Exploring Large Language Models for PERMA-based Psychological Well-being Assessment

Julianne Andrea Vizmanos, Ethel Ong

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Exploring Large Language Models for PERMA-based Psychological Well-being Assessment

Julianne Andrea Vizmanos and Ethel Ong
College of Computer Studies
De La Salle University
Manila, 1004 Philippines
julianne_vizmanos@dlsu.edu.ph, ethel.ong@dlsu.edu.ph

Abstract

This paper explores the potential of leveraging Large Language Models (LLMs), specifically ChatGPT-4, LLaMa 3-8B, and Gemini-1.5-pro, in PERMA-based psychological well-being assessment. Utilizing the ISEAR dataset, 7,431 utterances were processed then classified into the five well-being states: *excelling*, *thriving*, *surviving*, *struggling*, and *in-crisis*. In the absence of a ground truth, intercoder agreement was applied as the metric to compare the performance of the LLMs with one another and with the rule-based PERMA lexicon. Analysis of the results revealed that 9.45% of the dataset showed no agreement among the LLMs, 60.93% showed partial agreement, and 29.62% showed full agreement. The mode of the LLMs's labels then served as the standard for comparison, resulting in an intercoder agreement of 32.54% for PERMA lexicon, 72.86% for ChatGPT, 78.95% for Gemini, and 68.36% for LLaMa. These findings highlight that while the LLMs demonstrate substantial agreement, the discrepancies unveil the challenges in capturing nuanced emotional expressions - necessitating further refinements to enhance the LLMs' accuracy and reliability in psychological well-being assessments.

1 Introduction

Mental health is a state of well-being that exists on a complex continuum and can vary greatly among individuals (Gautam et al., 2024). Albeit fundamental aspect of overall well-being, it remains one of the leading global health challenges not only from the after effects of the COVID-19 pandemic (Duden et al., 2022), but also everyday stressors. If left unmanaged, this psychological distress can lead to lower quality of life, unrealized potentials, poor academic and work performance, and negative emotions. As such, the importance of proper detection and management of psychological well-being has grown significantly in recent years.

Emotional expression is the process of conveying one's emotions through verbal or non-verbal manner. It is a complex indicator of one's mental state and integral to psychological well-being. A study by Pennebaker (1997) revealed the importance of emotional expression in reducing psychological distress. Expressing emotions effectively can act as a coping mechanism that lowers stress levels, reduces depressive symptoms, improves mental health, and enhances psychological well-being. Conversely, emotional suppression is the inhibition of emotional expression. It is linked to lower levels of well-being and higher levels of depression and anxiety (Gross and John, 2003). Barrett et al. (2011), however, argue that emotional expressions are ambiguous as they can vary significantly depending on the context, individual differences, and cultural background.

The emergence of large language models (LLMs) with the ability to understand and generate fluent human language enables them to respond dynamically and coherently to a user's prompts. Some LLMs are also equipped with user-friendly interfaces and conversational capabilities that enable them to function as empathetic chatbots with applications in mental healthcare. These studies are devoted to building empathetic language models capable of understanding human emotions through language analysis (Shin et al., 2019; Zhou et al., 2020) and generating empathetic responses (Lee et al., 2022; Lin et al., 2020; Morris et al., 2018) in order to offer individualized emotional support. However, while emotion detection is a necessary component in generating empathetic responses, it is only one of the five dimensions that comprise an individual's mental health and well-being.

The PERMA model, proposed by Seligman (2010), is a psychological framework aimed at understanding well-being through its five dimensions: Positive Emotions (P), Engagement (E), Relationships (R), Meaning (M), and Accomplishment (A).

This can provide a better assessment of an individual's flourishing state. PERMA emphasizes that to be flourishing does not merely mean the absence of mental illness but the presence and sustained cultivation of positive states that contribute to long-term well-being. Moreover, unlike models that focus on a single aspect of well-being, PERMA recognizes that well-being is multi-faceted; thus, capturing multiple dimensions that are essential for overall well-being. Even though it is a holistic model, the use of PERMA for well-being detection and assessment has not been extensively explored in NLP research. Moreover, while there are publicly available datasets commonly used for emotion and stress detection, there is none for PERMA well-being assessment.

