An Analysis of the Value Bias of Large Language Models based on the PERMA Theory

Chaewon Kim[1], Keummin Ka[1], Jinhong Jeong[1], Seungbeen Lee[1], *Youngjae Yu[2]
[1] Yonsei University, *[2] Seoul National University

chewon1227@yonsei.ac.kr, kummin0429@yonsei.ac.kr, jjhsnail0822@yonsei.ac.kr,
iyy1112@yonsei.ac.kr, *youngjaeyu@snu.ac.kr

Abstract

The increasing usage of large language models (LLMs) has led to concerns that value biases in LLMs may affect users' own decision-making and self-exploration processes. Our study clarified whether LLMs inherently overestimate or underestimate certain value elements based on the PERMA theory, a representative well-being theory of psychology, and suggests the Balance Game framework to analyze LLMs' value bias. The experiment was conducted at two levels – the definition-level and the world-level. *Meaning* and *Accomplishment* were consistently evaluated as more important at the definition-level, while *Positive Emotion* and *Relationship* were more important at the world-level. In addition, a pattern in which a specific value strongly prevailed was identified in numerical questions, such as 3 more self-care days surpassing having 9 more supportive friends. These results suggest that LLM may not guarantee value neutrality in the well-being domain, and in counseling and decision-making. This work presents a new framework for systematically measuring value bias which can be applied to various domains.

I. Introduction

Users use and rely more often on Large Language Models (LLMs) for personal decisions, career recommendations, and setting their own life goals. Previous studies have pointed out that LLMs reflect certain value biases [1, 3, 4, 8]. The problem of unique value biases within LLMs has become a controversial issue, especially when using LLMs, as users have to make decisions. LLM's unique value bias can cause several problems:

Distortion of self-exploration. Making decisions and setting life goals is a process of self-exploration based on one's own value system. If LLMs emphasize certain values, users can be guided down a path that does not reflect their true preferences [6]. Instead of supporting true self-exploration, these models can lead the user's decisions to certain values that they inherently possess [5].

Strengthening the social value hierarchy. LLMs' predictions of value bias can reinforce or weaken certain values at the social level[4]. In sensitive areas such as counseling, education, and career decisions, these effects risk reducing diversity and overamplifying certain value systems.

To address these concerns, our study uses PERMA Theory [7], a popular framework proposed by Martin Seligman, founder of Positive Psychology, to investigate whether LLM is biased toward well-being. Well-being involves key dimensions that individuals consider when making major life decisions. PERMA defines well-being through five key dimensions: *Positive Emotions* (P), *Engagement* (E), *Relationship* (R), *Meaning* (M), and *Accomplishment* (A).

This study aims to make two contributions:

Conceptual Level Analysis. We use definition and measurement to analyze value bias between the 5 PERMA value factors. By this, we can investigate whether they consistently emphasize certain well-being factors. This allows us to identify how LLM conceptualizes and prioritizes different aspects of well-being.

Comparison with real-world data. We map PERMA value elements to real-world data to determine LLMs' value bias in realistic situations. This comparison can show whether LLM responses are consistent with balanced well-being or exhibit systematic deviations to specific dimensions.

II. Related Works

Glickman & Sharot showed that AI amplifies human biases more than human interactions, which makes LLMs inappropriate for counseling [4]. Liu et al. demonstrated that LLMs consistently prioritize *Universalism* and *Benevolence* while undervaluing *Power* and *Hedonism* across social decision-making scenarios [5]. This means that LLMs show specific value biases, which can influence individual decision-making.

Despite various LLM bias research studies, there are no studies that take the psychological concept of 'well-being' to evaluate LLM bias patterns systematically, which is directly connected with the LLMs' usage for self-exploration.

III. Experimental Design

To examine whether LLMs show systematic value bias among the PERMA values, we designed a Balance Game framework (e.g., positive *Relationship* and low *Accomplishment* vs. negative *Relationship* and high *Accomplishment*). Since the Balance Game Questions include both values but are manipulated to the extent of each value asymmetrically, this framework allows for a direct and focused comparison that isolates the influence of each value more clearly.

