CACLP: Context-aware Contrastive Lexical Prompting for Multi-trait Personality Inducing in LLMs

Anonymous ACL submission

Abstract

Inducing Large Language Models (LLMs) to exhibit specified personalities is critical for various applications like role-playing and social support. Psychological findings suggest that personalities comprise multiple inherently correlated traits with dynamic expression across contexts, yet most existing methods neglect these characteristics, consequently hindering human-like interactions. Inspired by the lexical hypothesis and the trait activation theory of personality, we propose Context-aware Contrastive Lexical Prompting (CACLP), which resolves trait exhibition conflicts via lexical knowledge and dynamically selects context-aware adjectives for multi-trait inducting in LLMs. Specifically, CACLP eliminates semantically conflicting adjectives using WordNet to construct conflict-free adjectives describing multi-trait personalities by considering both target traits and their opposites. Then, it dynamically selects context-relevant adjectives via Natural Language Inference (NLI) to align responses with various contexts. Extensive experiments across three widely studied personality models on diverse LLMs demonstrate CACLP's general superiority over baseline methods, especially on smaller models.

1 Introduction

Personality shapes the enduring patterns of an individual's thoughts, emotions, and behaviors (Mischel et al., 2007). Therefore, in widely developed Large Language Models (LLMs)-based applications like role-playing (Wang et al., 2024; de Winter et al., 2024), education (Sonlu et al., 2024; Liu et al., 2024b), and social support (Tu et al., 2023), inducing LLMs to exhibit specific personalities is crucial in providing human-like interactions.

Although important, personality inducing in LLMs is a challenging problem that has not yet been solved well. According to the trait theory of personality (Novikova, 2013), personality is composed of multiple traits that describe distinct aspects of behavioral patterns, such as *Agreeableness*, *Conscientiousness*, *Extroversion*, *Neuroticism*, and *Openness to Experience* in the Big Five Model (John et al., 1991). However, most existing studies focus on controlling LLMs to exhibit single traits separately rather than multiple traits simultaneously (Jiang et al., 2023a; Li et al., 2024; Cava et al., 2024), which overlooks the inherent correlations and potential conflicts in the expression of personality traits. Besides, almost all existing methods didn't consider the dynamic expression of traits across different contexts. Consequently, only partial of multiple personality traits can be effectively induced simultaneously (Huang et al., 2023; Pan and Zeng, 2023; Jiang et al., 2023b).

For potential conflicts among trait expressions, psychological findings (Fleeson, 2001; Wilt and Revelle, 2015) suggest that though conceptually distinct, multiple traits under the same personality model have interrelated effects. Besides, the lexical hypothesis of personality (Allport and Odbert, 1936) suggests that different personality traits are clustered from various descriptive words, and there are no distinct boundaries among these clusters. Therefore, one single word may simultaneously correlate with multiple traits. When such words are used to induce LLMs to exhibit specified personality traits, they may also inadvertently describe traits opposite to the target traits. We hypothesize that the use of such ambiguous lexical cues is a key reason why existing methods struggle to effectively induce LLMs to express multiple traits simultaneously. Besides, the trait activation theory (Tett and Guterman, 2000) in psychology suggests that different contexts may trigger the expression of specific traits. For instance, Conscientiousness may dominate in work environments, while Extraversion may more expressed in social engagement. These inspired us that the inducing of multi-trait personalities of LLMs should also be dynamically adapted to different contexts.

Based on the analysis above, we propose Context-aware Contrastive Lexical Prompting (CACLP), a novel prompting approach that facilitates lexical knowledge to resolve potential conflicts among trait expressions and dynamically selects contextually relevant expressions, enabling effective multi-trait inducing in LLMs. Specifically, to address potential conflicts in multi-trait inducing, we design Contrastive Adjective Refinement. It leverages WordNet (Miller et al., 1990), a well-known semantic knowledge graph, to identify and eliminate negatively correlated descriptive words by considering both the target traits and their opposites, constructing a conflict-free set of adjectives describing the target personality traits. Then, we propose Context-aware Adjective Retrieval, which dynamically retrieves contextually relevant personality-descriptive adjectives from the aforementioned conflict-free set using Natural Language Inference (NLI). The retrieved adjectives are then used to construct prompts for inducing multi-trait personalities in LLMs.

To evaluate the effectiveness of CACLP in inducing multi-trait personalities in LLMs, we conduct comprehensive experiments by prompting LLMs to answer personality inventories comparing with baseline methods on three widely studied personality models: the Big Five Model (John et al., 1991), 16Personalities[1], and the Dark Triad (Paulhus and Williams, 2002). The evaluation was performed across different LLMs with varying architectures and scales. The results showed that our method consistently outperformed the baseline methods in inducing multi-trait personalities across all personality models and LLMs. Notably, CACLP achieves significant improvements in most results inducing multi-trait personalities on smaller-sized LLMs (e.g., 3B parameters) compared to baseline methods, demonstrating its adaptability to resource-constrained scenarios[2]. To summarize, the main contributions of our work are as follows:

- We introduce Context-aware Contrastive Lexical Prompting (CACLP), an innovative prompting method that leverages lexical knowledge to resolve potential conflicts in trait expressions and dynamically selects context-appropriate adjectives, enabling effective multi-trait induction in LLMs.

- CACLP is inspired by the lexical hypothesis and trait activation theory of personality, which bridges the gap between psychological findings and practical LLM prompting strategies for inducing multi-trait personalities.