LLMs are capable of language comprehension, contextual understanding, and scalability that traditional machine learning models fall short of. Studies have also demonstrated the abilities of LLMs to perform annotations on textual data (Pangakis et al., 2023). However, LLMs are still limited in fully understanding nuanced human emotions. As such, Zhang et al. (2024) built the Agent for STICKERCONV (Agent4SC) to account for the limited abilities of LLMs in performing empathetic annotations.

In this paper, we describe our experiments in leveraging multiple LLMs, specifically ChatGPT-4 (OpenAI, 2023), LLaMa 3-8B (Touvron et al., 2023), and Gemini-1.5-pro (Team, 2024) for PERMA well-being assessment. Our study makes the following contributions:

1. Application of Seligman's PERMA model in psychological well-being assessment;
2. Comparison of the performance of ChatGPT-4, LLaMa 3-8B, and Gemini-1.5-pro in PERMA well-being assessment; and,
3. Utilization of intercoder agreement to derive the ground truth which can be used to label existing datasets with PERMA.

2 Related Works

Early works in well-being assessment focused on sentiment analysis through simply detecting the overall tone of an utterance, and emotion detection that captures a wider range of emotional states which is crucial in understanding the user's feelings. Both tasks are integral for empathetic dialogue generation that requires understanding the overall tone of the utterance and the emotional state of the user to respond empathetically. The use of LLMs for sentiment analysis and emotion detection are briefly presented in this section to provide the essential foundation of well-being assessment.

2.1 LLMs for Sentiment Analysis

Krugmann and Hartmann (2024) explored LLMs' performance in sentiment analysis. Specifically, their study evaluated the performance of three state-of-the-art LLMs: GPT-3.5, GPT-4, and LLaMa 2 for zero-shot binary and three-class sentiment classification tasks, as opposed to traditional learning models. Results showed that GPT-4 surpassed the LLMs for binary sentiment analysis, except the fine-tuned transfer-learning model SiEBERT. While GPT-4 dominated the three-class sentiment analysis for three out of four datasets, RoBERTa outperformed GPT-4 by 15% on the Twitter dataset. Although the LLMs demonstrated their prowess in zero-shot sentiment analysis, their study also highlights that fine-tuned transfer-learning models are able to surpass LLMs in certain contexts.

Sun et al. (2023) proposed a multi-LLM negotiation framework for sentiment analysis to address the challenge that single-round in-context learning of a single LLM may not generate accurate response. The multi-LLM negotiation framework involves a generator LLM that generates the sentiment and a discriminator LLM that evaluates the credibility of the generated sentiment by the generator LLM. Results showed that using two different LLMs such as GPT-3.5 and GPT-4 yield significant performance as opposed to one LLM (self-negotiation). Moreover, introducing a third LLM to settle disagreements between the two LLMs further improved the performance on sentiment analysis.

2.2 LLMs for Emotion Detection

Nedilko (2023) probed the utilization of generative pretrained transformers for multi-class emotion classification. Specifically, ChatGPT was employed to classify code-mixed Roman Urdu and English SMS messages into one of the twelve pre-defined emotion labels. Results showed that ChatGPT exceeded the baseline XGBClassifier and BERT-base-multilingual-cased model. Moreover, it was also observed that the ChatGPT's performance is reliant on the prompt.

Bhaumik and Strzalkowski (2024) introduced an approach that jointly addresses emotion detection and emotion reasoning as a generative question-

answering (QA) task. Their approach includes prompting the LLM to generate a context, then the context is subsequently utilized for the LLM to generate step-by-step reasoning through the chain-of-thought (CoT) prompting, and the emotion label. Results showed that this approach (QA prompting) excelled in emotion detection as opposed to regular prompting and CoT prompting.

