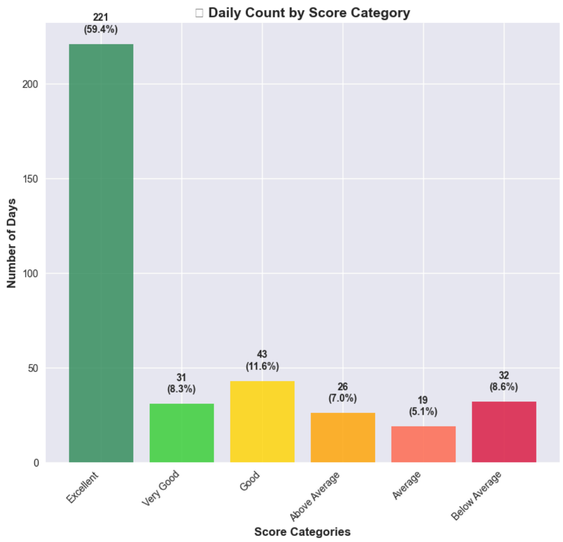
Keywords: emotional forecasting, large language models, mood tracking, longitudinal analysis, predictive modeling

KAGGLE DATASET PROFILE

This profile presents a comprehensive analysis of a 372-day longitudinal dataset (April 16, 2020 - April 16, 2021) from a single subject, pseudonymized as "Arwin," utilizing personal analytic data collected through daily mood tracking and activity logging within the ‘Daylio’ app, https://daylio.net/. The complete dataset can be found here: https://www.kaggle.com/datasets/kingabzpro/daylio-mood-tracker.

SUBJECT OVERVIEW & PSYCHOLOGICAL PROFILE

Arwin presents as a psychologically resilient individual with exceptional emotional regulation capabilities, demonstrated through consistent high well-being scores and adaptive activity patterns. This subject represents an exemplary case for studying, including optimal psychological functioning during challenging periods such as the pandemic of 2020.



Study Period: April 16, 2020 - April 16, 2021
Data Points: 372 consecutive daily entries
Occupation: Knowledge worker with significant engagement in coding, learning, and creative activities.

Arwin exhibits remarkably high emotional stability with an average daily well-being score of 8.63/10 (SD = 1.42).

1. Analysis Reveals Distinct Weekday Patterns

- Peak Performance Days: Thursday (8.89/10) and Friday (8.78/10)
- Challenging Days: Tuesday (8.45/10) and Wednesday (8.44/10)
- Weekend Performance: Slightly higher than weekdays (+0.15 points)

2. Resilience Indicators

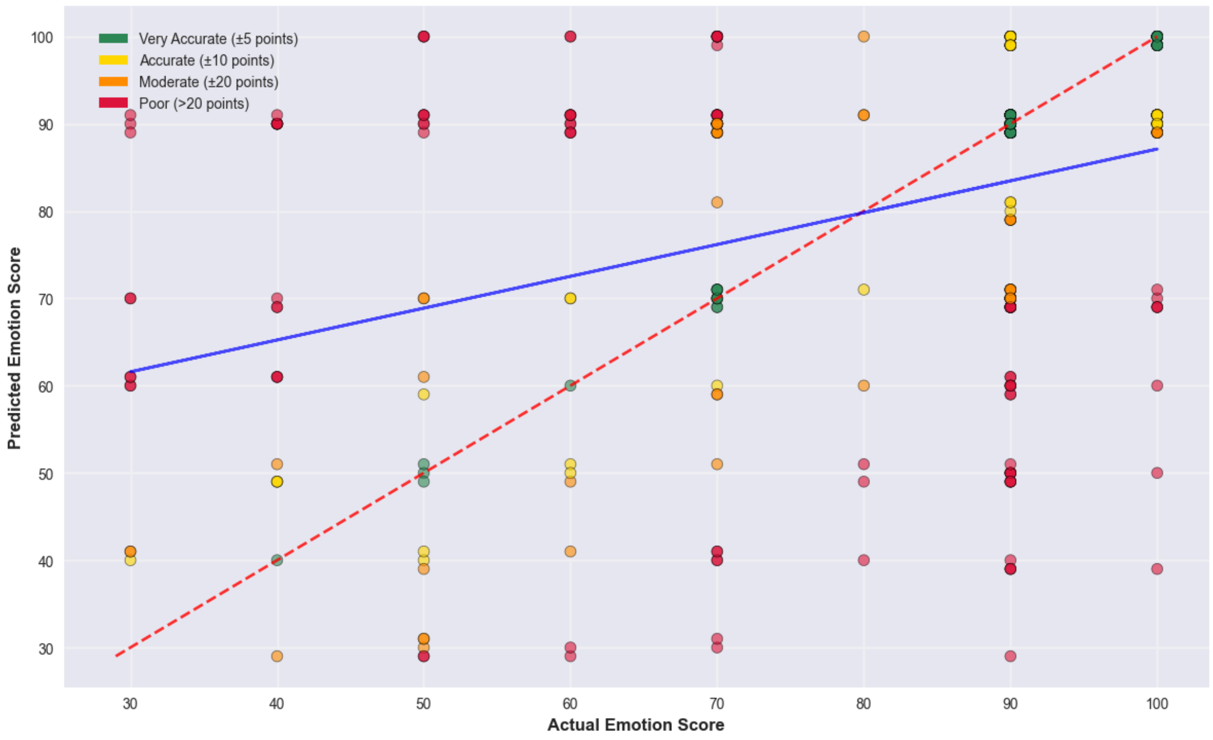
- Longest positive streak (≥8.0/10): 47 consecutive days
  Success rate: 80.4% of days scored ≥8.0/10
  Recovery patterns: Rapid bounce-back from challenging days