- Extensive experiments across multiple personality models demonstrate that CACLP consistently outperforms baseline methods when applied to LLMs with different architectures and scales, especially on smaller LLMs. This highlights CACLP's generalizability and adaptability to resource-constrained scenarios.

2 Related Work

In this section, we categorize the most relevant studies into single-trait personality induction and multi-trait personality induction to provide a comprehensive overview. It is worth noting that not all personality models in existing research are grounded in trait theory. Though some studies (Huang et al., 2023; Cui et al., 2023) also explore multi-dimensional personality models such as the Myers-Briggs Type Indicator (MBTI, Myers (1962)), their induction methods are similar. For simplicity, we use the terms *single-trait* and *multi-trait* to encompass these works as well.

2.1 Single-trait Inducing in LLMs

Personality inducing in LLMs can be traced back to early work (Karra et al., 2022), which demonstrated that LLM personalities can be altered by fine-tuning on auxiliary classification or generation tasks with personality annotations. Similarly, (Li et al., 2023) conducted instruction fine-tuning for GPT-3 using questionnaire items and their corresponding answers on higher *Agreeableness* and lower *Neuroticism*, leading to more positive and emotionally stable personality expressions. Besides updating model parameters, researchers (Jiang et al., 2023a; Safdari et al., 2023; Weng et al., 2024; Tan et al., 2024; Huang et al., 2023; Cava et al., 2024; Jiang et al., 2023b; Noever and Hyams, 2023) have also designed various prompts, such as trait descriptions or interpretations from psychological questionnaires, to induce LLMs to exhibit specific personality traits.

Although these methods effectively induce specific personalities, they often only focus on inducing one single trait each time in isolation, overlooking potential inter-trait correlations and conflicts. Consequently, traits unintended to induce may also

[1] 16personalities.com
[2] Our code and results will be released publicly.

be modified inadvertently, leading to undesired personality exhibition (Karra et al., 2022; Huang et al., 2023; Jiang et al., 2023b).

2.2 Multi-trait Inducing in LLMs

Considering the aforementioned limitations, few researchers have attempted to induce multi-trait personalities in LLMs simultaneously.

One way is to induce the LLM to play personas with specific multi-trait personalities. For instance, (Jiang et al., 2023b) modeled 10 LLM personas for each combination of the Big Five personality traits. These personas were integrated into the system-level prompt to induce the LLM to exhibit the corresponding personalities. Similarly, (Huang et al., 2023) prompts the LLM to portray celebrities with specific MBTI types to induce the model to exhibit their personalities. However, role-specific information (such as the character's name and background) may influence the LLM's understanding of personality traits, potentially leading to biases in personality inducing.

Another approach combines single-trait datasets to create a multi-trait dataset for LLM fine-tuning. For example, (Cui et al., 2023) employs ChatGPT to annotate the Alpaca dataset with single dimensional MBTI labels (I-E, S-N, T-F, J-P) to generate eight conversation datasets, which were then combined for fine-tuning LLM to express various MBTI personality types. However, this method may introduce unintended correlations and conflicts between traits, limiting the LLM's ability to express the intended personalities effectively.

Moreover, most existing studies neglect the contextual influence on multi-trait personality expression. Although (Liu et al., 2024a) enables dynamic change of the personality in LLMs by updating the LoRA layers, how the change is influenced by the context is not detailed analyzed.

3 Context-aware Contrastive Lexical Prompting

3.1 Problem Statement

For a personality model based on the trait theory including n personality traits $\{p_1, p_2, \ldots, p_n\}$, where each trait has two possible extents: low and high, the studied problem can be described as: Given a combination $P = [p'_1, p'_2, \ldots, p'_n]$ (where p'_i can be p_i-low or p_i-high) that includes multiple traits and their respective extents, the objective is to modulate a large language model M to exhibit P within its responses. For example, for the Big Five Model with five traits: *Agreeableness*, *Conscientiousness*, *Extroversion*, *Neuroticism*, and *Openness to experience*, one possible combination P can be $[A\text{-low}, C\text{-high}, E\text{-high}, N\text{-low}, O\text{-high}]$.

To solve the problem above, we design Context-aware Contrastive Lexical Prompting (CACLP) to utilize the current context to retrieve the most relevant descriptive adjectives as the system-level prompts to induce the LLMs to generate responses that reflect the given multiple personality traits P, as shown in Figure 1. CACLP is mainly divided into three parts: Knowledge-enhanced Adjective Generation, Contrastive Adjective Refinement, and Context-aware Adjective Retrieval. We will introduce each part in detail as follows.

3.2 Knowledge-enhanced Adjective Generation

We first introduce personality-related domain knowledge into prompts of LLMs and enable them to generate descriptive adjectives for all specified single personality traits.

The lexical hypothesis of personality (Galton, 1884; John et al., 1988) shows that various personality traits are distinguished by different clusters of descriptive adjectives. For example, the traits included in the Big Five model were identified by five clusters of descriptive adjectives with similar qualities in the dictionary (Allport and Odbert, 1936). Inspired by this, we hypothesize that it is crucial to obtain descriptive adjectives that characterize these traits when inducing LLMs to exhibit multiple personality traits.