3 Task Description

In this study, PERMA well-being assessment is projected as a text classification task. Given the PERMA label **L = {excelling, thriving, surviving, struggling, in crisis}** which is a set containing all possible well-being states defined by (Delphis, 2020) and **U** which is the set of all input utterances, the well-being assessment task is a function $f : U \rightarrow L$ to classify each utterance $u \in U$ with a label $l \in L$ that best represents the well-being state of the utterance $\boldsymbol{u}$. This label is the output of the PERMA well-being assessment task. Figure 1 depicts the five well-being states.

Figure 1: Well-being States Defined by Delphis (2020).

4 Methodology

We outline our procedure in pre-processing the dataset, data annotation, and the experiments to validate the performance of three LLMs, namely ChatGPT-4, Gemini-1.5-pro, and LLaMa 3-8B, on the PERMA-based well-being assessment task.

4.1 ISEAR Dataset

The International Survey on Emotion Antecedents and Reactions (ISEAR) dataset serves as a benchmark for emotion classification. It contains 7,666 records of phrases, sentences, and short paragraphs that were sourced from a survey where participants described their emotional experiences for particular situations (Scherer and Wallbott, 1994).

The ISEAR dataset is chosen in this study because of the emotional experiences transcribed that is closely related to the PERMA model. As such, the emotional responses recorded in the *content* column of the dataset is utilized in our experiments.

Pre-processing included the removal of special characters and duplicate entries from the dataset. Records with non-informative content such as variants of "no response," ""not applicable," "no description," and "nothing" were excluded. After pre-processing, the ISEAR dataset is reduced to 7,475 rows of utterances.

4.2 PERMA Lexicon

The PERMA Lexicon is a tool designed to measure well-being based on the PERMA model. This lexicon associates scores to each token in an input utterance, enabling the automated assessment of well-being from textual data (Schwartz et al., 2016). Prior works (Beredo and Ong, 2022; Ong et al., 2024) employed the PERMA Lexicon to facilitate the assessment of users' mental health for chatbots to generate affective responses. The reliance on dictionaries, however, limits the dynamic handling of new contexts and utterances that use figurative languages (Belal et al., 2023). This prompted the exploration of PERMA in LLMs as it offers the ability to understand context in a way that traditional lexicons cannot.

4.3 Prompt Formulation

Following the work of Vizmanos et al. (2024), prompts were formulated such that they specify the role of the LLM, the well-being assessment task to be performed, the utterance $u \in U$ which serves as the input, and the target labels **L** which serve as options for the output to be generated. These prompts were sent to the respective LLMs from which the LLMs will respond with a label $l \in L$ for each utterance u.

4.4 Large Language Models

The LLMs employed to label the ISEAR dataset according to the PERMA model are ChatGPT-4, Gemini-1.5-pro, and LLaMa 3-8B.

4.4.1 ChatGPT-4

ChatGPT-4 is a transformer model built from GPT-4. It is pre-trained to predict the next token in a sequence using diverse publicly available and third-party licensed datasets. It is then fine-tuned through Reinforcement Learning from Human Feedback (RLHF) (OpenAI, 2023).

The web-interface of ChatGPT-4[1] is employed to label each row of the ISEAR dataset with the well-being states. Because of its 40-prompt limitation every 3 hours, multiple accounts were used in this

[1]https://chatgpt.com/

study to send prompts to the model to label the ISEAR dataset.

4.4.2 Gemini-1.5-pro

The Gemini models are built on top of Transformer decoders with several architectural enhancements and optimizations to support training and optimized inference on Google's Tensor Processing Units (TPUs). The models were then trained on multimodal and multilingual datasets that include data from web documents, PDFs, books, codes, images, charts, audio, and video data using TPUv5e and TPUv4. RLHF was applied post-training to align the model's responses with human preferences (Team, 2024).

The Gemini-1.5-pro API from Google AI for Developers[2] is utilized as this is the latest stable version of the model. Because access to this model is limited to 120 requests per minute with a recommendation to not exceed 1 request per second, the code is implemented to sleep 20 seconds for every request sent. Moreover, Gemini has strict safety guidelines for hate speech, harassment, sexually explicit, and dangerous contents. This hindered 44 utterances from being labeled due to the presence of sensitive content. As such, these 44 entries were removed from the dataset to achieve uniformity across all LLMs.

