

# The Utility of LLM Text Generation in Longitudinal Psychological Datasets

Jari Zegers<sup>1</sup>, Bennett Kleinberg<sup>1,2</sup>

<sup>1</sup>Tilburg University

<sup>2</sup>University College London

[j.zegers@tilburguniversity.edu](mailto:j.zegers@tilburguniversity.edu), [bennett.kleinberg@tilburguniversity.edu](mailto:bennett.kleinberg@tilburguniversity.edu)

## Introduction

Natural Language Processing (NLP) techniques provide enormous potential to study human behavior (Boyd & Schwartz, 2021; Feuerriegel et al., 2025). The introduction of large language models (LLMs) has sparked a new wave of interest in textual data for psychological research. However, one of the methodological challenges that come with an increased adoption of LLMs pertains to their utility in understanding human written texts. One potential application in longitudinal datasets is to use LLMs to quantify the predictability of human written texts in later waves.

In this ongoing project, we use a multi-year panel dataset of rich narratives on the emotional responses during the COVID-19 pandemic to prompt an LLM with temporal sequences of human-written narratives and instruct the model to generate the textual response of a subsequent measurement moment. We then assess how similar the generations of the LLM are compared to “true” human data. In doing so, we assess the model’s ability to incorporate relevant information from previous waves to generate a new text. As a secondary aim, we then examine whether the degree to which the generated texts correspond to the ground truth texts can be quantified as a variable of “psychological surprise”. As we believe LLMs will generate texts that continue the narrative trajectory of previous waves, strong deviation from what is generated could indicate changes in psychological functioning, while ease of generation could indicate a rather stable trajectory through the pandemic. We formally test this idea by correlating changes in reported emotions of the participants with the generation discrepancy.

## Aims

The aims of this project are therefore as follows: First, we investigate whether LLM-generated texts for future measurement moments are more similar to ground truth texts than would be expected at chance level. Second, we examine whether the similarity between text pairs (generated and ground truth) is associated with relevant psychological variables. Third, we examine how generated texts differ from ground truth texts using embedding-based topic modelling.

## Methods

### Real World Worries Dataset

We used a panel dataset on the COVID-19 pandemic – “The Real-World Worries” dataset (RWW; Van Der Vegt & Kleinberg, 2023). Data were collected in April of 2020 ( $n=2441$ ), 2021 ( $n=1716$ ), 2022 ( $n=1152$ ), and 2023 ( $n=868$ ) and included Likert scale ratings of eight emotions and free text responses where participants expressed how they felt about the pandemic at the time of data collection (see Figure 1 for an overview). The mean number of tokens across all waves was 123.37 ( $SD= 32.20$ ).

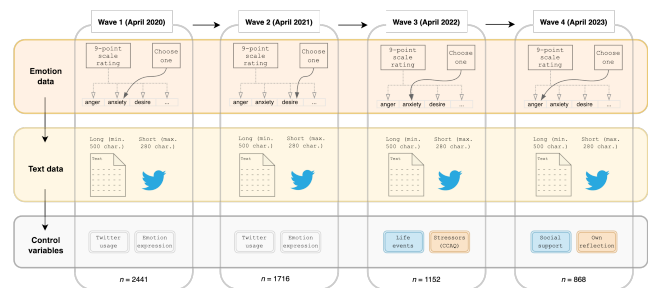


Figure 1. Data collection procedure across all waves.

### Text Generation via LLM prompts

Texts were generated from the gpt-4o model through the OpenAI API (OpenAI, 2024). Responses were generated with a temperature setting of 0.7 and a maximum response length 10% greater than the ground truth wave 4 text. The model was provided the following prompt:

"What would this person write in 2023? IMPORTANT: Please make sure to finish your response with a full sentence. 2020: [insert text from 2020]  
2021: [insert text from 2021]  
2022: [insert text from 2022]  
2023:"

## Measuring Generation Discrepancy

Embeddings for both the generated and ground truth wave 4 texts were obtained using OpenAI's model "text-embedding-3-large" with 256 dimensions. Next, we calculated the cosine similarities between the embeddings of the LLM generated text and ground truth wave 4 texts for each participant. To test whether the mean cosine similarity was greater than would be expected at chance level, a permutation approach was used. We created a null distribution by randomly pairing generated and true texts 10,000 times and calculating the mean cosine similarity for each permutation. The proportion of permutation means higher than the sample mean is analogous to the p-value. A statistical significance threshold of alpha of 0.05 was used. Methods for the generation of texts and carrying out the permutation test were pre-registered here: <https://aspredicted.org/qzkg-r2v8.pdf>.

