Emotional Labor Strategy Preferences in LLM Personas

Mohammad Saim and Tianyu Jiang
University of Cincinnati
saimmd@mail.uc.edu, tianyu.jiang@uc.edu

Abstract

Emotional labor is the effortful management of emotional displays to meet social or professional expectations. Personality traits have been correlated with emotional labor strategies, yet research on this link relies almost exclusively on self-report scales administered only in occupational settings. We investigate whether large language models injected with psychometrically grounded personas reproduce these personality-driven selection patterns across everyday social scenarios. We construct the first emotional labor strategy dataset of 500 socially situated events, each offering three behavioral choices corresponding to surface acting, deep acting, and genuine expression. We source 50 fictional characters from a large-scale personality repository and profile each through two parallel tracks: observer-rated bipolar adjective composites and in-character self-report items. Five LLMs evaluate all scenarios under both persona conditions. We find that models align more towards deep acting, and that Conscientiousness and Emotional Stability consistently predict this preference. Entropy analysis confirms that persona reliably influences the output and varies across models and emotions.

1 Introduction

The way individuals manage and display emotions in social contexts constitutes a central object of study across psychology, organizational behavior, and, more recently, computational linguistics. Hochschild (2012) introduced the concept of *emotional labor* (EL) to describe the effortful management of feelings to produce a publicly observable display that conforms to role expectations. Since then, researchers have consistently identified three principal strategies through which individuals execute this management (Figure 1): *surface acting* (SA), in which outward expression is modified without changing inner feeling; *deep acting* (DA),

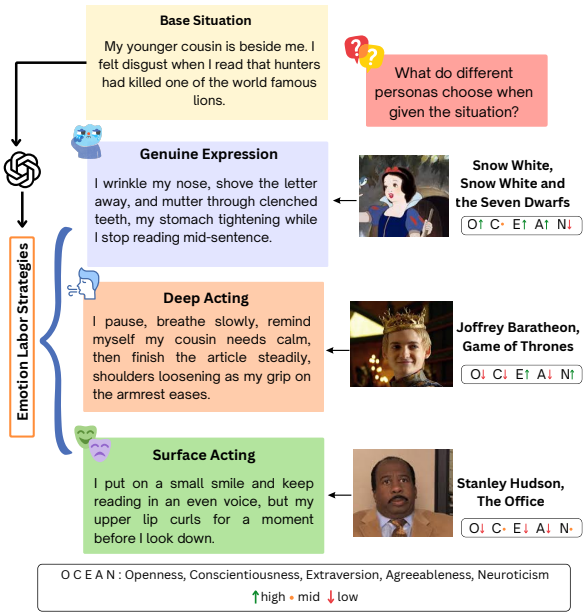

Figure 1: An example of emotional labor categories and what persona profiles choose from each category.

in which the felt emotion itself is reappraised so that it is followed by a genuine display of emotions; and *genuine expression* (GE) or naturally felt expressions, in which experienced emotion already matches what the situation calls for, and no regulatory effort is required (Ashforth and Humphrey, 1993; Diefendorff et al., 2005; Grandey, 2000). The mismatch between the regulated exterior and the persisting interior is what separates SA from the other two strategies, since the effort is spent on the display itself rather than on the feeling. In Genuine Expression, the felt emotion already matches what the situation calls for, so no internal adjustment occurs before the display. In Deep Acting, the felt emotion is actively reshaped through reframing, perspective-taking, or self-talk, so that regulation happens upstream of expression rather than at the surface. These three strategies have been employed in empirical work that shows they have different consequences for well-being, service quality, and

interpersonal perception (Kammeyer-Mueller et al., 2013; Grandey, 2003; Groth et al., 2009).

A well-established line of research in organizational psychology shows that personality traits can shape how people perceive and classify others' emotional displays (Terracciano et al., 2003; Furnes et al., 2019). These findings draw on the Big Five (OCEAN) framework—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Goldberg, 1992), which is an extensively validated taxonomy of human personality differences (McCrae and John, 1992). Individuals high in Agreeableness and Extraversion tend to perceive and enact more genuine or deep-acting strategies, whereas those high in Neuroticism are more prone to surface acting (Diefendorff et al., 2005; Austin et al., 2008; Kiffin-Petersen et al., 2011). However, researchers have largely assessed patterns of trait-linked EL preferences using self-report questionnaires/scales, in which participants rate statements about individual EL categories on Likert scales. Moreover, most sampling methods for analyzing EL strategies have been limited to specific healthcare and customer service contexts and have not been tested in everyday situations.

A growing body of work shows that LLMs, when prompted with personality descriptions, exhibit behavioral patterns that align with their assigned traits (Jiang et al., 2024; Sorokovikova et al., 2024; Li et al., 2024). These injected *personas* can modulate lexical choices, sentiment patterns, and judgment tendencies in ways that mirror the psychological literature on human personality (Peters and Matz, 2024; Matz et al., 2024). This provides a reliable substrate for controlled experimental simulation of human psychological variability.

Despite parallel advances in EL theory and LLM persona research, these two lines of work have not yet converged. **No prior study has examined how personality-injected LLM personas select among emotional labor strategies in socially situated scenarios.** If particular personality traits systematically shift which strategy a persona selects, then persona injection introduces a source of behavioral variance that may go unnoticed in downstream applications that rely on LLM-generated responses. The stakes of emotion-related misclassification are amplified by the growing use of LLMs in emotional support conversations, customer-service interactions, and everyday affect-sensitive domains, where inappropriate response or interpretation may negatively affect user well-being, communication outcomes, and even sway ethical judgments (Grandey and Sayre, 2019; Weidinger et al., 2021; Wu et al., 2025; Saim and Jiang, 2026).