4.4.3 LLaMa 3-8B

The LLaMa models are based on the transformer architecture with several modifications. The first modification is pre-normalization inspired by GPT-3. The RMSNorm normalizing function was used to normalize the input for each of the transformer sub-layer. The second modification is replacing ReLU with SwiGLU activation function inspired by PaLM. The last modification is replacing absolute positional embeddings with rotary positional embedding (RoPE) inspired by GPTNeo.

The LLaMa models were trained on a diverse set of publicly available datasets. This includes the English CommonCrawl, C4, Github, Wikipedia, Gutenberg and Books3, Arxiv, and Stack Exchange. The data were tokenized with the byte-pair encoding algorithm through the Sentence-Piece tokenizer. As such, the entirety of the training dataset contains roughly 1.4 trillion tokens (Touvron et al., 2023).

The LLaMa 3-8B is chosen for this study as it is currently the most capable and accessible version of the LLM (meta llama, 2024). The entirety of the LLaMa 3-8B model is 16.07GB; but due to hardware constraints, the quantized version of LLaMa 3-8b is obtained from the Ollama[3] library which is only 4.7GB. There are also no request limitations as the LLaMa 3-8B model was executed locally.

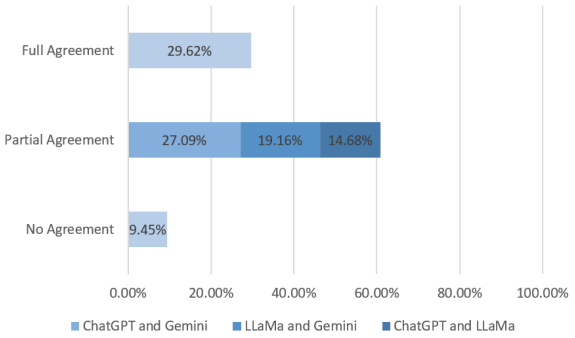

Figure 2: Agreement Percentages of the PERMA Labels Generated by LLMs.

4.5 Evaluation Metric

The Intercoder Agreement is the measure of agreement between annotators in the absence of ground truth. Specifically, the consistency of the PERMA labels across the PERMA Lexicon, ChatGPT-4, Gemini-1.5-pro, and LLaMa 3-8B is analyzed and used as the basis for evaluating the performance of the models.

5 Results and Analysis

We performed two types of analysis using intercoder agreement to evaluate the performance of the LLMs: consistency in PERMA labels and agreement to the reference label.

5.1 Consistency in PERMA Labelling

To determine the consistency of the LLMs in associating PERMA labels to an input utterance, we compare their output label $l \in L$ for each utterance u according to three (3) agreement levels: *no agreement*, *partial agreement*, and *full agreement*. The percentages of agreement are shown in Figure 2.

5.1.1 Full Agreement

The analysis revealed that ChatGPT, Gemini, and LLaMa fully agreed on labeling 29.62% of the dataset as observed in Figure 2. This represents the most reliable classification and showcases the LLMs' ability to consistently identify certain aspects of well-being. Further analysis on their agreement showed that unambiguous utterances with

[2]https://ai.google.dev/gemini-api/docs/api-key

[3]https://ollama.com/library

clear emotional cues and straightforward language are universally recognized by the LLMs. Given an utterance "*I had a summer job in Sweden, and my boyfriend came to meet me on my birthday*," ChatGPT, Gemini, and LLaMa unanimously agreed on the label *excelling*.

5.1.2 Partial Agreement

Partial agreement refers to instances when two of the LLMs agreed on the PERMA labels. As seen in Figure 2, results revealed that ChatGPT, Gemini, or LLaMa partially agreed in their PERMA labels for 60.93% of the dataset. Specifically, ChatGPT and Gemini agreed on labeling 27.09% of the dataset, followed by LLaMa and Gemini with 19.16%, and ChatGPT and LLaMa with 14.68% of the dataset. The most common pair of LLMs with partial agreement is ChatGPT and Gemini. The findings further suggest that Gemini has a higher tendency to come into consensus with both ChatGPT and LLaMa.