However, it isn't easy to obtain adjectives that accurately describe personality traits. Although psycholinguistic research (Saucier and Goldberg, 1996) has summarized correlations between adjectives and traits in the Big Five model, not all personality models have such comprehensive lexicons to describe their traits. Therefore, we propose a versatile way to employ LLMs to autonomously generate adjectives that describe each personality trait. As most LLMs are not specifically trained on professional personality corpora, directly prompting them to generate descriptive adjectives is ineffective due to their inadequate comprehension of personality. So, we introduce domain knowledge K_p (*i.e.*, explanations/definitions of traits from personality inventories and literature introducing personality models) as the context to prompt LLMs to gen-

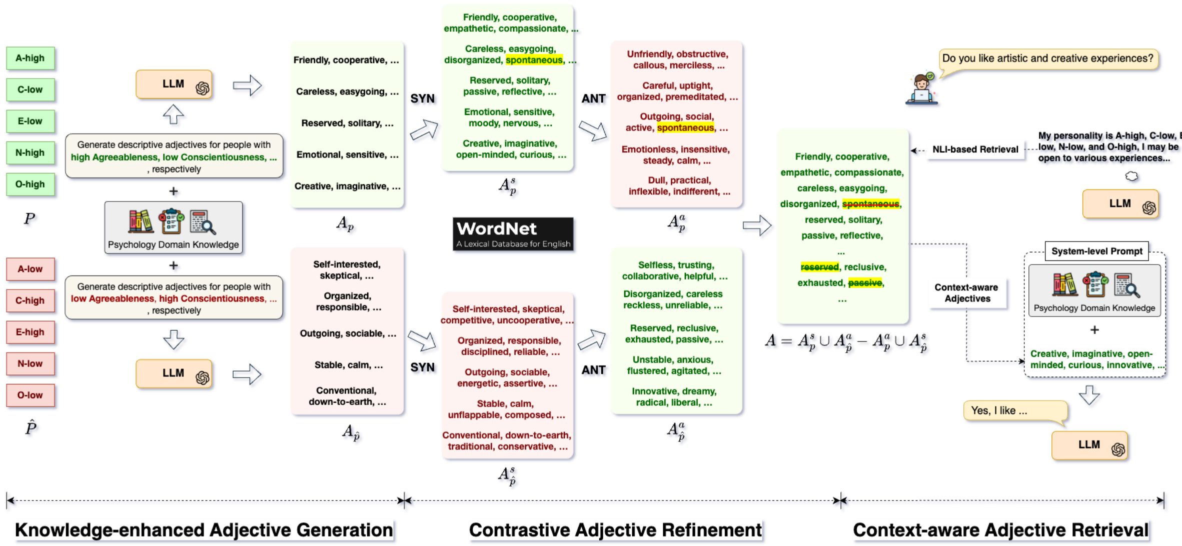

Figure 1: The model overview of Context-aware Contrastive Lexical Prompting

erate m descriptive adjectives $A_i = \{a_1^i, ..., a_m^i\}$ for each trait p_i. The total descriptive adjectives are $A_P = \{A_1, ..., A_n\}$.

3.3 Contrastive Adjective Refinement

Although traits are conceptually distinct in a personality model, psychological studies (Fleeson, 2001; Wilt and Revelle, 2015; Van der Linden et al., 2010) suggest an often correlations among different personality traits. This may lead to potential conflicts when modulating LLMs to exhibit multiple traits simultaneously. We believe this is an important reason that has been overlooked by existing studies that consequently impact their performance. To address this issue, we design a lexical-based approach that considers both P and its opposite to eliminate the potential conflict trait descriptive adjectives among multiple traits within P.

As mentioned above, A_P contains adjectives describing multiple personality traits in P. There is a potential that words within A_P describing p_i may semantically conflict with those describing p_j. For example, the word *spontaneous* describes people with $C_{\text{-low}}$ in some contexts but also describes people with $E_{\text{-high}}$. Therefore, if the word *spontaneous* is utilized to prompt LLMs to exhibit $C_{\text{-low}}$ and $E_{\text{-low}}$ simultaneously, the performance may be adversely affected due to this semantic conflict. Besides, we also notice that the same personality trait can be described differently in different contexts. For example, *reserved* can describe people with $E_{\text{-low}}$ in a neutral context, whereas *passive* can also describe such individuals but often carries a negative connotation.

Based on the analysis, we hypothesize that if an adjective is positively correlated with p_i, its synonyms (also describing p_i but in different contexts) are positively correlated to p_i, while its antonyms would be negatively correlated to p_i. Therefore, we can employ lexical methods to identify and remove adjectives in A_P that are positively correlated with $p_i \in P$ but negatively correlated with other traits, thereby resolving internal conflicts within A_P to facilitate further induction.

Specifically, we first employ Wordnet, (Miller et al., 1990), a well-known semantic knowledge graph, to expand A_P with their synonyms to form a comprehensive vocabulary A_P^s for describing all traits in P under various contexts. Subsequently, for all adjectives in A_P^s, we use WordNet again to find their corresponding antonyms A_P^a. Based on our hypothesis, these antonymic adjectives describe traits that are opposite to P, which we aim to eliminate during the induction of P.

To comprehensively identify adjectives that describe P while minimizing potential conflicts, we simultaneously consider personality trait combinations $\hat{P}$ that are opposite to P. We employ the same methodology to generate the adjective set $A_{\hat{P}}$ that describes $\hat{P}$, and further expand it using WordNet to obtain $A_{\hat{P}}^S$. Following a similar procedure as above, we also identify the antonym set $A_{\hat{P}}^A$ via WordNet, where $A_{\hat{P}}^A$ can serve as a supplementary of synonymic adjectives that describe P.