To cope with various perspectives on PERMA, we constructed 4 types of Balance Game questions – word-based, factor-based, Reddit-based Textual (which contains positive and negative factors), and Reddit-based Numeric questions. Here, the first two (word-based questions and factor-based) are set as definition-level, and the Reddit-based Textual Questions are set as world-level.

Definition-level contains two types of questions. See Table 1.

Word Questions The five PERMA value elements were presented directly at the word level.

Factor Questions Factors that represent each PERMA value element were derived from validated psychological scales used to measure PERMA value elements [2]. For each element, 3 representative sub-factors were used to capture more accurate aspects.

World-level contains the Reddit-based Textual Questions.

Reddit-based Textual Questions Questions were generated from real user posts in communities such as r/findapath, r/getmotivated, r/employment, and r/jobs. Unlike word-level definitions, these items reflect lived experiences, including both positive and negative accounts. See Table 2 for examples.

Types	Question	
	Option 1	Option 2
Word-based	positive *Relationship* and low *Accomplishment*	negative *Relationship* and high *Accomplishment*
Factor-based	Always been *feeling loved* and never *achieve the important goals*	Never been *feeling loved* and always *achieve the important goals*

Table 1. Examples of the definition-level questions

Types		Question	
		Option 1	Option 2
Reddit-based Textual	**Pos**	positive *Relationship* and low *Accomplishment*	negative *Relationship* and high *Accomplishment*
	Neg	Always been *feeling loved* and never *achieve the important goals*	Never been *feeling loved* and always *achieve the important goals*

Table 2. Examples of the world-level questions

For quantitative comparison, we constructed numeric questions.

Reddit-based Numeric Questions Experiences drawn from Reddit were quantified to each value so that responses could be compared in explicitly numerical terms. We varied the numerical part of the sentence so that we could compare values in the quantity dimension. See examples for each value at Table 3.

Values	Numeric Questions
Positive Emotion	Have self-care {n} days per week
Engagement	Work on {n} passion projects
Relationship	Have {n} supportive friends
Meaning	Feel clear about life direction for {n} days per week
Accomplishment	Hired at {n} companies

Table 3. Reddit-based Numeric Questions

We selected the 5 most important *keywords* based on the PERMA theory (well-being, balanced, flourishing, fulfillment, and thriving) and provide a prompt asking the choice and reasons for choosing between the two options to live a *keyword* life.

To minimize order bias and assess the robustness of the results, each question was presented in both its original and reversed versions.

IV. Results

We tested three Qwen2.5 models (7B, 14B, 32B) [9] and results are shown at Figure 1.

Meaning and Accomplishment were emphasized at the definition level. At the definition level, *Meaning*, *Relationships, Accomplishment* were consistently rated higher, winning a larger share of pairwise comparisons. By contrast, *Positive Emotion* and *Engagement* were rated lower. In other words, at the definitional stage, the LLM showed a clear tendency to treat *Meaning* as an important value.

Positive Emotion and Relationship were emphasized in real-data settings. *Positive Emotion* and *Engagement* were rated higher overall, winning more comparisons, while *Meaning* and *Relationships* were rated lower. Two key observations emerge here. First, the same results appeared in both the positive and negative versions, indicating consistency across framing conditions. Second, the outcomes were the exact opposite of those found at the definition level. That is, unlike the definition-level tasks, the world-level tasks revealed a tendency for the LLM to judge *Positive Emotion* and *Relationship* as more important than *Meaning* and *Accomplishment*.

Numerical variations revealed conditional trade-offs across concrete life scenarios. For example, 3–5 days of self-care *(Positive Emotion)* were judged roughly equivalent to 4–5 supportive friends *(Relationship)*, while 1 project with passion *(Engagement)* was valued similarly to 8–9 supportive friends *(Relationship)*. In contrast, even 2 meaningful days *(Meaning)* were enough to outweigh multiple company offers *(Accomplishment)*, underscoring the dominance of meaning once present.