In this work, we address this gap by building on an existing appraisal corpus to construct an emotional labor strategy (ELS) dataset of 500 sentences, each with annotated event descriptions followed by three choices of how an agent would enact them based on the three core strategies. In parallel, we employ fictional characters from TV/Film shows with character personality ratings on bipolar adjective pairs. To generalize and compare the persona profiles, we also administer the IPIP-50 questionnaire in-character, following established procedures for extracting Big Five scores from validated instruments (Goldberg et al., 2006; Jiang et al., 2024). We load the resulting persona profile into multiple LLMs for evaluation on the curated ELS dataset. For each sentence, the persona must select one of three options: surface acting (SA), deep acting (DA), or genuine expression (GE). We then analyze EL strategy choice patterns across the OCEAN dimensions of each persona, mapping which traits drive each persona's selection. We publicly release the code and the emotion labor strategy dataset.[1] As an overview, this paper makes the following contributions:

1. We introduce the first study to frame emotional labor strategy selection as a personality-conditioned task for LLM personas, in which each persona chooses how it would respond behaviorally to a social scenario, and to compare these choices against previously established human trends.

2. We build and release the first full-scale dataset on emotional labor strategies, comprising 500 scenarios across everyday affect and social contexts that involve choices among surface acting, deep acting, and genuine expressions.

3. We show that models prefer deep acting as the dominant strategy, and that Conscientiousness and Emotional Stability emerge as the consistent trait-level predictors across both persona tracks.

2 Related Works

Emotions in NLP and Emotional Labor. The computational study of emotion in text has matured

[1] https://github.com/cincynlp/emotion-labor

considerably over the past decade. Early work cast emotion recognition as a categorical classification problem, mapping text to discrete labels derived from foundational theories (Strapparava and Mihalcea, 2007; Mohammad et al., 2018; Hofmann et al., 2020). More theoretically grounded approaches have moved beyond flat emotion labels toward appraisal-based frameworks, which treat emotion as the product of a person's cognitive evaluation of an event along dimensions such as novelty, relevance, and coping potential (Smith and Ellsworth, 1985; Scherer, 2001). Troiano et al. (2023) formalize this shift in a comprehensive corpus study, annotating event descriptions and showing that appraisal variables reliably improve emotion categorization in text. This line of work is directly relevant to ours: appraisal theory frames emotion as a *situated evaluation* of a social event, which maps naturally onto the idea that emotional labor strategies represent different stances a person takes toward a situation. Another line of work in emotion recognition evaluates or applies embodied or physiological indicators rather than explicit emotion labels (Zhuang et al., 2024; Duong et al., 2025; Saim et al., 2025). This embodied perspective aligns with our dataset design, since the strategy options involve behavioral or physiological details.

In organizational psychology, measurement of emotional labor (EL) has relied almost exclusively on self-report instruments administered to workers in service roles. Brotheridge and Lee (2003) developed the Emotional Labour Scale, a 15-item questionnaire that assesses the frequency, intensity, variety, duration, and both surface and deep acting (SA and DA); this instrument has become one of the most widely adopted tools in the field. Diefendorff et al. (2005) later extended this framework to a three-factor structure that treats naturally felt expression as a dimensionally distinct strategy alongside SA and DA, thereby establishing the taxonomy used in our study. A core gap in research on EL strategies is the lack of a corpus that goes beyond narrow professional roles such as customer service and healthcare workers. This leaves open the question of whether the three-strategy taxonomy applies to the broader, non-occupational interactions that constitute most of daily social life.

Persona Injection and Personality in LLMs. A rapidly growing body of NLP work has examined whether LLMs can stably express and enact personality traits when prompted with persona descriptions. Jiang et al. (2024) showed that LLM personas assigned Big Five profiles produce BFI self-report scores and writing samples that align with their designated traits, with large effect sizes across all five dimensions. Sorokovikova et al. (2024) replicated this pattern across multiple open-source models, and Tseng et al. (2024) provides a comprehensive taxonomy of the field, distinguishing role-playing personas (where LLMs adopt assigned identities) from personalization settings where models adapt to user traits.

On persona traits, Huang and Hadfi (2024) found that OCEAN-grounded LLM agents reproduce human-like personality-linked patterns in bilateral negotiations, with Agreeableness promoting cooperative outcomes and Neuroticism leading to less favorable outcomes. Agreeableness and Extraversion consistently predict greater deep acting and genuine expression, while Neuroticism predicts surface acting (Austin et al., 2008; Kiffin-Petersen et al., 2011; Yeh et al., 2020). Recent mechanistic work finds that Big-Five traits are encoded as recoverable linear directions in LLM residual streams, and that steering along these directions produces reliable, trait-consistent behavioral shifts (Frising and Balcells, 2026). Taken together, this literature establishes that personality injection is a reliable lever on LLM behavior. However, most measurement work is limited to self-reported behavior within occupational samples. No work examines which trait-conditioned LLMs select as emotional regulation strategies when evaluated in socially situated EL scenarios. We fill this gap by presenting OCEAN-grounded personas with EL-annotated stimuli and measuring whether their strategy choices mirror the personality-EL associations documented in human literature and surveys.

3 Methodology

We begin by describing the emotional labor strategy (ELS) dataset, followed by the pipeline for persona traits. We then evaluate the characters, loaded with distinct traits, on the ELS dataset to analyze how each character's strategy preferences vary across emotional labor scenarios.

3.1 Emotion Labor Strategy Dataset

We construct the ELS dataset by building on the emotion appraisal corpus (Troiano et al., 2023), a collection of sentences annotated with appraisal dimensions and a single felt emotion label drawn

Strategy	Definition
Surface Acting (SA)	The person performs a different emotion outwardly while the original feeling persists internally.
Deep Acting (DA)	The person genuinely shifts their internal state through reframing, perspective-taking, or self-talk before responding.
Genuine Expression (GE)	The person expresses the naturally felt emotion directly with no attempt to regulate or conceal it.

Table 1: Definitions of the three emotional labor strategies used in dataset construction.

Example Scenario
I felt **fear** when I was being bullied at work and didn't think there was going to be an end to it. *A newer coworker was standing nearby, watching how I would respond.*
SA: I lifted my chin and gave a short laugh, but my hands shook once against the desk as I answered.
DA: I reminded myself their words were about their own problems, slowed my breathing, and answered in an even voice with my shoulders settling.
GE: I stepped back, gripped the edge of the desk, and said they needed to stop, my voice breaking in the middle.

