Well-Being Assessment Using ChatGPT-4: A Zero-Shot Learning Approach

Julianne Vizmanos
De La Salle University
Manila, Philippines
julianne_vizmanos@dlsu.edu.ph

Ethel Ong
De La Salle University
Manila, Philippines
ethel.ong@dlsu.edu.ph

Jackylyn Beredo
De La Salle University
Manila, Philippines
jackylyn.beredo@dlsu.edu.ph

Remedios Moog
De La Salle University
Manila, Philippines
remedios.moog@dlsu.edu.ph

ABSTRACT

Traditional approaches of using self-report questionnaires and emotion-based lexicon pose limitations in assessing the well-being states from dialogue utterances which consequently impact the generation of appropriate empathetic responses. The development of ChatGPT unveiled the potential of applying large language models in various domain of text understanding tasks, including well-being assessment. In this paper, we present our investigation of using ChatGPT-4 to measure well-being based on Seligman's PERMA model. The International Survey on Emotion Antecedents and Reactions (ISEAR) dataset was manually annotated with the elements of PERMA. Zero-shot ChatGPT-4 is then employed to label the dataset across five well-being states: *excelling, thriving, surviving, struggling,* and *in crisis*. In the absence of a reference gold standard to serve as baseline, we compared our results with those produced from using the PERMA Lexicon. Because there is ambiguity in the boundary between neighboring well-being states, we reduced the labels to three with *excelling, thriving and surviving* comprising one state, while *struggling* and *in crisis* remain as separate states. Results from applying the intercoder agreement yielded 37.22%, 39.61%, and 50.11%, respectively. Our findings highlight the challenges of automating this inherently subjective task without a PERMA-labeled dataset to serve as a basis for ground truth and lays the groundwork in employing ChatGPT-4 for psychological well-being assessment.

KEYWORDS

Well-being assessment, PERMA model, PERMA lexicon, ChatGPT, zero-shot learning

1 INTRODUCTION

Psychological well-being plays an important role in a person's mental health which can impact our emotions, relationships, productivity, and overall satisfaction with life. Seligman's PERMA model [15] is one of the most influential models in psychological well-being that measures its five core elements, namely positive emotion (P), engagement (E), relationships (R), meaning (M), and accomplishments (A). Enhancing each of these five (5) elements can enable an individual to achieve a more fulfilled and satisfied life.

Psychological well-being assessment is usually measured using self-report questionnaires such as the PERMA Profiler [3]. This instrument contains 15 questions that covers the five (5) elements of PERMA and 8 questions that focus on overall health, negative emotion, happiness, and loneliness. Prior studies have reported its reliability in measuring well-being [4, 7]. However, these tools are resource-intensive and pose challenges in scalability.

With the availability of conversational agents or chatbots, researchers began exploring the use of these technologies in delivering mental health and well-being support services [6, 8–10]. Most of these chatbots are designed primarily to generate appropriate empathetic responses, but very few works have focused on measuring the support seeker's well-being. The PERMA Lexicon is an attempt to automate well-being measurement [2, 14] but faced some shortcomings. These lexicon-based approaches heavily rely on a pre-defined set of words with their corresponding PERMA score and may lead to inaccuracies if a word to be processed is not found in the dictionary. Moreover, lexicons are limited with their inability to understand context and handle figurative languages such as irony and sarcasm [1].

The emergence of large language models (LLMs) such as GPT-3.5, GPT-4 [11], and LLaMa [16] offered potential benefits in healthcare applications through the generation of human-like responses that are coherent and textually relevant to the user's prompts pose. Developed by OpenAI, ChatGPT is trained on GPT-3.5 using the Reinforcement Learning from Human Feedback (RLHF) technique to align its responses to humans [12]. Research work are also exploring its application in sentiment analysis [1, 18, 19] and emotion detection [17]. To perform such NLP tasks, models are trained with annotated datasets. The release of ChatGPT has opened a new area of research that focuses on employing these models in zero-shot learning - that is, without additional task-specific training.

In this paper, we describe our investigation of using ChatGPT-4 in measuring well-being based on Seligman's PERMA Model [15]. We manually annotated the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset [13] with the five (5) elements of PERMA. Zero-shot ChatGPT-4 is then employed to label the dataset across five well-being states: *excelling, thriving, surviving, struggling,* and *in crisis*. Our main contribution in this paper is to provide an evaluation of ChatGPT-4's performance in well-being assessment by comparing it with the results of the PERMA Lexicon [2]. Our findings highlight the challenges of automating an inherently subjective task without a PERMA-labeled dataset to serve as a basis for ground truth, and lays the groundwork in employing ChatGPT-4 for psychological well-being assessment.