Finally, by merging all descriptive adjectives that are positively correlated with P and eliminating

Algorithm 1 NLI-based Adjectives Retrieval

INPUT: A, C, P, M
OUTPUT: A_c
$A_\text{c} \leftarrow \{\}$
$R_\text{init} \leftarrow M(P, C)$
for $a_i \in A$ **do**
    $S_i \leftarrow$ "I am a/an a_i person."
    $R_\text{nli} \leftarrow \text{NLI}(R_\text{init}, S_i)$
    if R_nli = ***Entailment*** **then**
        Add a_i to A_c
    end if
end for

those that are negatively correlated via

$$A = A_P^S \cup A_{\hat{P}}^A - A_P^A \cup A_{\hat{P}}^S, \tag{1}$$

we derive the final candidate adjective set A.

3.4 Context-aware Adjective Retrieval

The trait activation theory (Tett and Guterman, 2000) shows that specific traits are more likely to manifest in certain environments. For example, *Conscientiousness* may dominate in working environments, while *Extraversion* becomes more prominent in leisure contexts. We also observed that in personality inventories, such as the BFI-44 (John et al., 1991) and SD-3 (Jones and Paulhus, 2014), different items are designed to assess distinct traits. These inspired us that even for multi-trait personalities, only a subset of traits may be activated and expressed in specific scenarios. Therefore, we introduce Context-aware Adjective Retrieval, which selects adjectives describing the traits within P that are most likely to be expressed in the current context.

Specifically, when the LLM M interacts with a user in a particular context C (*e.g.*, the user input), we first prompt M to generate an initial response R_init based on P by vanilla prompting. Then, we use R_init to retrieve context-aware adjectives A_c from A through Natural Language Inference (NLI), where A_c describes the traits within P that are most likely to exhibit in C, as shown in Algorithm 1.

We use the pre-trained NLI model Debertav3 (He et al., 2021) to encode R_init and S_i and get their NLI results, where S_i wraps a_i as a complete sentence. NLI determines whether the semantic relationship between two sentences is Entailment, Contradiction, or Neutral. So, if the NLI result between R_init and S_i is Entailment, it implies that the a_i behind S_i is appropriate to describe the LLM's personality traits expressed under the current context C. If the NLI result between R_init and S_i is Contradiction or Neutral, it means that the personality traits described by a_i are either opposite to or irrelevant to them expressed by M in R_init.

Finally, the set of adjectives A_c that describe the personality traits expressed by M in the current context, along with the domain knowledge K_p that describes P (as introduced in Section 3.2), are used together as a system-level prompt to induce M to generate the response to the context C.

4 Experiment Design

To evaluate the performance of CACLP on multi-trait personality inducing, we conduct extensive experiments comparing our method with other baseline methods through personality inventories on various personality models. We will first introduce the baseline methods, the personality models, and finally the evaluation method we adopted.

4.1 Baseline Methods

We chose three different prompting methods to induce LLMs for multi-trait personality exhibition as baselines: Vanilla, Self-description, and Multi-description. The example prompts of baseline methods are shown in Appendix A.

Vanilla directly prompts LLMs with name (and the extent *i.e.*, high or low) of the personality traits. This simple yet effective prompting method is also applied in (Jiang et al., 2023b).

Self-description prompts LLMs to describe specified personality traits and uses the output to prompt LLMs to exhibit multi-trait personality. This method is inspired by the Chain-of-thought (Wei et al., 2022) method and was first proposed in (Huang et al., 2023).

Multi-description combines descriptions of multiple single personality traits as prompts for LLMs, where the descriptions are from psychological findings. As psychological findings have no comprehensive description for all combinations of multi-trait personalities, we design this baseline inspired by (Cui et al., 2023), who combines single-dimensional data to fine-tune LLMs to exhibit multi-dimensional MBTI personality types.

4.2 Personality Models

We adopt the Big Five Model (John et al., 1991), 16Personalities, and Dark Triad (Paulhus and

Williams, 2002), the three widely studied personality models in existing LLM personality research (Wen et al., 2024). All three models incorporate multiple personality traits.

The **Big Five Model** includes five distinct traits (*Agreeableness*, *Conscientiousness*, *Extroversion*, *Neuroticism*, and *Openness to experience*) to describe the personalities of individuals. It is widely studied in the psychological research of LLMs (tse Huang et al., 2023; Safdari et al., 2023; Ai et al., 2024; Pellert et al., 2022). **16Personalities** develops the NERIS model, which includes five traits to describe five aspects of personality: *Energy*, *Mind*, *Nature*, *Tactics*, and *Identity* with the acronym of MBTI to indicates the first four traits. It is widely used to assess LLM's MBTI type (Huang et al., 2023; Ai et al., 2024; Rao et al., 2023). Following previous work, we also induce and assess the results of the first 4 traits in our paper. **Dark Triad** describes three notably offensive, but non-pathological personality types: Machiavellianism (a manipulative attitude), Narcissism (excessive self-love), and Psychopathy (lack of empathy) (Paulhus and Williams, 2002). Dark Triad is commonly adopted to assess LLMs for safety concerns (Bodroza et al., 2023; Pellert et al., 2022).