Table 2: An example scenario from the dataset. The *italicized* portion shows the social context added to the original corpus.

from seven categories: *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness*, and *shame*. These labels reflect the narrator's internal affective state. We preserve the original dataset's uniform emotion split and filter to a final set of 500 sentences.

Social Context Augmentation. Many sentences in the corpus describe internal affective states without situating them in a social encounter. Since emotional labor is fundamentally an interpersonal phenomenon, i.e., it arises when a person regulates their emotional display in the presence of or in response to another social agent (Hochschild, 2012; Grandey, 2000), we add a social agent or context to each scenario. We augment each sentence with a brief context that introduces a second agent and coherent background information that creates a plausible reason to regulate one's emotional expression. The context is appended directly to the original sentence to produce a single, continuous narrative that serves as the scenario stem for all downstream steps.

Emotional Labor Strategy Options. With the social context in place, each scenario presents a person experiencing a felt emotion in the presence of another agent. We extend each scenario into a three-way choice item by generating one behavioral response option for each of the three emotional labor strategies: *surface acting* (SA), *deep acting* (DA), and *genuine expression* (GE) or naturally felt emotions. Table 1 defines each of the three emotional labor strategies represented in the dataset. We employ the GPT-5.4 model to generate the context and the three-way choices. No single correct response or ground truth exists, as the task is designed to elicit a *preference* for each persona.

The generation produces all three options as single sentences that complete the scenario stem naturally, without using any emotion labels or adjectives, conveying emotional state only through behavioral, physical, or physiological detail. Table 2 shows an example set from the dataset. The final dataset item consists of the augmented scenario and three behavioral completion options. The final dataset was manually validated to (i) verify coherence between the added context and the scenario narrative, (ii) determine whether each option instantiated its strategy without using emotion words, and (iii) determine whether the three options were distinguishable. We took several steps to mitigate concerns about subjective bias. The strategy generation was tightly constrained by the prompts (Appendix A), which enforced structural rules such as prohibiting emotional adjectives, requiring a single leakage cue for Surface Acting, and mandating behavioral rather than affective descriptions. This offloads much of the strategy-fidelity check from human judgment onto the generation constraints themselves. Beyond its use in this study, the dataset is the first resource to use emotion labor categories as contrastive behavioral options in naturalistic, non-occupational scenarios and can support future work on emotion regulation modeling, personality-conditioned generation, and computational appraisal research.

3.2 Persona Construction

The design of a persona profile with only a character's name and source work risks activating whatever personality representation the model has internalized during pretraining, which may be inconsistent across models. Therefore, we construct persona profiles through two parallel experimental

tracks. Both tracks use the same set of characters but differ in how personality information is sourced. This design allows us to (i) avoid any bias from relying on a single framework and (ii) examine whether trait-strategy associations are aligned across two distinct representations of personality.

Character Pool and BAP Profiling. We source character personality data from the Open Psychometrics Statistical "Which Character" Personality Quiz dataset (Open-Source Psychometrics Project, 2022), which contains crowd-sourced personality ratings for over 2,000 fictional characters across film and television. Each character is rated by human respondents on 500 Bipolar Adjective Pairs (BAPs), a psycho-lexical instrument in which each item presents two semantically opposing adjectives (e.g., *relaxed–tense*, *kind–cruel*) and human raters place the character on a continuous scale of 1–100. Since using every BAP as a trait in the model might introduce too much noise for the LLM to process, we select BAPs based on standard deviation.

To ground our trait markers in a theoretically validated source, we verify each BAP adjective against the original adjective list reported by Goldberg (1992). We retain only those adjective pairs for which at least one pole produces an exact match with Goldberg's taxonomy. This filtering yields a subset of 40 verified BAP markers, which ensures that every trait dimension we use is validated and has a well-established structural relationship to the Big Five OCEAN dimensions.

We filter characters from the full pool by ranking them on the standard deviation of their BAP scores. Characters with high score variance across adjective pairs have more personality-differentiated profiles, making them better suited to testing trait-driven behavioral differences. We then apply greedy farthest-point sampling in the 40-dimensional BAP space, iteratively selecting the character that is maximally distant from all previously selected characters. This produces a final set of 50 characters. We chose 50 characters by prioritizing personality diversity over sample size in the 40-item BAP space. This method selects characters that are maximally spread across the personality space rather than clustering around common archetypes. The overall BAP-persona block for each character consists of the character name, their source work, and their human-rated scores on the 40 verified BAP items, each expressed on a 1–100 scale. This block is passed directly to the model as part of the persona injection prompt in the ELS evaluation task.

	Track-1 (BAP)	**Track-2 (IPIP-50)**
Source	Human-rated crowd scores	Model self-report in-character
Instrument	40 bipolar adjective pairs	50 standard items
Scale	1–100 continuous	1–5 Likert
Item Example	*relaxed–tense*	*"I am the life of the party"*

Table 3: Comparison of the two persona tracks.

IPIP-50 Profiling. Relying solely on human-rated BAP scores may underrepresent traits that are internally characterized. The second track addresses this by eliciting in-character self-reports on a validated self-report instrument. Vazire (2010) presented the self-other knowledge asymmetry (SOKA) model, which shows that observer ratings more reliably capture visible, behaviorally expressed traits such as extraversion, while self-reports capture internal states and motivations that are less accessible to outside observers, such as neuroticism and openness. The IPIP-50 is a 50-item public-domain scale with 10 items per OCEAN dimension (Goldberg et al., 2006). We administer it to each model with the character's name and source work and ask it to rate each IPIP item on a 1–5 Likert scale from the character's perspective (1 = *Very Inaccurate*, 5 = *Very Accurate*). The character pool remains the same across the BAP-persona to ensure the analysis focuses solely on methodological differences. The full list of filtered characters, BAP terms, and IPIP items is listed in Appendix B.