2 TASK DESCRIPTION

In this section, we provide a computational definition of the well-being assessment task, followed by the process of computing the PERMA scores using the PERMA Lexicon. We include a discussion of how we formulated the prompts to instruct ChatGPT to perform the required task.

2.1 Well-being Assessment

We formulate the well-being assessment task as a text classification problem similar to the approach by MHBot and VHope [2, 10]. Given an input utterance **u**, the well-being assessment annotates the utterance **u** to one of the predefined set **L** of well-being states where **L = {excelling, thriving, surviving, struggling, in crisis}**. Thus, the well-being assessment is a function $f : U \longrightarrow L$ that annotates each utterance $u \in U$ with a label $l \in L$ that best represents the well-being state of the utterance, where **U** is all the input utterances and **L** is all the possible well-being states.

2.2 PERMA Lexicon

The PERMA Lexicon is a collection of words with positive and negative scores representing the elements of PERMA. This dataset was used by the MHBot [10] and the VHope [2] chatbots to measure the well-being of an individual based on the lexical choices found in their messages, social media posts, or utterances. The formula shown in Equation 1 computes the total positive and negative PERMA score of a given input text.

Given an input text or utterance $\boldsymbol{u}$, the average *PERMA weight* is first computed from the sum of the *PERMA_weight(w_i)* for each token w_i in $\boldsymbol{u}$ divided by the total number of tokens $\boldsymbol{n}$. For each PERMA element, denoted by category $\boldsymbol{c}$, the lowest score min_c is subtracted from the resulting average *PERMA_weight* and then multiplied by 10, which corresponds to the total number of categories - 5 positive and 5 negative PERMA elements. The product is then divided with the difference of the category's max_c and min_c values as indicated in Table 1. From this process, each of the 10 categories yields a score that is totalled by group (*pos* and *neg*), leading to the final positive and negative PERMA score.

$$PERMA\_score(pos, neg) = \sum_{c=p}^{a} \frac{\left(\frac{\sum_{i=1}^{n} PERMA\_weight(w_i)}{n} - \mathrm{min}_c \right) * 10}{max_c - min_c} \tag{1}$$

where,

$PERMA\_score$ = total well-being score
pos = positive score
neg = negative score
w = tokens in the input
c = p, e, r, m, a (positive and negative)
n = total number of tokens in the input
min = minimum score of the category
max = maximum score of the category

Table 1: Minimum and maximum scores of each Category representing the positive and negative PERMA elements

	Minimum Score	Maximum Score	Mean
POS_P	-0.36639	0.76549	0.04172
POS_E	-0.30074	0.34065	0.03234
POS_R	-0.28884	0.78376	0.03824
POS_M	-0.16748	0.77167	0.02517
POS_A	-0.19784	0.55031	0.03990
NEG_P	-0.32731	0.70697	0.04705
NEG_E	-0.15230	0.84017	0.04354
NEG_R	-0.28648	0.62033	0.04045
NEG_M	-0.14987	0.31674	0.03416
NEG_A	-0.15369	0.24760	0.03426

The *PERMA_score* is then interpreted using the labels from the PERMA Profiler [3]: *very high functioning*, *high functioning*, *normal functioning*, *sub-optimal functioning*, and *languishing*. These labels may be vague to users of VHope, thus, they were renamed following the Mental Health Continuum [5] phases to support the belief that an individual's mental health is not binary but a continuously changing state. The revised labels are indicated in Table 2.

Table 2: PERMA Score Interpreter

Label	Positive Score	Negative Score
Excelling	7 and above	0 to 1
Thriving	6 to 6.9	1.1 to 2.5
Surviving	4.5 to 5.9	2.6 to 3.9
Struggling	3 to 4.4	4 to 4.9
In crisis	below 3	above 5

Initial testing of VHope [2] using the interpretation suggested by [3] showed inaccuracy in labeling the utterances based on the computed well-being score. The values were adjusted as shown in Table 2 and used during VHope's user testing phase. The accuracy was again re-assessed by comparing the user's computed well-being score from Equation 1 and their score derived manually from the self-report questionnaire. Out of the 21 well-being levels, the computed and the manually derived well-being scores agree on 12 well-being levels, or 57% accuracy.