Additional insights may also be observed from the pairwise agreement rates. The high agreement rate between ChatGPT and Gemini suggests that these models are more aligned and may have had similar training methodologies and datasets such as fine-tuning through the RLHF. On the other hand, the low agreement rates involving LLaMa may be attributed to the lost precision of the quantized model which could have influenced LLaMa's ability to capture emotional cues. Consider the utterance "*My daughter was two years when she went up to a colt tried to hit it. It turned on her and kicked her over the heart, sent her flying through the air. I left my mother and sister to deal with her as they are nurses. I felt I didn't want to know if she was going to die, it was just too much.*" While ChatGPT and Gemini both labeled this *in crisis* due to the intensified situation-driven emotional cue implying sadness, fear, and anxiety, LLaMa labeled this utterance *excelling*.

5.1.3 No Agreement

No agreement is used to refer to instances when the three LLMs generated differing PERMA labels. Results shown in Figure 2 revealed that the models did not agree on the PERMA labels for 9.45% of the dataset. This lack of agreement highlights the challenges in well-being assessment, and suggests that the ambiguous nature of emotional expressions in certain sentences were challenging for the LLMs to classify consistently (Barrett et al., 2011).

A closer examination of the dataset revealed that the disagreement between the LLMs occurred as the labels generated by each LLM are merely adjacent from each other. This is evident in Figures 4, 5, and 6 where the concentration of values is along the diagonal and the adjacent cells. While the highest concentration shown through the heat map is along the diagonal that represents agreement, the minimal concentration on the adjacent cells indicate that even when the LLMs disagreed, their assessment were often close. This mirrors real-world scenarios where different psychologists may give varying diagnosis based on their own interpretation and respective biases, highlighting the complexity and subjectivity in well-being assessment.

Further analysis of the variance among the LLMs' labels showed that 8.33% of the dataset has a high variance, meaning the labels assigned by the LLMs are not adjacent, but at least two well-being states away. For instance, the utterance "*The day I was happiest was the day when I received a phone call from Eve's Weekly to inform me that I had won the first prize of the All India Essay competition. I had won this prize when I was an undergraduate when even post graduates had participated. I had been judged by eminent judges and political scientists*" was labeled by ChatGPT, Gemini, and LLaMa as *surviving*, *excelling*, *excelling* respectively. On the other hand, 91.67% of the dataset exhibited low variance. That is, the LLMs assigned either similar or adjacent labels to a given utterance. Given an utterance "*A bus drove over my right leg. The event itself was not very frightening, but when I had to wait in the emergency ward for three hours and then my leg began to swell, I was frightened.*," ChatGPT, LLaMa, and Gemini labeled the utterance as *struggling*, *struggling*, and *surviving*. This suggest that while the LLMs may align in well-being assessments, slight difference on interpreting utterances may still occur.

5.2 Reference Label

A reference label is a predefined label used as the standard in evaluating the performance of a machine learning model in tasks such as classification. This serves as the "ground truth" from which the outputs generated by the model are compared with. Because the ISEAR dataset does not have a reference PERMA label, the most common label generated between the three LLMs, which we termed as the "**mode**", was adopted to be the reference label in this study. We used this model to perform further analysis on the performance of the each PERMA

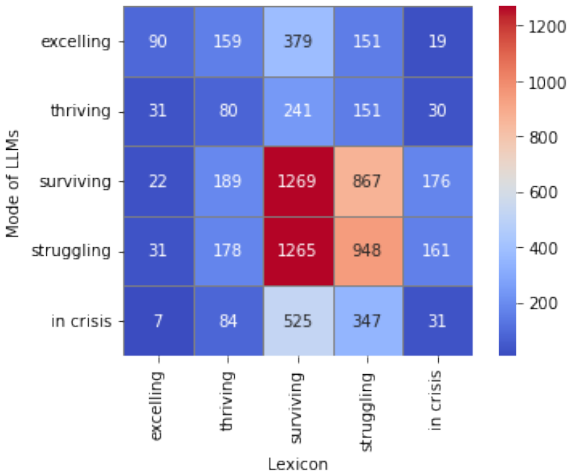

Figure 3: Mode of LLMs vs. PERMA Lexicon.