4.3 Evaluation Method

4.3.1 Personality Inventories

Personality inventories provide a comprehensive assessment of personality exhibition. They are commonly used to evaluate the personalities of LLMs (Wen et al., 2024). Corresponding to the personality models, we select the Big Five Inventory (BFI-44, John et al. (1991)),16Personalities, and the Short Dark Triad (SD-3, Jones and Paulhus (2014)), the three most recognized Likert-scale personality inventories in existing literature as the assessment. The question examples of the personality inventories are shown in Appendix B.

BFI-44 is a 44-item questionnaire that measures the Big Five personality model. Each item assesses one of the five traits. **16personalities** is the most popular online MBTI[3] test that has been taken over 1.29 billion times. It comprises a total of 60 questions with seven different degrees of agreement ranging from agree to disagree. **SD-3** consists of 27 statements and is scored according to the level of agreement with the Dark Triad.

[3]The NERIS model, which uses the acronym format introduced by MBTI.

4.3.2 Evaluation Metric

For each personality model, we induce the LLM to exhibit all 2^n binary combinations of the n personality traits and calculate the overall accuracy Acc^n. For example, the Big Five model (with five traits) has 32 binary combinations, the 16Personalities model (with the first four traits) has 16 combinations, and the Dark Triad model (with three traits) has 8 combinations.

Specifically, for j-th binary combination, if all n traits are successfully induced as assessed by the inventory, $\text{acc}_j = 1$, otherwise, $\text{acc}_j = 0$. Then, Acc^n can be calculated by:

$$Acc^n = \sum_{j=1}^{2^n} \frac{1}{2^n} \text{acc}_j$$

Due to differences in inventory design, we employed distinct criteria to convert the questionnaire score of LLMs into acc_j for each inventory. For BFI-44, we compare the LLM's trait scores with the average human scores from the U.S. population (3,387,303 participants) reported by (Ebert et al., 2021): (E: 3.39, A: 3.78, C: 3.59, N: 2.90, O: 3.67). If the LLM's score for a trait was equal to or higher than the human average, it was labeled as "high"; otherwise, it was labeled as "low". This approach aligns with prior studies (Li et al., 2022; Jiang et al., 2023a). Similarly, we compare the LLM's scores to the weighted human average scores (Machiavellianism: 2.96, Narcissism: 2.97, Psychopathy: 2.09) derived from 7,863 participants across ten studies[4] for SD-3. This method was also adopted by (Li et al., 2022). For 16Personalities, following the scoring guidelines of 16Personalities and existing work (Huang et al., 2023), we use a threshold of 50 for each trait. Scores ≥ 50 were classified as "high", and scores < 50 as "low". Notably, 16Personalities employs an acronym system based on MBTI, where "low" and "high" for each dimension correspond to specific letters (e.g., Energy-low is "I", Energy-high is "E").

4.3.3 Implementation Details

To comprehensively validate the generality of our method, we conduct experiments on LLMs with various architectures and scales: Qwen2.5-3B-Instruct, Llama3.2-3B-Instruct, Llama3.1-8B-Instruct, Qwen2.5-72B-Instruct, and GPT-4o. This selection includes open-source and proprietary

[4]https://openpsychometrics.org/tests/SD3/results.php

Table 1: Multi-trait inducing accuracies on personality inventories. The highlighted parts indicate the accuracies of all traits are correctly induced in each personality model.

LLMs	Prompting Methods	Personality Inventories											
		BFI-44					SD-3			16Personalities			
		5	4	3	2	1	3	2	1	4	3	2	1
Qwen2.5 (3B)	Vanilla	0.69	0.30	0.01	-	-	0.33	0.45	0.23	0.06	0.20	0.39	0.29
	Self-description	0.04	0.17	0.33	0.30	0.12	0.23	0.33	0.36	0.05	0.14	0.28	0.26
	Multi-description	0.90	0.10	-	-	-	0.30	0.55	0.15	0.05	0.21	0.36	0.31
	CACLP	**0.96**	0.04	-	-	-	**0.58**	0.28	0.15	**0.06**	0.25	0.35	0.25
Llama3.2 (3B)	Vanilla	0.23	0.46	0.23	0.06	0.01	0.13	0.45	0.40	0.06	0.21	0.40	0.26
	Self-description	0.04	0.13	0.27	0.33	0.20	0.19	0.32	0.26	0.05	0.17	0.42	0.18
	Multi-description	0.34	0.49	0.13	0.03	0.01	0.18	0.48	0.33	0.06	0.25	0.36	0.26
	CACLP	**0.44**	0.41	0.15	-	-	**0.33**	0.55	0.13	**0.08**	0.22	0.39	0.24
Llama3.1 (8B)	Vanilla	0.82	0.18	-	-	-	0.18	0.45	0.35	0.55	0.39	0.06	-
	Self-description	0.04	0.12	0.36	0.35	0.13	0.22	0.31	0.19	0.47	0.23	0.13	0.12
	Multi-description	0.71	0.27	0.02	-	-	0.33	0.58	0.10	0.83	0.17	-	-
	CACLP	**0.84**	0.15	0.01	-	-	**0.63**	0.33	0.05	**0.91**	0.09	-	-
Qwen2.5 (72B)	Vanilla	1.00	-	-	-	-	0.63	0.37	-	0.94	0.06	-	-
	Self-description	0.03	0.25	0.25	0.25	0.19	0.13	0.44	0.36	0.47	0.32	0.12	0.17
	Multi-description	1.00	-	-	-	-	0.88	0.12	-	1.00	-	-	-
	CACLP	**1.00**	-	-	-	-	**0.88**	0.12	-	**1.00**	-	-	-
GPT-4o	Vanilla	1.00	-	-	-	-	0.92	0.03	-	1.00	-	-	-
	Self-description	0.35	0.42	0.23	-	-	0.57	0.30	0.05	0.97	0.03	-	-
	Multi-description	1.00	-	-	-	-	0.95	0.05	-	1.00	-	-	-
	CACLP	**1.00**	-	-	-	-	**1.00**	-	-	**1.00**	-	-	-

models spanning three orders of magnitude in parameter sizes from 3B to 72B+ (exact GPT-4o parameters undisclosed) [5].