## Association with Psychological Variables

Using the cosine similarity for all ground truth-generated text pairs, we investigated whether text similarity was associated with psychological variables in the dataset using three approaches.

First, we calculated the Euclidian distance between emotion vectors (i.e., an eight-dimensional vector where each element corresponds to a participant's score for an emotion at a particular wave) at wave 3 and wave 4 for each participant. High distances indicate greater changes in emotion scores between waves. We calculated the Spearman correlation between the Euclidian distance and text similarity.

Second, we regressed emotion scores on waves 1 to 3 for each participant and predicted wave 4 scores. The absolute prediction error for wave 4 was averaged across emotions per participant. The rationale of that procedure is that on the resulting measure (the emotion trend error) those who deviate more strongly from their trend in previous waves will obtain higher scores. This procedure mirrors conceptually the procedure used to generate texts (i.e., using waves 1-3 to predict wave 4). We computed Spearman correlation between the emotion trend error and text similarity. Negative correlations would suggest that individuals with more volatile emotion scores show higher discrepancy between generated and ground truth texts.

Third, previous work has used latent class trajectory analysis to model latent classes in the trajectories of emotion scores (Zegers & Kleinberg, 2024). That work concluded that there were important differences in the emotion trajectories that could be captured in the form of six latent classes. We tested whether class membership was associated with cosine similarity using an ANOVA. A Bonferroni corrected alpha was used for significance testing.

## Topics

We used embeddings-based topic modelling on the joint corpus of ground truth and generated texts for wave 4. For each text, we obtained the embeddings (GPT "text-embedding-3-large", 256 dimensions), used a UMAP dimensionality reduction to ten dimensions and then applied density-based clustering on the UMAP representations. Specifically, we used DBSCAN with a maximum of 10% of the texts labelled as topic outliers. The resulting topic model produced nine topics that were represented as n-grams with a TF-IDF weighting by topic. The top terms were used to label the topics. We then used a Chi-square test to examine whether topic membership was statistically associated with the text being generated or a ground truth human-written text.

## Results

### Similarity between Generated and Ground Truth Texts

The permutation test indicated that the mean cosine similarity between pairs of generated and ground truth wave 4 texts was significantly higher than expected at chance level ( $M=0.558$ ,  $SD=0.09$ ,  $p<.001$ ). The LLM was thus able to use texts from the first three waves to generate the fourth.

### Associations between Psychological variables and Text Similarity

There was no significant association between text similarity and the Euclidian distance between wave 3 and 4 emotion vectors,  $r(860)=-0.06$ ,  $p=.065$ . However, there was a significant negative correlation between text similarity and emotion trend error,  $r(860)=-0.09$ ,  $p=.010$ . This represents a small effect size where 0.81% of the variance in text similarity could be explained due to emotion trend error scores. There was no significant effect of class membership on text similarity,  $F(1, 860)=1.43$ ,  $p=.232$ . There is thus limited evidence for the mapping of text-pair similarity onto psychological variables.

### Topic Differences

There was a significant association between topic and text origin (generated vs ground truth),  $\chi^2(10)=1638.3$ ,  $p<.001$ . The standardized residuals indicated the association was driven by ground truth texts being overrepresented as outliers and in a topics that express complex emotions and relate to hygiene protocols. Conversely, LLM-generated texts were overrepresented in topics about an optimistic outlook and self-reflection about the pandemic. These findings reveal substantial differences in how the LLM completed the

text for 2023 the ground truth texts from 2020-2022, compared to how what participants actually wrote and how the pandemic unfolded for them.

## Discussion

As part of this ongoing work, we prompted an LLM with three waves of texts from a longitudinal panel dataset to generate a text for wave 4. We compared generated to ground truth texts using cosine similarity on embeddings and tested whether text similarity was associated with psychological variables. We found limited evidence for an association but do find differences in the topics used in generated versus ground-truth texts. An explanation for differences in text similarities remains the subject of ongoing investigation.

## References

- Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, 40(1), 21–41. <https://doi.org/10.1177/0261927X20967028>
- Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 4(2), 96–111. <https://doi.org/10.1038/s44159-024-00392-z>
- GPT-4o*. (2024). [Computer software]. OpenAI. <https://openai.com>
- Van Der Vegt, I., & Kleinberg, B. (2023). A multi-modal panel dataset to understand the psychological impact of the pandemic. *Scientific Data*, 10(1), 537. <https://doi.org/10.1038/s41597-023-02438-y>
- Zegers, J., & Kleinberg, B. (2024, October). *Beyond the sample: Individual differences in psychological responses to the COVID-19 pandemic*. PsyArXiv. <https://doi.org/10.31234/osf.io/2px7h>