3. Activity-Wellbeing Correlations

- Prayer/Spiritual practices: +0.82 correlation
- Creative activities (art, writing): +0.64 correlation
- Learning activities: +0.58 correlation
- Physical wellness activities: +0.52 correlation

FORECAST METHODS

We employed OpenAI gtp-4o-mini LLM API to predict 7-day forward emotional states from historical mood tracking data derived from the ‘Daylio Mood Tracker’ dataset. Each prediction analyzed synthetically generated daily summaries (≤250 characters) based on the input provided within the dataset and containing mood ratings, activities, and reflections, along with the previous week's entries for temporal context.



Technical Architecture

Input: Target date summary + 4-day historical window
Processing: 376 asynchronous predictions via Convex database functions
Output: Emotion score (0-100) with [1]fixed 85% confidence level

Performance Results

The system achieved modest but statistically significant predictive capability ($r = 0.369$, $p < 0.05$):

- 52.2% of predictions within ±10 points
- 15.0 mean absolute error
- No systematic bias (mean error ≈ 0)

KEY FINDINGS

The model excels at predicting stable high-performing periods but struggles with emotional volatility and sudden mood shifts. Perfect predictions occurred during consistent behavioral patterns, while errors up to 61 points emerged during abrupt emotional transitions.

Performance varied significantly by emotional state range:

- Excellent states (91-100): 9.6 point average error
- Low states (30-50): 18.3 point average error

Limitations

This proof-of-concept demonstrates LLM viability for emotional forecasting while highlighting the need for dynamic confidence scoring and enhanced temporal modeling as follow:

- Fixed 85% confidence regardless of prediction uncertainty
- Limited 7-day context window
- Equal weighting of all activities despite varying correlations (prayer: +0.82, entertainment: -0.15)
- Compression toward mean values

FORECAST PERFORMANCE REVIEW

The AI forecasting system shows moderate overall performance with a correlation coefficient of 0.577 between predicted and actual scores:

- Mean Absolute Error (MAE): 13.2 points (out of 100)
- Overall accuracy: 65.5% of predictions fall within ±10 points
- Very accurate predictions: 32.8% fall within ±5 points
- Poor predictions: 11.5% have errors exceeding 20 points

The analysis reveals that perfect or near-perfect predictions occur when:

- Stable patterns: The subject maintains consistent routines and activities
- High baseline scores: When actual scores are in the 80-90 range (excellent days), predictions are more accurate
- Momentum effects: During positive streaks, the AI correctly identifies continuation patterns
- The best prediction example showed only 0.1 points error, demonstrating the model's capability when conditions are optimal.

[1] Fixed confidence issue: All predictions use 85% confidence regardless of certainty, meaning for the purpose of this study the model doesn't adapt its confidence based on pattern stability.

The worst predictions occur due to:

- Sudden mood shifts: The model struggles with abrupt changes in emotional state

The model performance range:

- Low scores (30-50): Average error of 18.3 points
- Medium scores (51-70): Average error of 16.8 points
- High scores (71-90): Average error of 11.8 points
- Excellent scores (91-100): Average error of 9.6 points

KEY INSIGHTS FOR PREDICTION METHODS

The model excels at predicting high-scoring days (excellent emotional states). Shows consistent slight under-prediction bias (-1.8 points average). Performs best when the subjects maintain regular patterns.

The model struggles with poor adaptability to volatility, and cannot handle sudden emotional changes with high accuracy. The model may be limited from a shortened context understanding, and may not fully capture the impact of specific activities or events without more context. The model tends to predict toward the mean, missing extreme values, or potentially lowering the accuracy of less-predictable events.