Guidance counselors also reviewed the PERMA labels assigned to utterances. Of the 97 PERMA labels extracted from 43 conversation logs, only 57 labels or 59% were noted as appropriate. But even with the low accuracy, the counselors noted that the PERMA labels assigned by VHope were able to dynamically adapt to the user's changing well-being state throughout a conversation. This makes the PERMA Lexicon a sufficient basis for comparing the well-being assessment generated by ChatGPT.

2.3 Prompt Formulation

We adapted the prescribed prompt formulation template defined by OpenAI[1] for the well-being assessment task to indicate the *role* to be portrayed by the LLM, the *task* or instruction to be performed, the *input* text, and the target *labels*.

[1] https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

Following the prompt engineering strategies, the formulated prompt designates a role for the LLM:

> *You are a PERMA expert.*

specifies the task to be performed:

> *Your task is to categorize the sentences into the five zones of The Mental Health Continuum by Delphis: Excelling, Thriving, Surviving, Struggling, and In Crisis.*

and defines the target labels from [5]. The desired length of the output is also specified in the prompt:

> *Given the explanation for each category, output the category only. Number each row, but NO explanation.*

3 METHOD

We cleaned the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset by removing special characters and duplicate entries. Nuisance rows such as those containing variants of *no*, *not applicable*, *no description*, and *nothing* were also removed. The cleaned dataset yielded 7,475 rows. The emotional responses recorded in the *content* column of the data is utilized in this study.

The PERMA Lexicon is employed to annotate the ISEAR dataset with the labels: *excelling, thriving, surviving, struggling,* and *in crisis*. Each entry as tokenized to derive the corresponding PERMA weights. Tokens without associated weights in the lexicon are assigned with a value of zero. The overall PERMA well-being score is then computed for each entry and assigned the corresponding label indicated in Table 2.

Zero-shot ChatGPT-4 is employed to annotate the ISEAR dataset. In the zero-shot setting, the pre-trained model is not provided with any additional task-specific training. There was a noticeable significant variance in the execution time between the web interface and API of ChatGPT-4. Because of this, the web interface of zero-shot ChatGPT-4 is employed to annotate the ISEAR dataset.

Without a reference gold standard to serve as the ground truth, we utilized the Intercoder Agreement to measure the consistency of annotation labels between the PERMA Lexicon and ChatGPT-4. We deemed it inappropriate to use the annotations derived from the PERMA Lexicon as the ground truth since VHope reported achieving only 57% accuracy when using this approach [2].

4 PRELIMINARY RESULTS

In consultation with a guidance counselor specializing in PERMA, we performed three (3) comparative analyses of our results by clustering the PERMA labels as shown in Table 3. The **5-Label Analysis** compares the performance of ChatGPT with that of the PERMA Lexicon by looking at each of the labels independently. The **4-Label Analysis** explores the influence of aggregating the neighboring labels *excelling* and *thriving* to the performance of the models in well-being assessment. Lastly, the **3-Label Analysis** extends the aggregation of neighboring labels to include *surviving* with *excelling* and *thriving*. Because well-being assessment is highly subjective, our approach addresses the ambiguity in the boundary between neighboring well-being states.

Table 4 presents the annotation results derived using the PERMA Lexicon and ChatGPT-4. The matrix highlights the common annotation labels along the main diagonal line. This serves as the basis for comparing the performance of the two approaches using the 5-Label, 4-Label, and 3-Label Analyses.

Table 3: Comparative Analyses

	Labels				
5-Label	excelling	thriving	surviving	struggling	in crisis
4-Label	excelling + thriving		surviving	struggling	in crisis
3-Label	excelling + thriving + surviving			struggling	in crisis

4.1 5-Label Analysis

As shown in Figure 1, *Struggling* has the highest percent agreement at 43.99%, followed by *surviving* at 41.47% and *excelling* at 34.07%. The *thriving* and *in-crisis* labels have the lowest, at 12.77% and 1.67% agreement, respectively. Overall, there is a 37.22% intercoder agreement in the 5-label analysis.

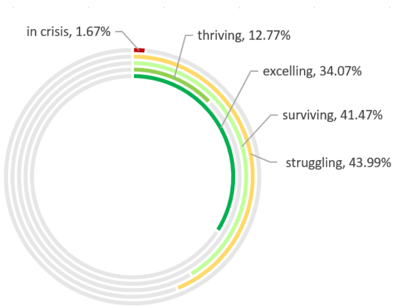

Figure 1: Intercoder Agreement for each well-being label in the 5-Label Analysis.