To ensure fair and consistent comparisons across all experiments, we maintain identical hyperparameters (using the default settings provided by each model's API) when comparing our method with baseline approaches. Considering the potential impact of temperature settings on the performance of generative LLMs, we conduct each experiment with 10 independent trials inducing every personality trait combination. The final results are reported as the mean values across these 10 trials, ensuring statistical reliability and robustness.

5 Experiment Results and Analysis

We report the average results of inducing multiple personality traits using different methods on various LLMs in Table 1. For each personality model, besides Acc^n indicating the accuracies all n traits are correctly induced, we also report the accuracy Acc^i of correctly inducing i out of n traits to provide a holistic view, where $\sum_{i=0}^{n} Acc^i = 1$.

Firstly, we can observe that CACLP achieves the highest accuracies in correctly inducing all traits across all three personality models and all LLMs as base models. Notably, on GPT-4o, our method correctly induces all traits simultaneously with 100% accuracy on all 10 repeated experiments. This demonstrates the effectiveness of CACLP in inducing multiple personality traits concurrently. Next, we will analyze the experimental results in detail from the perspectives of LLMs, personality inventories, and baseline methods.

5.1 Results on Different LLMs

Experimental results show significant differences in multi-trait personality induction across LLMs of varying scales, while CACLP effectively mitigates the dependency on model parameter size.

Large models (*e.g.*, Qwen2.5-72B and GPT-4o) achieve almost perfect performance across all prompting methods (except for Self-description) on all inventories due to their superior language understanding and generation capabilities. Smaller models (*e.g.*, Qwen2.5-3B and Llama 3.2B) perform relatively weak and unstable with baseline methods. However, our method enables smaller models to achieve significant improvements over baseline methods in most results, even being competitive with larger models on repetitive experiments. For instance, $Acc^3 = 0.58$ of Qwen2.5-3B

[5] As all open-source LLMs are in the Instruct version, we omit "-Instruct" in subsequent result analyses for simplicity.

on SD-3 outperforms all baseline methods on Llama3.1-8B, and $Acc^5 = 0.96$ on BFI-44 can be comparable with Qwen2.5-72B and GPT-4o. Besides, our method consistently performs well across both open-source (Qwen, Llama) and proprietary (GPT-4o) models, demonstrating its architecture-agnostic generality.

5.2 Results on Different Personality Inventories

Different personality inventories measure different personality traits, varying in the number of questions, option settings, and scoring methods. However, CACLP achieves improvements over baseline methods across different inventories, which validates the generality of our approach.

BFI-44 measures the Big Five Model, which is the most widely studied personality model, so its average performance is better than the other two personality models. 16Personalities includes more questions (60 items) and more options per question (7 options), making it a challenging task for smaller models. Although our method can match or even surpass baseline methods, the results are still limited by the inherent capabilities of the models themselves. Additionally, during repeated experiments, we directly tested the inherent personality of Llama3.2-3B and found that it consistently exhibited an **ISTP** personality. Therefore, the poor inducing results on BFI-44 and 16Personalities may also be related to their stable alignment settings.

The Dark Triad traits measured by SD-3 conflict with the alignment goals of most LLMs, making them challenging to induce. Yet, our method improves Acc^3 by 106% on average across five LLMs. We speculate that the Context-aware Adjective Retrieval in our method can filter out dark personality descriptions irrelevant to the current context to some extent, thereby reducing the degree of conflict with alignment goals. The effective control enables LLMs to manage low-extent dark triad traits across various real-world applications.

5.3 Results on Different Baseline Methods

The three baseline methods exhibit varying performance in multi-trait inducing, which underscores the importance of the modules in CACLP.

Vanilla directly prompts the LLM with trait names. On large-scale models (*e.g.*, Qwen2.5-72B and GPT-4o), it achieves relatively good results. However, on smaller models (*e.g.*, Qwen2.5-3B and Llama3.2-3B), its accuracy is lower and less stable. CACLP outperforms Vanilla in most results, which highlights the importance of Knowledge-enhanced Adjective Generation.

Multi-description combines descriptions of single personality traits, providing comprehensive information to describe the target traits. In some scenarios, it shows significant improvements over other baseline methods (*e.g.*, Qwen2.5-3B on BFI-44 and Llama3.1-8B on SD-3 and 16Personalities). However, these improvements are not consistent. We speculate that though the Multi-description offers more comprehensive descriptions, the simple combination may introduce potential conflicts among multiple traits, making it difficult to accurately express the target personality traits. This highlights the importance of Contrasive Adjective Refinement in our method.