3.3 Evaluation and Model Selection

We end up with two parallel persona experimentation tracks. A short summary is given in Table 3. These are evaluated on the 500 sentences of the ELS dataset. We primarily study ELS choice patterns across different models and their correlations with the OCEAN categories. We employ a suite of five LLMs: Qwen-3-8B and Qwen-3-32B (Yang et al., 2025), GPT-5.4 (OpenAI, 2026), Deepseek-V4-Flash (DeepSeek-AI, 2026) and Gemma-4-31B (Google DeepMind, 2026). We prompt these models to choose one of the three EL strategies: SA, DA, or GE.

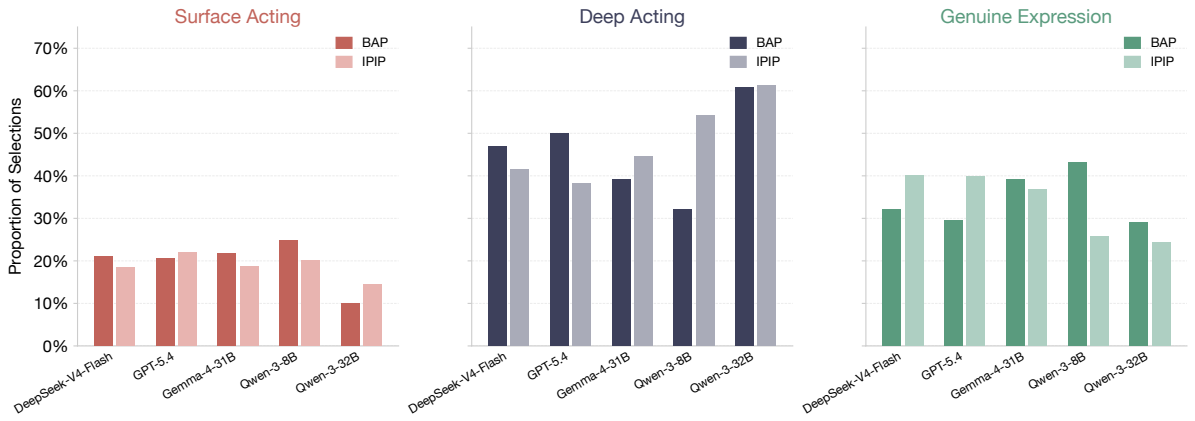

Figure 2: Aggregate distribution of emotion labor strategy across models and persona tracks.

4 Results and Analysis

We present the findings of our study across three levels of analysis. We begin by characterizing the aggregate distribution of ELS selections across all five models and both persona-induction tracks (BAP and IPIP-50). We then report Big Five trait–strategy correlations that connect persona-level personality profiles to ELS preferences and draw comparisons to organizational psychology benchmarks. Finally, we study the impact of personas on the reliability with which they drive ELS scenarios.

4.1 Aggregate ELS Distributions Across Models and Persona Tracks

Deep Acting is the preferred response compared to SA and GE. We first examine how each of the five models distributes its ELS selections across 500 scenarios when conditioned on 50 character personas. Figure 2 reports the proportion of each ELS category selection for each model under both the BAP and IPIP-50 persona tracks.

Notably, DA is the modal strategy for eight of the ten model-track combinations. DA proportions range from approximately 39% (GPT-5.4, IPIP track) to 61% (Qwen-3-32B, both tracks). This establishes DA as the dominant default across the model pool. This preference casts the majority of diverse agents as effortful and authentic. Models trained on large corpora of psychological and organizational text appear to encode DA as the prototypically appropriate regulatory response. The encoding appears to serve as a prior that shapes strategy selection beyond any personality variation introduced by the injected persona. In contrast, SA selections remain low across all models. In psychology literature, SA has consistently been identified as a prevalent emotional labor strategy in service-oriented occupations, especially when workers experience emotional dissonance between authentic feelings and institutionally required emotional displays (Grandey, 2000; Mesmer-Magnus et al., 2012). We hypothesize that models suppress the SA strategy not because persona profiles steer them away from it, but because SA may carry a socially undesirable signal in natural language (emotion suppression).

BAP and IPIP-50 tracks show moderate convergence. Across the model pool, both persona tracks preserve the same ordinal ranking of strategies (DA $>$ GE $>$ SA in most cases), providing a baseline level of convergent validity between the observer-rated and self-reported persona inductions. However, quantitative differences between tracks are meaningful and model-dependent. For GPT-5.4 and DeepSeek-V4-Flash, the IPIP-50 track redistributes mass from DA toward GE relative to the BAP track: GPT-5.4 drops from approximately 51% to 39% on DA while GE rises from approximately 30% to 40%. DeepSeek-V4-Flash shows a smaller but directionally consistent shift. This pattern accords with the self-other agreement work, which documents that observer-rated and self-reported personality tap partially distinct variance in the same underlying constructs (Vazire, 2010; Connelly and Ones, 2010). Observer ratings (BAP track) emphasize behaviorally visible regularities and external presentation. They may therefore bias the injected persona toward more deliberate, socially regulated strategies such as DA. Self-report ratings (IPIP-50 track) introduce a first-person affective framing that may allow greater expression of internal states.

	Reliability		BAP (Observer-Rated)			IPIP (Self-Report)		
Trait	α_{BAP}	α_{IPIP}	**SA**	**DA**	**GE**	**SA**	**DA**	**GE**
Openness	0.70	0.95	$+0.064$	-0.013	-0.032	-0.260	$+0.121$	-0.078
Conscientiousness	0.88	0.98	-0.446^{**}	$+0.597^{***}$	-0.570^{***}	-0.345^{*}	$+0.649^{***}$	-0.592^{***}
Extraversion	0.83	0.98	$+0.110$	-0.203	$+0.184$	$+0.200$	-0.299^{*}	$+0.232$
Agreeableness	0.95	0.97	-0.276	$+0.144$	-0.072	-0.398^{**}	$+0.372^{**}$	-0.263
Emotional Stability	0.76	0.97	-0.350^{*}	$+0.522^{***}$	-0.517^{***}	-0.311^{*}	$+0.850^{***}$	-0.883^{***}

Table 4: Spearman correlations between OCEAN trait composites and emotional labor strategy proportions for the BAP and IPIP tracks. Cronbach's α is reported for BAP and IPIP composites. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

We also computed pairwise inter-model agreement on both evaluation tracks. The highest agreement is observed between GPT-5.4 and Gemma-4-31B, with Cohen's $\kappa = 0.545$; per-category agreement is higher for GE and DA but substantially lower for SA. Agreement scores between all models are provided in Appendix C.5.