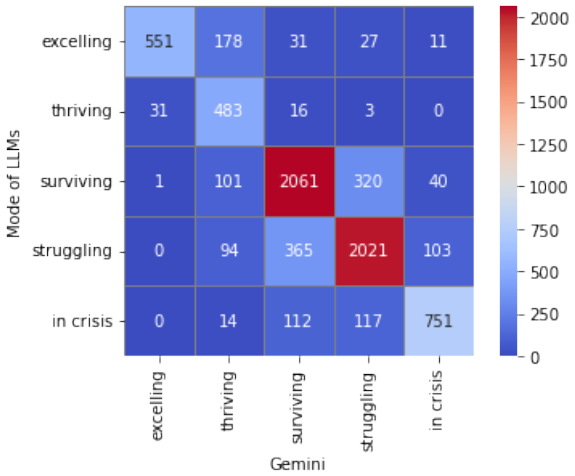

Figure 5: Mode of LLMs vs. Gemini-1.5-pro.

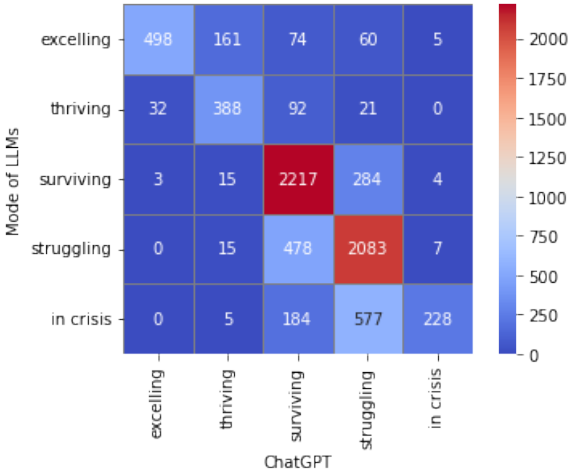

Figure 4: Mode of LLMs vs. ChatGPT-4.

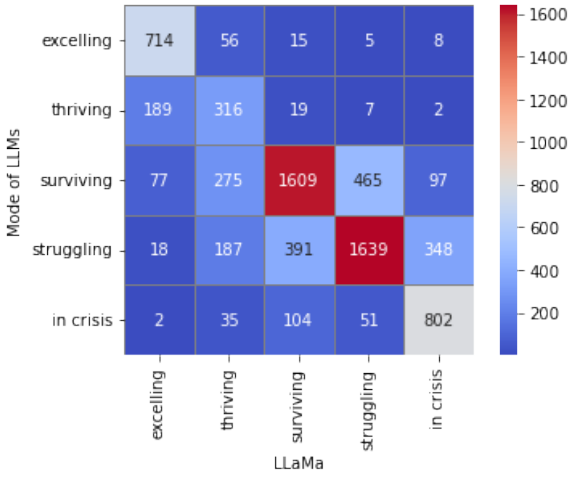

Figure 6: Mode of LLMs vs. LLaMa 3-8B.

annotator namely PERMA Lexicon, ChatGPT-4, Gemini-1.5-pro, and LLaMa 3-8B.

5.2.1 PERMA Lexicon

The PERMA Lexicon achieved an intercoder agreement of 32.54%, the lowest amongst the annotators employed in this study. It is observed in Figure 3 that the lexicon struggles to classify *excelling* and *in crisis* states, but rather classifies the utterances as *surviving* instead. This may be attributed to the context-dependent nature of language describing high and low emotional states that the lexicon fails to capture because of its static dictionaries. An example of this would be words that are positive in one context, but are negative in another. Consider the utterance "*I am dying out of laughter!*" While this utterance is conveying excessive joy and used the word *dying* to express this intensified feeling, the PERMA Lexicon labeled this as *struggling* because of the negative score associated with the word *dying*. This exemplifies the lexicon's inability to understand context, causing it to miss subtle cues, misinterpret the utterance, and ultimately misclassify the well-being states.