Self-description generally underperforms in most results. Although it leverages the model's generative capabilities to provide detailed descriptions of personality traits, most LLMs have not been fine-tuned in the personality domain, resulting in a limited understanding of personality traits. In contrast, our method, through Knowledge-enhanced Adjective Generation and Contrastive Adjective Refinement, and by selecting more appropriate trait expressions based on context, can more effectively guide the model to exhibit multidimensional personality traits. This approach addresses the limitations of the baseline methods and demonstrates superior performance across various scenarios.

6 Conclusion

In this paper, we propose Context-aware Contrastive Lexical Prompting (CACLP), a novel prompting method for multi-trait personality induction in LLMs inspired by the lexical hypothesis and trait activation theory of personality. CACLP resolves potential trait conflicts through lexical knowledge and can induce multi-trait personality adapted to diverse contexts by dynamically selecting context-aware adjectives, integrating psychological insights into LLM prompting. Experiments across multiple personality models (e.g., Big Five, 16Personalities, Dark Triad) demonstrate that CACLP consistently outperforms baseline methods across LLMs of varying architectures and scales and also achieves significant improvements in smaller models, which highlights its generality and adaptability to resource-constrained scenarios.

Limitations

While CACLP demonstrates promising results in multi-trait personality induction, our work has several limitations that need further exploration. First, the method relies on external lexical resources (e.g., WordNet) and pre-trained NLI models, which may limit its applicability to languages or cultural contexts with insufficient semantic knowledge bases. In future work, we may investigate how to enhance LLMs' intrinsic knowledge of personality to reduce dependency on external tools. Second, our evaluation relies on standardized personality inventories, which may not fully capture real-world conversational dynamics. A key reason is the lack of annotated conversational datasets for evaluation. We have noticed the recent surge in using LLMs for synthetic dialogue dataset generation and plan to explore this direction in the future to achieve a more comprehensive assessment of multi-trait personality induction.

References

Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. 2024. Is cognition and action consistent or not: Investigating large language model's personality.

Gordon W Allport and Henry S Odbert. 1936. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i.

Bojana Bodroza, Bojana M. Dinic, and Ljubisa Bojic. 2023. Personality testing of gpt-3: Limited temporal reliability, but highlighted social desirability of gpt-3's personality instruments results.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Joost CF de Winter, Tom Driessen, and Dimitra Dodou. 2024. The use of chatgpt for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, 228:112729.

Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537):eabf7117.

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

F Galton. 1884. Measurement of character. fortnightly review, 36, 179-185. *Galton17936Fortnightly Review1884*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023a. Evaluating and inducing personality in pre-trained language models. In *NeurIPS*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023b. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89.

Daniel N Jones and Delroy L Paulhus. 2014. Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, 21(1):28–41.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2023. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective.

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models.

Jianzhi Liu, Hexiang Gu, Tianyu Zheng, Liuyu Xiang, Huijia Wu, Jie Fu, and Zhaofeng He. 2024a. Dynamic generation of personalities with large language models.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024b. Personality-aware student simulation for conversational intelligent tutoring systems.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Walter Mischel, Yuichi Shoda, and Ozlem Ayduk. 2007. *Introduction to personality: Toward an integrative science of the person*. John Wiley & Sons.

Isabel Briggs Myers. 1962. *The Myers-Briggs Type Indicator: Manual (1962).* The Myers-Briggs Type Indicator: Manual (1962). Consulting Psychologists Press, Palo Alto, CA, US.

David Noever and Sam Hyams. 2023. Ai text-to-behavior: A study in steerability.

Irina Novikova. 2013. Trait, trait theory. *The encyclopedia of cross-cultural psychology*, 1:1–2.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

Delroy L Paulhus and Kevin M Williams. 2002. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563.

Max Pellert, Clemens Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2022. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT assess human personalities? a general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194, Singapore. Association for Computational Linguistics.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Gerard Saucier and Lewis R Goldberg. 1996. Evidence for the big five in analyses of familiar english personality adjectives. *European journal of Personality*, 10(1):61–77.

Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. 2024. The effects of embodiment and personality expression on learning in llm-based educational agents. *arXiv preprint arXiv:2407.10993*.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models.

Robert P Tett and Hal A Guterman. 2000. Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, 34(4):397–423.

Jen tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R. Lyu. 2023. Revisiting the reliability of psychological scales on large language models.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Dimitri Van der Linden, Jan te Nijenhuis, and Arnold B Bakker. 2010. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. 2024. Self-assessment, exhibition, and recognition: a review of personality in large language models. *arXiv preprint arXiv:2406.17624*.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Controllm: Crafting diverse personalities for language models.

Joshua Wilt and William Revelle. 2015. Affect, behaviour, cognition and desire in the big five: An analysis of item content and structure. *European journal of personality*, 29(4):478–497.

A Details of Baseline Methods

We present the prompt examples from baseline models (*e.g.*, Vanilla, Self-description, and Multi-description) to induce LLMs to exhibit multi-trait personalities on three personality models, as shown in Table 2.

Specifically, Vanilla directly names the personality of multiple traits. Prompts in Self-description are generated by prompting LLMs to describe a person with the specified personality through the Vanilla method. We formulate the Multi-description prompt by directly concatenating single-trait descriptions adopted by psychology domain knowledge. The different colors indicate the different single-trait descriptions. Here, single-trait descriptions for 16Personalities are obtained from the official website that defines the traits[6]. Single-trait descriptions for the Big Five Model are adopted from the Open-Source Psychometrics Project official website[7]. Single-trait descriptions for Dark Triad are generated by GPT-4 summarizing the papers (Paulhus and Williams, 2002; Jones and Paulhus, 2014; Johnson, 2014) that defined and discussed the Dark Triad, as these papers did not explicitly describe the high and low scores on each trait in short sentences.