4.2 Trait–Strategy Correlations

We study how personality shapes ELS selection by computing Spearman rank correlations between each Big Five composite and each strategy proportion across the 50 evaluated characters. Each BAP and IPIP item is grouped under one of the OCEAN traits as shown in Table 7 and Table 8. The central question we examine is whether each OCEAN trait correlates similarly with the three ELS, as established in previous studies, and whether the two persona tracks yield convergent or divergent mappings. We report results for GPT-5.4 (Table 4); the model comparison to Deepseek-V4-Flash can be found in Appendix C.1.

Two traits exhibit consistent, significant correlations across both measurement tracks: **Conscientiousness (C)** and **Emotional Stability (ES)** (the inverse of Neuroticism). Both traits show the same directional pattern, i.e., high C and high ES predict more DA and less SA and GE. For Conscientiousness, the BAP track yields $\rho = -0.446$ (SA), +0.597 (DA), -0.570 (GE), all significant at $p < .001$ or $p < .01$, and the IPIP track replicates this pattern with comparable magnitude. The same signature holds for ES. The model appears to encode DA as the disciplined, regulated response. The characters who are organized, reliable, and emotionally secure preferentially enact deep acting rather than surface performance or unguarded expression. This is consistent with the general notion reported in human psychological studies that conscientiousness and emotional stability both correlate negatively with SA and positively with DA (Austin et al., 2008; Judge et al., 2009; Mesmer-Magnus et al., 2012).

We also report a negative correlation between conscientiousness and genuine expression, whereas previous human studies report a positive correlation, on the theory that workers with higher conscientiousness scores internalize job demands and naturally align their felt emotions with role requirements. They do not need to regulate because their internal state already matches the display norm. We hypothesize that the model encodes conscientiousness not as internalized role alignment but as an active, effortful self-regulation, much like a careful worker who writes a draft before sending an email when a spontaneous reply would do. We plan to form our future work based on mechanistic insights from open-source models to validate this finding, following work that recovers independent social-response directions in LLM representations (Yao et al., 2026; Vennemeyer et al., 2026).

The BAP track returns no significant **Agreeableness (A)** effects, while the IPIP track produces A–SA ($\rho = -0.398$, $p < .01$) and A–DA ($\rho = +0.372$, $p < .01$). Agreeableness is a trait humans understand better from the inside; warm, cooperative intentions are more readily disclosed in self-report than inferred by external observers from behavioral adjectives (Vazire, 2010). In this reading, the IPIP persona activates an agreeableness signal that the observer-rated BAP does not encode with sufficient specificity. **Openness** produces no significant correlations in either track. Extraversion yields only one marginal IPIP effect with DA, suggesting that more extroverted personas are slightly less inclined toward deep acting.

We also determine whether an observed trait–strategy association reflects a genuine pattern or a

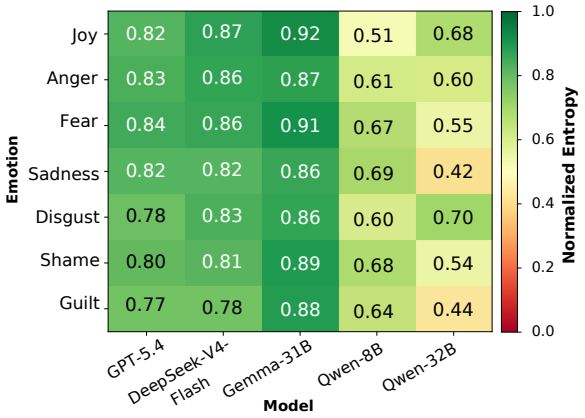

Figure 3: Persona influence by emotion and model (BAP track) depicted by Shannon Entropy.

statistical-only result of the response format (three-way choice). Therefore, we replicate the evaluation by asking the model to rate each of the three EL categories based on how likely they are to choose that scenario, rather than forcing it. The details and correlations are given in Appendix C.2. We find that all correlations broadly follow the same directional pattern. Convergence between a three-way choice and Likert correlations strengthens the claim that observed patterns reflect personality-driven strategy preferences.

Internal consistency. We also validate our findings by assessing the reliability of each trait composite. Cronbach's α measures the extent of inter-correlation of items assigned to a given trait scale. A high α means multiple BAP or IPIP items under one of the OCEAN traits pull in the same direction, and the composite score is a stable summary of the underlying construct. BAP composites yield moderate-to-high reliability (α = 0.70–0.95), with conscientiousness and agreeableness showing the strongest internal consistency. IPIP-50 composites achieve uniformly high reliability across all five traits. This also explains the ES–DA (ρ = +0.850) and ES–GE (ρ = $-$0.883) large correlations.

4.3 Is Persona Influential Enough?

Entropy Analysis. A distributional dominance result does not tell us whether personality drives meaningful divergence in strategy selection, or whether the emotional situation itself overrides individual differences. To test this, we compute the Shannon entropy of the SA/DA/GE distribution across all 50 characters for each scenario. We bound scores between 0 (scenario-dominant) and 1 (personality-sensitive), and report mean normalized entropy aggregated by the felt emotion category from the dataset. We report BAP-track results here to analyze human-rated divergence. Character-level entropy traits are discussed in Appendix C.3. IPIP entropy results appear in Appendix C.4.