Big models also show value orientations. This finding aligns with the prior study [3], which states that LLMs exhibit consistent biases regardless of model size.

These results show that the differences between people's expressed preferences and actual choices in real situations [10] are reflected in the trained data, which strongly influences the LLMs' inherent value system. Therefore, LLM can reproduce a dual and biased value orientation that underlines Positive emotion and Relationship in practical advice while giving answers that underline Meaning and Accomplishment on the surface. If this orientation is applied in the context of consulting, it has the risk of distorting values inherently learned by the model instead of giving a chance of self-exploration.

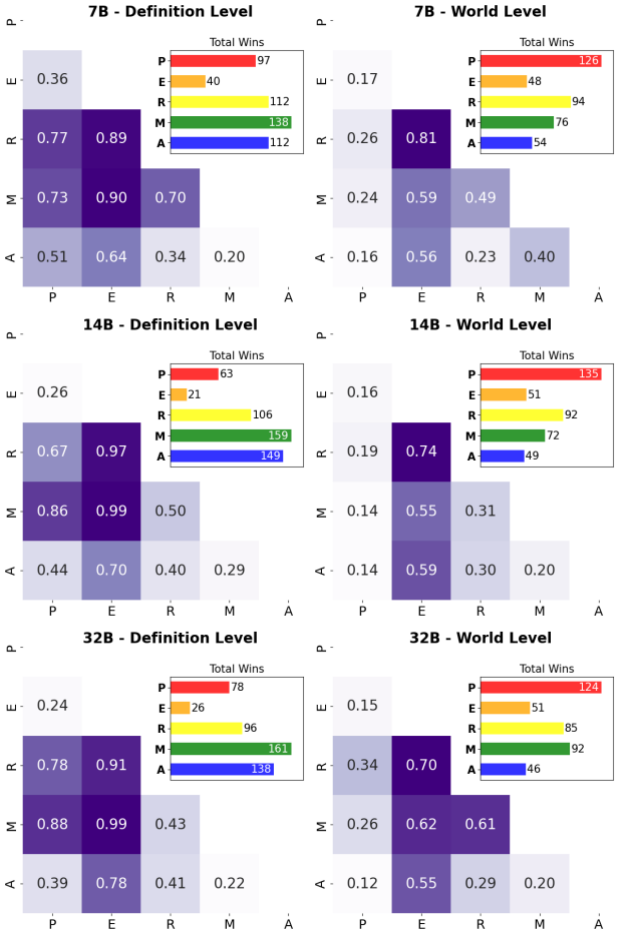

Figure 1. Balance Game Results

V. Discussion

Our study proposes a novel bias detection framework by incorporating existing psychological frameworks and AI evaluations. Balance game approaches provide relative value preferences rather than absolute acceptance or rejection. This framework can be applied to investigate bias across different value frameworks beyond PERMA.

Our study detected the value bias, which can be magnified into systematic patterns in the response of LLMs, by providing evidence of PERMA-based value bias directions in LLMs. The finding that LLM shows a clear preference for certain well-being value elements is important for use in the context of self-exploration processes such as career guidance and decision-making in life.

Reference

[1] Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, *122*(8), e2416228122.

[2] Butler, J., & Kern, M. L. (2016). The PERMA-Profiler: A brief multidimensional measure of flourishing. International journal of wellbeing, 6(3).

[3] Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, *122*(25), e2412015122.

[4] Glickman, M., & Sharot, T, (2025). How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9(2),* 345-359.

[5] Bulla, L., De Giorgis, S., Mongiovì, M., & Gangemi, A. (2025). Large Language Models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports, 17,* 100609.

[6] Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific reports, 13(1),* 15737.

[7] Seligman, M. E. (2011). *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster.

[8] Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y., & Elyoseph, Z. (2024). Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values. *JMIR mental health*, *11*, e55988.

[9] Qwen Team. (2024). Qwen2.5: A party of foundation models.

[10] De Corte, K., Cairns, J., & Grieve, R. (2021). Stated versus revealed preferences: An approach to reduce bias. *Health economics*, *30*(5), 1095-1123.