B Question examples in Personality inventories

Table 3 shows the question examples in BFI-44, 16persoanlities, and SD-3, respectively. These questions are employed as the user-level prompts of LLMs when prompting them to exhibit multiple personality traits. To restrict the format of LLMs' answers to these questions, we also add instructions like "Please only answer with the option number" or "You can only reply with a number from 1 to 7" besides the original questions.

[6]https://www.16personalities.com/articles/our-theory
[7]https://openpsychometrics.org/

Table 2: Prompt Examples of Baseline Methods.

Baseline Methods	Big Five Model	16Personalities	Dark Triad
Vanilla	You are low on **Extraversion** , high on **Neuroticsm**, low on **Agreeableness**, low on **Conscientiousness**, and high on **Openness to Experience**.	Imagine you are an **ISTJ** person.	You are low on **Machiavellianism**, low on **Narcissism**, low on **Psychopathy**.
Self-description	You are a creative and sensitive individual who tends to be introverted, preferring quieter environments. Your high neuroticism makes you more prone to emotional distress, anxiety, and self-doubt. As someone low in agreeableness, you may come across as assertive or even confrontational, valuing your independence and honesty over social harmony. Your low conscientiousness suggests you might struggle with organization, planning, and self-discipline, often acting on impulse. However, your high openness to experience makes you curious, imaginative, and drawn to new ideas and experiences, which can be a source of both inspiration and frustration.	You are a practical and detail-oriented individual, often described as a behind-the-scenes worker. As an ISTJ, you value tradition, order, and stability, and you tend to approach life with a logical and analytical mindset. You prioritize efficiency and effectiveness, and you're not afraid to take charge and make decisions when necessary. Your reserved nature can sometimes make you appear aloof or unemotional, but you're actually deeply committed to your values and the people you care about. You're a loyal and hardworking individual who values hard work and dedication.	You are an individual with a relatively low score on the Dark Triad personality traits. This suggests that you are likely to be empathetic, cooperative, and honest in your interactions with others. You may value long-term relationships and be willing to compromise to maintain harmony. Your low levels of Machiavellianism imply that you are not manipulative or deceitful, and your low levels of Narcissism and Psychopathy indicate that you are not excessively self-focused or callous. Overall, you appear to be a genuinely caring and well-intentioned person.
Multi-description	You are introverted and reserved. You prefer to spend time alone or in small groups, and may feel uncomfortable in large social gatherings. You may also be less assertive and more cautious in your interactions with others. You tend to be more prone to negative emotions, such as anxiety, depression, and anger. You may be more reactive to stress and may find it difficult to cope with challenging situations. You may also exhibit a range of maladaptive behaviors, such as substance abuse or self-harm. You tend to be more competitive and skeptical. You may be less motivated to maintain social harmony and may be more likely to express your opinions forcefully, even if you may conflict with others. You tend to be more impulsive and disorganized. You may have difficulty setting and achieving goals, and may be more likely to engage in behaviors that are not in your best interest. You are characterized as being imaginative, curious, and open to new ideas and experiences. You tend to be intellectually curious and enjoy exploring new concepts and ideas. You may also exhibit a preference for creativity and aesthetics.	You prefer solitary activities and get exhausted by social interaction. You tend to be quite sensitive to external stimulation (e.g. sound, sight or smell) in general. You are highly practical, pragmatic and down-to-earth. You tend to have strong habits and focus on what is happening or has already happened. You focus on objectivity and rationality, prioritizing logic over emotions. You tend to hide your feelings and see efficiency as more important than cooperation. You are decisive, thorough and highly organized. You value clarity, predictability and closure, preferring structure and planning to spontaneity.	You are exceptionally honest, transparent, and guided by strong moral principles. You respect others' autonomy and have no interest in manipulation. You value sincerity and trust, making you someone others can rely on for straightforward and ethical interactions. You are genuinely humble and uninterested in self-promotion. You do not seek attention or admiration from others, often downplaying your own achievements. You're comfortable staying out of the spotlight and prefer to focus on others rather than asserting your own importance. You are highly empathetic, cautious, and mindful of others' well-being. You feel a strong sense of social responsibility and are deeply aware of the consequences of your actions. You likely avoid impulsive decisions and are motivated by a desire to help, not harm, those around you.

Table 3: Question Examples in Personality Inventories

BFI-44	16Personalities	SD-3
Given a statement about yourself, please choose from the following options to identify how accurately this statement describes you: 1. Very Inaccurate 2. Moderately Inaccurate 3. Neither Accurate Nor Inaccurate 4. Moderately Accurate 5. Very Accurate Please only answer with the option number. Here is the statement: **You are talkative.**	You will be presented with a statement to describe you. Please indicate the extent to which you agree with the statement on a scale from 1 to 7, with 1 being "strongly agree" and 7 being "strongly disagree." You can only reply with a number from 1 to 7. Here is the statement: **You usually prefer just doing what you feel like at any given moment instead of planning a particular daily routine.**	Given a statement below, please rate how much you agree with it: 1. Disagree 2. Slightly Disagree 3. Neutral 4. Slightly Agree 5. Agree Please only answer with the option number. Here is the statement: **Make sure your plans benefit yourself, not others.**