5.2.2 ChatGPT-4

ChatGPT-4 recorded an intercoder agreement of 72.86%. ChatGPT-4 mostly misclassified *in crisis* labels as *struggling*. However, it was able to excel in classifying other nuanced states like *surviving* and *struggling* as observed in Figure 4. Despite its high agreement rate, its occasional misclassification highlights the need for further training due to the sensitive nature of psychological well-being.

5.2.3 Gemini-1.5-pro

Gemini-1.5-pro achieved an intercoder agreement of 78.95% which is the highest amongst the annotators. It is particularly able to classify most of

the well-being states in consensus with the other LLMs as shown in Figure 5. This suggests that Gemini-1.5-pro may have understood the context of the utterances more compared to the other LLMs. Though not explicitly mentioned, Gemini's architecture, training, and fine-tuning may have aided it in capturing the subtle emotions which led to its high agreement with other LLMs.

5.2.4 LLaMa 3-8B

LLaMa 3-8B was able to record an intercoder agreement of 68.36%. LLaMa 3-8B excelled in classifying the extremities of the well-being states compared to the other LLMs. Specifically, LLaMa 3-8B was able to accurately classify *excelling* and *in crisis* more than ChatGPT-4 and Gemini-1.5-pro as shown in Figure 6. However, LLaMa 3-8B also recorded the lowest performance in classifying *thriving*, *surviving*, and *struggling* states as opposed to ChatGPT-4 and Gemini-1.5-pro. As mentioned before, the lost precision from employing the quantized LLaMa 3-8B model could have affected its capability in capturing context-dependent texts and subtle nuances of expressions.

5.3 Discussion

Analysis of the results revealed the distinct strengths and weaknesses of each LLM in the well-being assessment task. ChatGPT-4 excelled in classifying intermediate states *surviving* and *struggling*, but encountered challenges in classifying *excelling* and *in crisis* states as shown in Figure 7. Conversely, LLaMa 3-8B proficiently classified the extremities of the well-being states *excelling* and *in crisis*, although it performed the worst in classifying *thriving*, *surviving*, and *struggling* states. Despite Gemini-1.5-pro achieving the highest intercoder agreement, it was only able to outperform the other LLMs in classifying the *thriving* state. Further analysis revealed that utterances with ambiguous language or mixed emotions resulted in disagreement between the LLMs, while utterances with clear emotional cues resulted in agreement amongst the LLMs.

Barrett et al. (2011) previously highlighted the significance of context in emotional expressions, revealing that there is significant variability in how emotions are expressed and perceived which is rooted on personal experiences, societal expectations, and cultural norms. This emphasizes that the interpretation of emotional expressions is highly variable and deeply influenced by contextual factors. Barrett suggests researchers to account for the variability and context-dependence of emotions. This has direct implications for NLP classification tasks where context can drastically alter the meaning of the language used.

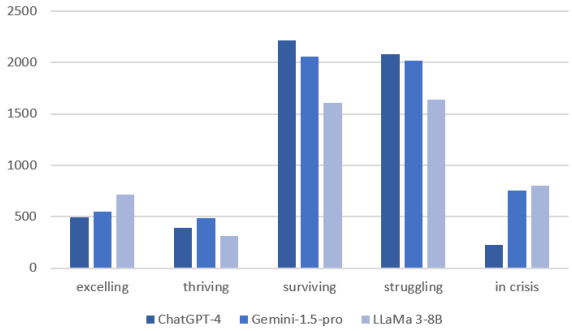

Figure 7: Performance of the LLMs in PERMA Well-Being Assessment Task.