Surviving and *struggling* have the highest contribution to the overall intercoder agreement at 55.10% and 39.22%, respectively. These values suggest that the labels are not only prevalent in the dataset but that both the PERMA Lexicon and ChatGPT-4 performed well in labeling instances of these states. In the case of *excelling*, although the intercoder agreement is significant, its contribution to the overall percentage agreement is only 2.23%, indicating that both approaches tend to agree in labeling this state despite its low occurrence in the dataset.

The high instances of intercoder disagreements in general, and for the labels *excelling*, *thriving*, and *surviving* in particular, can be attributed to the subjective nature of well-being assessments and the fuzzy boundary between neighboring labels. Looking at Table 4,

Table 4: Annotation matrix for the Well-being labels derived using the PERMA Lexicon and ChatGPT-4.

		ChatGPT-4				
		excelling	thriving	surviving	struggling	in crisis
Lexicon	excelling	62	54	33	32	1
	thriving	125	89	230	232	21
	surviving	258	264	1533	1519	123
	struggling	83	156	1052	1091	98
	in crisis	7	24	205	176	7

there is a 63.20% disagreement rate between the neighboring labels. The occurrence of classification biases is similar to how different psychologists may give different labels to an individual's well-being due to how they interpret a situation.

It is observed that intercoder disagreements typically occur in utterances with contradicting statements. The utterance *"I had lied to a person because I had thought that I could not tell him the truth. When he found out he was not angry but understanding. We talked the whole thing over"* presents a situation wherein an individual lied to a person because he thought that the truth cannot be shared. The person did not explode in fits of anger upon discovery of the truth, but was rather understanding in discussing and resolving this issue. This complex narrative is a challenge for the labeling task as the initial deceit could suggest a *crisis* while the resolution leans towards *survival.* As such, an utterance encompasses a wide range of emotions and states of well-being - ultimately relying on the coders to decide on a label that best represents the overall state of well-being.

4.2 The 4-Label Analysis

As observed in Figure 2, *Struggling* still has the highest agreement rate at 43.99%, followed by *surviving* at 41.47%. Meanwhile, the combined *excelling* and *thriving* label has achieved a 37.54% agreement, and *in crisis* label remained the lowest at 1.67%. Overall, there is a 39.61% intercoder agreement for the 4-label analysis.

Surviving and *struggling* have a significant impact to the overall intercoder agreement at 51.77% and 36.85%, respectively. This suggests a strong consensus between PERMA Lexicon and ChatGPT-4 in assigning these labels because of clearer distinction between these well-being states. The consolidation of *excelling* and *thriving* achieved 37.54% intercoder agreement and contributed 11.14% to the overall agreement rate, implying that closely-related positive states may be easily agreed upon by both annotators as opposed to its finer states. Its consolidation also lowered the disagreement rate between the neighboring labels to 59.29%.

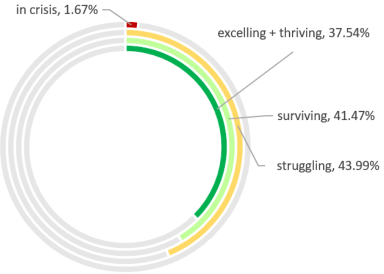

Figure 2: Intercoder Agreement for each well-being label in the 4-Label Analysis

4.3 The 3-Label Analysis

The aggregation of *excelling*, *thriving*, and *surviving* label achieved the highest percent agreement at 57.87%. This is followed by *struggling* at 43.99%. *In crisis* remains at 1.67% as exemplified in Figure 3. Overall, there is a 50.11% agreement rate for the 3-label analysis.

The combination of *excelling*, *thriving*, and *surviving* label recorded a 57.87% intercoder agreement and significantly influenced the overall agreement at 70.69%. This could be attributed to the more evident boundaries between the labels and reinforces that positive states of well-being is distinct as opposed to its negative counterpart. However, the high disagreement rate for *in crisis* at 98.33% suggests that the PERMA Lexicon and the language model might need further refinement, or the threshold employed by the PERMA Lexicon should be re-examined.

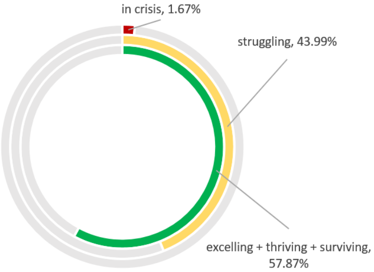

Figure 3: Intercoder Agreement for each well-being label in the 3-Label Analysis.