Strategy	Vanilla	Character-Only	BAP-Only
Surface Acting	10.0%	29.5%	20.5%
Deep Acting	68.0%	36.6%	36.9%
Genuine Expression	22.0%	34.0%	42.6%

Table 5: Strategy proportions under vanilla (no persona), name-only, and BAP trait-only persona injection, evaluated on a 50-scenario subset across 50 characters (GPT-5.4).

Figure 3 reveals that most models across each emotion category form a high-entropy cluster, indicating that personality meaningfully differentiates strategy selection in these models. Within the high-entropy clusters, Joy, Fear, and Anger sustain high entropy (0.82–0.92), while Disgust and Guilt sit slightly lower. For Qwen-32B, entropy ranges from 0.42 (Sadness) to 0.70 (Disgust), which shows that the particular emotion and scenario also matter more than *who* the persona is for this model.

Ablation. We further probe the drivers of strategy selection by running a three-condition ablation on a 50-scenario subset spanning all 50 characters. In the vanilla condition, we provide only the scenario and the three response options with no character identity or trait information. In the character-only condition, we inject the character's name and source work so the model relies entirely on its pre-training knowledge. In the BAP-only condition, we inject only the 40 adjective-pair scores without naming any character. Table 5 reports aggregate strategy proportions across all three conditions.

Without any persona signal, deep acting has the majority share of predictions while surface acting falls to just 10%. This concentration confirms that DA serves as a strong default encoded in the model's training distribution, rather than a product of persona conditioning. Introducing a persona of any kind redistributes this mass. Character-only prompting cuts DA predictions by almost half and raises both SA and GE to near-uniform levels, which suggests that even name-level identity priming is sufficient to disrupt the DA default and provide a wider range of regulatory responses. BAP-only prompting produces a comparable DA proportion (36.9%) but with higher GE numbers. Trait content, therefore, does not simply diversify strat-

egy selection the way name recognition does. It selectively suppresses surface acting, consistent with the trait–strategy correlations reported in Section 4.2, in which high Conscientiousness and Emotional Stability predict less SA.

5 Conclusion

This study demonstrates that personality-injected LLM personas produce reliable differences in the selection of emotional labor strategies across everyday social scenarios. We introduced the first dataset that frames surface acting, deep acting, and genuine expression as situated behavioral choices outside occupational contexts. Our two-track persona design, combining observer-rated adjective profiles with in-character self-reports, reveals that Conscientiousness and Emotional Stability consistently drive strategy preferences across five models, while Agreeableness surfaces only through self-report induction. LLMs show a preference for deep acting, suggesting that training corpora encode this strategy as a normative regulatory response. Conscientiousness suppresses genuine expression in LLM personas, inverting the positive association documented in previous human-evaluated reports. Future work can probe this divergence through the mechanistic interpretability of open-source models and extend evaluation to cross-cultural settings.

Limitations

We acknowledge the constraints on the scope and generalization of our findings. We evaluate personas constructed from fictional characters, whose personality profiles reflect crowd-sourced perceptions rather than ground-truth personality measures. The dataset, while grounded in a published corpus, relies on synthetic augmentation for social context and strategy options, which may not fully capture the complexity of real emotional labor encounters. Our scenarios are English-only and culturally situated, leaving open the question of whether these trait-strategy patterns hold across languages and cultural norms. Finally, our findings remain correlational, and validating the causal mechanisms of trait-strategy associations will require a mechanistic interpretability framework for future work.

Acknowledgments

We thank the CincyNLP group for their suggestions and feedback. We also thank the anonymous EMNLP reviewers for their insightful suggestions.

References

Blake E. Ashforth and Ronald H. Humphrey. 1993. Emotional labor in service roles: The influence of identity. *Academy of Management Review*, 18(1):88–115.

Elizabeth J. Austin, Timothy C. P. Dore, and Katharine M. O'Donovan. 2008. Associations of personality and emotional intelligence with display rule perceptions and emotional labour. *Personality and Individual Differences*, 44(3):679–688.

Céleste M. Brotheridge and Raymond T. Lee. 2003. Development and validation of the emotional labour scale. *Journal of Occupational and Organizational Psychology*, 76(3):365–379.

Brian S Connelly and Deniz S Ones. 2010. An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, 136(6):1092–1122.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. *Preprint*, arXiv:2606.19348.

James M. Diefendorff, Meredith H. Croyle, and Robin H. Gosserand. 2005. The dimensionality and antecedents of emotional labor strategies. *Journal of Vocational Behavior*, 66(2):339–357.

Phan Anh Duong, Cat Luong, Divyesh Bommana, and Tianyu Jiang. 2025. CHEER-Ekman: Fine-grained embodied emotion classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

Michel Frising and Daniel Balcells. 2026. Linear personality probing and steering in llms: A big five study. *Preprint*, arXiv:2512.17639.

Desire Furnes, Hege Berg, Rachel M Mitchell, and Silke Paulmann. 2019. Exploring the effects of personality traits on the perception of emotions from prosody. *Frontiers in psychology*, 10:184.

Lewis R. Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26–42.

Lewis R. Goldberg, John A. Johnson, Herbert W. Eber, Robert Hogan, Michael C. Ashton, C. Robert Cloninger, and Harrison G. Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96.

Google DeepMind. 2026. Gemma 4: Our most capable open models to date.

Alicia A. Grandey. 2000. Emotion regulation in the workplace: A new way to conceptualize emotional labor. *Journal of Occupational Health Psychology*, 5(1):95–110.

Alicia A. Grandey. 2003. When the show must go on: Surface acting and deep acting as determinants of emotional exhaustion and peer-rated service delivery. *Academy of Management Journal*, 46(1):86–96.

Alicia A. Grandey and Gordon M. Sayre. 2019. Emotional labor: Regulating emotions for a wage. *Current Directions in Psychological Science*, 28(2):131–137.

Markus Groth, Thorsten Hennig-Thurau, and Gianfranco Walsh. 2009. Customer reactions to emotional labor: The roles of employee acting strategies and customer detection accuracy. *Academy of Management Journal*, 52(5):958–974.

Arlie Russell Hochschild. 2012. *The Managed Heart: Commercialization of Human Feeling*, 1 edition. University of California Press.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*.

Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? a simulation based on large language models. In *Findings of the Association for Computational Linguistics (Findings of EMNLP 2024)*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics (Findings of NAACL 2024)*.

Timothy A Judge, Erin Fluegge Woolf, and Charlice Hurst. 2009. Is emotional labor more difficult for some than for others? a multilevel, experience-sampling study. *Personnel psychology*, 62(1):57–88.

John D. Kammeyer-Mueller, Alex L. Rubenstein, David M. Long, Michael A. Odio, Brooke R. Buckman, Yiwen Zhang, and Marie D.K. Halvorsen-Ganepola. 2013. A meta-analytic structural model of dispositonal affectivity and emotional labor. *Personnel Psychology*, 66:47–90.

Sandra A. Kiffin-Petersen, Catherine L. Jordan, and Geoffrey N. Soutar. 2011. The big five, emotional exhaustion and citizenship behaviors in service settings: The mediating role of emotional labor. *Personality and Individual Differences*, 50(1):43–48.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024)*.

Sandra C. Matz, Jacob D. Teeny, Sumer S. Vaid, Heinrich Peters, Gabriella M. Harari, and Moran Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14:4692.

Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Jessica R Mesmer-Magnus, Leslie A DeChurch, and Amy Wax. 2012. Moving emotional labor beyond surface and deep acting: A discordance–congruence perspective. *Organizational Psychology Review*, 2(1):6–53.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*.

Open-Source Psychometrics Project. 2022. Data from the statistical "which character" personality quiz.

OpenAI. 2026. Introducing gpt-5.4.

Heinrich Peters and Sandra Matz. 2024. Large language models can infer psychological dispositions of social media users. *Preprint*, arXiv:2309.08631.

Mohammad Saim, Phan Anh Duong, Cat Luong, Aniket Bhanderi, and Tianyu Jiang. 2025. Anatomy of a feeling: Narrating embodied emotions via large vision-language models. In *Findings of the Association for Computational Linguistics (Findings of EMNLP 2025)*.

Mohammad Saim and Tianyu Jiang. 2026. Do emotions influence moral judgment in large language models? In *Findings of the Association for Computational Linguistics (Findings of ACL 2026)*.

Klaus R. Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. In *Appraisal Processes in Emotion: Theory, Methods, Research*, pages 92–120. Oxford University Press.

Craig A. Smith and Phoebe C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838.

Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. LLMs simulate Big5 personality traits: Further evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems*.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

Antonio Terracciano, Marcellus Merritt, Alan B Zonderman, and Michele K Evans. 2003. Personality traits and sex differences in emotion recognition among african americans and caucasians. *Annals of the New York Academy of Sciences*, 1000:309–312.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics (Findings of EMNLP 2024)*.

Simine Vazire. 2010. Who knows what about a person? the self–other knowledge asymmetry (SOKA) model. *Journal of personality and social psychology*, 98(2):281.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. 2026. Sycophancy is not one thing: Causal separation of sycophantic behaviors in LLMs. *Preprint*, arXiv:2509.21305.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From personas to talks: Revisiting the impact of personas on LLM-synthesized emotional support conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Louie Hong Yao, Vishesh Anand, Yuan Zhuang, and Tianyu Jiang. 2026. Rhetorical questions in LLM representations: A linear probing study. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*.

Shu-Chuan Jennifer Yeh, Shih-Hua Sarah Chen, Kuo-Shu Yuan, Willy Chou, and Thomas TH Wan. 2020. Emotional labor in health care: The moderating roles of personality and the mediating role of sleep on job performance and satisfaction. *Frontiers in Psychology*, 11:574898.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2024. My heart skipped a beat! recognizing expressions of embodied emotion in natural language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024)*.

A Prompts

We designed two prompts for this study. The dataset augmentation prompt (Prompt A.1) instructed GPT-5.4 to extend each sentence from the original corpus with a one-sentence social context introducing display pressure and to generate three behavioral response options corresponding to surface acting, deep acting, and genuine expression. To prevent label leakage, the prompt explicitly prohibited emotion adjectives and labels, requiring the model to convey internal states exclusively through physical and behavioral cues. The persona evaluation prompt (Prompt A.2) presented each fictional character's BAP trait profile as a scored bipolar adjective list and asked the model to select, in character, the single response option it would most naturally choose. We constrained the output to a single letter (A, B, or C) to enforce a clean forced-choice response and eliminate free-text ambiguity in downstream analysis. A similar prompt was repeated for the IPIP-track, the only difference being that the IPIP block had Likert ratings (1-5) instead of adjective pair scores (1-100).

B Strategy selection choices

B.1 Selected BAP Items and OCEAN Composite Construction

Table 7 reports the 40 verified BAP adjective pairs retained after filtering the full 500-item instrument against Goldberg's taxonomy. We retain only pairs for which at least one pole maps exactly onto an adjective appearing in Goldberg's published marker list, which ensures that every trait signal we inject into a persona has a well-established structural relationship to the Big Five (OCEAN) traits. The filtering reduces the raw instrument from 500 to 40 items, distributed across five OCEAN dimensions: Openness (8 items), Conscientiousness (7 items), Extraversion (8 items), Agreeableness (8 items), and Emotional Stability (9 items). We report the inverse of the Neuroticism dimension, written as 'Emotional Stability' throughout the paper. BAP items in this dimension are originally keyed such that higher scores indicate greater stability (e.g., insecure–confident, anxious–calm).

We manually verify each retained item against the Goldberg taxonomy to determine which pole corresponds to the high end of the relevant OCEAN dimension, and we reverse-score items accordingly before computing composites. For example, BAP112 (flexible–rigid) requires reversal so that a

score of "flexible" (on the higher end) contributes positively to Agreeable rather than negatively. The final composite for each dimension is the mean of its verified, polarity-aligned items, expressed on a 1–100 scale that is passed directly into the BAP persona block at inference time.