Ghosal et al. (2021) quantified the role of context in emotion, act, and intent detection for utterance-level dialogue understanding. Findings revealed that inter-speaker context had the most significant impact on the model's performance, followed by the context shuffling of the order of an utterance in a dialogue. Moreover, replacing the utterance with its paraphrased version led to a minimal decrease in the model's performance, indicating that the overall meaning conveyed by the utterance is what primarily contributed to the accurate classification rather than the precise wording. Meanwhile, Chatterjee et al. (2019) developed EmoContext that handles the ambiguity of emotional expressions by leveraging contextual information from dialogue history. EmoContext, however, still faced challenges in differentiating the *happy* class from the *neutral* class due to the inherent ambiguity between these classes. A greeting like "*Happy Morning*" can be interpreted by some as conveying a happy emotion, while being interpreted as neutral by others. These challenges that continue to baffle emotion detection research, combined with the multi-faceted dimensions of well-being, will be addressed in future studies that seek to build LLMs able to perform PERMA-based well-being assessment.

6 Conclusion

This paper explored the potential of LLMs in detecting psychological well-being through the PERMA model. The findings revealed that while LLMs offer additional contextual understanding and there is a substantial agreement among the LLMs, fur-

ther research and development or refinement must be done to enhance the accuracy and reliability of LLMs for psychological well-being assessments. Moreover, the utilization of the intercoder agreement as a metric to establish ground truth that facilitated the comparison of the LLMs' performance in the absence of labeled data. This approach is particularly crucial in research areas where annotated datasets are scarce.

The insights gained from this study can contribute to the ongoing research of LLMs in mental health and psychological well-being assessments. Future works will focus on refining the LLMs, exploring additional LLMs, and incorporating human validation to enhance the reliability of psychological assessments. Additionally, a middle-layer architecture that will function as a decision-making module may be developed to optimize the distinct strengths of each LLM in classifying well-being states. Lastly, emotion embeddings may be explored to represent the user's emotional state to aid the LLMs in capturing the complexity and nuances of human emotions.

References

Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current directions in psychological science*, 20(5):286–290.

Mohammad Belal, James She, and Simon Wong. 2023. Leveraging chatgpt as text annotation tool for sentiment analysis. *Preprint*, arXiv:2306.17177.

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).

Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a generative approach for emotion detection and reasoning. *Preprint*, arXiv:2408.04906.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Delphis. 2020. The mental health continuum is a better model for mental health. https://delphis.org.uk/mental-health/continuum-mental-health/.

Gesa Solveig Duden, Stefanie Gersdorf, and Katarina Stengler. 2022. Global impact of the covid-19 pandemic on mental health services: A systematic review. *Journal of Psychiatric Research*, 154:354–377.

Shiv Gautam, Akhilesh Jain, Jigneshchandra Chaudhary, Manaswi Gautam, Manisha Gaur, and Sandeep Grover. 2024. Concept of mental health and mental well-being, it's determinants and coping strategies. *Indian Journal of Psychiatry*, 66(Suppl 2):S231–S244.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online. Association for Computational Linguistics.

James J. Gross and Oliver P. John. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does GPT-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13622–13623.

meta llama. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Robert R. Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M. Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, 20(6):e10148.

Andrew Nedilko. 2023. Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 616–620, Toronto, Canada. Association for Computational Linguistics.

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).

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *Preprint*, arXiv:2306.00176.

James W. Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162–166.

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):310.

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*, pages 516–527. World Scientific.

Martin Seligman. 2010. Flourish: Positive psychology and positive interventions. *The Tanner Lectures on Human Values*, 31(4):1–56.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience lookahead. *arXiv preprint arXiv:1906.08487*.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations. *Preprint*, arXiv:2311.01876.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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. *Preprint*, arXiv:2302.13971.

Julianne Vizmanos, Ethel Ong, Jackylyn Beredo, and Remedios Moog. 2024. Well-being assessment using chatgpt-4: A zero-shot learning approach. In *Proceedings of the 24th Philippine Cmoputing Science Congress*. Computing Society of the Philippines.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, Lingshuai Wang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. STICKERCONV: Generating multimodal empathetic responses from scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733, Bangkok, Thailand. Association for Computational Linguistics.

Ming Zhou, Minlie Huang, and Xiaoyan Zhu. 2020. Emotion-aware chatbots: A survey of recent advances and future research directions. *Information Fusion*, 59:103–127.