4.4 Overall Analysis

The 5-Label Analysis reported the lowest intercoder agreement at 37.22%. However, an increase of 2.39% is observed for the 4-Label Analysis, resulting to 39.61%. Ultimately, the 3-Label Analysis yielded the highest intercoder agreement of 50.11% with a significant boost of 12.89%. This suggests that the subtle nature of well-being states are more challenging to classify because of their overlapping characteristics, but the aggregation of labels exemplified that the states of well-being are more distinct when considered in a wider scope.

Collectively, the analyses highlighted the complexity in classifying the continuous nature of well-being - reinforcing its subjectivity. As such, it is observed that classifying general well-being states gained significant results as opposed to distinguishing between its varying levels. This alludes to the blurred boundaries between its varying levels as opposed to its clearer limits when consolidated with states sharing similar characteristics. It also implies that the PERMA Lexicon or the ChatGPT-4 model may need further development, and the threshold for PERMA Lexicon's classification may need to be re-assessed.

5 FURTHER WORK

The insights gained from this study can contribute to the expanding knowledge of psychology in Natural Language Processing (NLP); as well as contribute to the exploration of automated well-being assessment tools through LLMs. Future works should examine ChatGPT-4 in one-shot and few-shot settings to determine the influence of the setting on the intercoder agreement rate. Balancing of the distribution of instances for each well-being state may also improve the performance of LLMs. Alternative metrics that consider the subjective nature of well-being assessment can also be explored to provide additional insights to the model's efficacy in this classification task.

REFERENCES

[1] Mohammad Belal, James She, and Simon Wong. 2023. Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis. arXiv:2306.17177 [cs.CL]

[2] Jackylyn L. Beredo and Ethel Ong. 2022. Analyzing the Capabilities of a Hybrid Response Generation Model for an Empathetic Conversational Agent. *International Journal of Asian Language Processing* 32, 4 (2022). https://doi.org/10.1142/S271755452350008X

[3] Julie Butler and Margaret L. Kern. 2016. The PERMA-Profiler: A Brief Multidimensional Measure of Flourishing. *International Journal of Wellbeing* 6, 3 (2016).

[4] Thainá Ferraz de Carvalho, Sibele Dias de Aquino, and Jean Carlo Natividade. 2021. Flourishing in the Brazilian Context: Evidence of the Validity of the PERMA-Profiler Scale. *Current Psychology* 42 (2021), 1828—-1840.

[5] Delphis. 2020. The Mental Health Continuum is a Better Model for Mental Health. https://delphis.org.uk/mental-health/continuum-mental-health/.

[6] Vanshika Gupta, Varun Joshi, Akshat Jain, and Inakshi Garg. 2023. Chatbot for Mental health support using NLP. In *Proceedings of the 2023 4th International Conference for Emerging Technology (INCET)*. https://doi.org/10.1109/INCET57972.2023.10170573

[7] Margaret L. Kern, Lea E. Waters, Alejandro Adler, and Mathew A. White. 2015. A Multidimensional Approach to Measuring Well-being in Students: Application of the PERMA Framework. *Journal of Positive Psychology* 10, 3 (2015), 262–271.

[8] Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. 2018. Can a Machine Tend to Teenagers' Emotional Needs? A Study with Conversational Agents. In *Proceedings (ACM) Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Montreal, Canada, 1–6.

[9] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure Through a Chatbot. In *Proc, (ACM) 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Hawaii, USA, 1–12.

[10] Ethel Ong, Melody Joy Go, Rebecalyn Lao, Jaime Pastor, and Lenard Balwin To. 2024. Investigating Shared Storytelling with a Chatbot as an Approach in Assessing and Maintaining Positive Mental Well-Being among Students. *International Journal of Asian Language Processing* 33, 3 (2024). https://doi.org/10.1142/S2717554523500170

[11] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[12] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[13] Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology* 66, 2 (1994), 310.

[14] H. Andrew Schwartz, Maarten Sap, Margaret L. Kern, Johannes C. Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, Michal Kosinski, Martin E.P. Seligman, and Lyle H. Ungar. 2016. Predicting Individual Well-being through the Language of Social Media. In *Biocomputing 2016: Proceedings of the pacific symposium*. World Scientific, 516–527.

[15] Martin Seligman. 2010. Flourish: Positive Psychology and Positive Interventions. *The Tanner Lectures on Human Values* 31, 4 (2010), 1–56.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[17] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Bias in Emotion Recognition with ChatGPT. arXiv:2310.11753 [cs.RO]

[18] Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. arXiv:2304.04339 [cs.CL]

[19] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. arXiv:2304.10145 [cs.AI]