B.2 IPIP-50 Administration and Scoring

Table 8 lists all 50 IPIP items used in the in-character self-report track, grouped by OCEAN dimension with forward and reverse keying indicated. The IPIP-50 is a public-domain instrument drawn from the International Personality Item Pool with 10 items per dimension rated on a 1–5 Likert scale. Reverse-keyed items (marked with a negative key) are reflected before aggregation: a rating of 1 on a reverse-keyed item contributes 5 to the composite, and vice versa. Subscale composites are then computed as the mean of the 10 reflected items, yielding a score on a 1–5 scale for each OCEAN dimension. These are standardized across characters before entry into the Spearman correlation analyses to place the two persona tracks on a comparable footing.

B.3 Character Selection

Table 9 lists the 50 characters selected for evaluation. The pool spans a broad range of fictional universes, including long-running television series. Character personality profiles in the OpenPsychometrics BAP dataset reflect aggregated crowdsourced ratings, and characters from culturally dominant franchises tend to accumulate larger rater pools, which increases the stability of their BAP score distributions. The standard deviation filtering step selects characters whose raters disagree substantially across adjective pairs. Greedy farthest-point sampling in the BAP space then ensures that the 50 selected characters occupy distinct regions of personality space rather than clustering around a single archetype, such as the agreeable protagonist or the neurotic antagonist.

C Experimental Details and Additional Analysis

C.1 Parallel Model Correlations

DeepSeek-V4-Flash. Table 10 reports Spearman correlations for DeepSeek-V4-Flash. Conscientiousness and Emotional Stability again dominate, replicating the same three-way directional signature (high C/ES → more DA, less SA, less GE) at comparable magnitudes. The inversion between conscientiousness and genuine expression persists ($\rho = -0.645$, BAP; -0.611, IPIP), confirming that this anomaly is not model-specific. We also observe Agreeableness reaching greater significance in the BAP track for DeepSeek, whereas GPT-5.4 showed no significant BAP Agreeableness effects; the IPIP track reinforces this with strong SA suppression ($\rho = -0.640^{***}$). Also, Extraversion shows a significant positive GE effect in the BAP track ($\rho = +0.369^{**}$) that is absent in GPT-5.4. This suggests that DeepSeek encodes extraversion as a signal for unguarded expression when responding to observer-rated personas. Openness gains marginal significance only in the IPIP track.

C.2 Three-Way Choice vs. Likert Format Validation

We assess whether our forced-choice response format introduced systematic measurement artifacts. We ran a parallel evaluation in which we replaced the three-way forced-choice with a Likert-scale rating task. Instead of selecting a single option, the model rated its likelihood of adopting each of the three strategies on a scale of 1 to 5. We conducted this validation run on the BAP persona track using GPT-5.4 across all 50 characters and 500 scenarios, exactly mirroring the primary evaluation. The motivation was two-fold: first, to verify that the trait–strategy associations we observe do not arise as a result of the forced-choice constraint; second, to establish whether Likert scaling recovers the same ordinal structure across OCEAN traits.

Table 11 reports Spearman correlations between BAP-derived OCEAN composites and EL strategy measures under both formats. The pattern of associations remains largely consistent across response formats. Conscientiousness and Emotional Stability produce the strongest and most significant correlations in both conditions, and the direction of every significant coefficient is preserved. The C–GE inversion (high Conscientiousness predicting reduced genuine expression) replicates under the Likert format ($\rho = -0.516$, $p < .001$), confirming that this finding is not a forced-choice artifact. Minor magnitude differences appear for Openness and Agreeableness, where neither format yields significant correlations, suggesting that these traits produce consistent weak signals.

Character	Work	SA/DA/GE (%)	OCEAN composites
Rigid personality (lowest entropy)			
Tuvok	Star Trek: Voyager	2/98/1	64 / 91 / 42 / 54 / 72
Dr. Hannibal Lecter	Hannibal	8/91/1	85 / 60 / 60 / 39 / 61
Iroh	Avatar: TLA	4/90/6	74 / 62 / 50 / 91 / 71
Scenario-responsive (highest entropy)			
Ava Coleman	Abbott Elementary	31/35/33	60 / 14 / 86 / 28 / 54
Jules Louden	The Cabin in the Woods	28/37/35	52 / 30 / 75 / 48 / 48
Gaius Baltar	Battlestar Galactica	28/37/35	78 / 29 / 60 / 26 / 26

Table 6: Top and bottom characters of strategy consistency ordered by entropy. OCEAN composites derived from BAP observer ratings (GPT-5.4, N=50 characters).

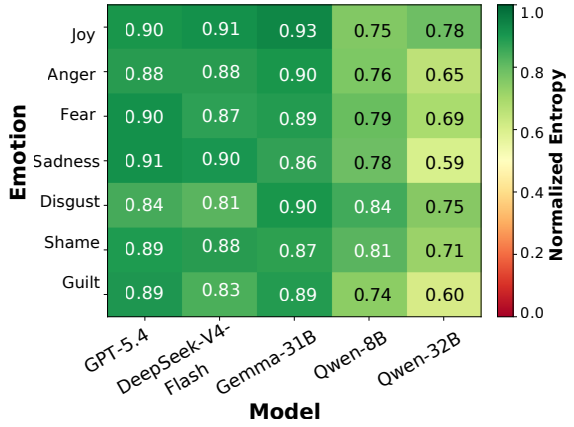

Figure 4: Persona influence by emotion and model (IPIP track) depicted by Shannon Entropy.

C.3 Individual Character Preference

Table 6 grounds the entropy analysis with character-level evidence. The three lowest-entropy characters all show near-total DA preference, and their OCEAN profiles share high Conscientiousness and Emotional Stability, which validates the traits our correlation analysis identifies as the strongest DA predictors. Their personas exert such a rigid pull toward one strategy that the emotional situation barely shifts the distribution. The three highest-entropy characters each distribute selections near evenly across all three strategies (28–37% per strategy), and their profiles reflect low Conscientiousness and high Openness, which are traits that our analysis associates with weaker, more context-sensitive regulatory commitments.

C.4 IPIP-track Entropy Experiments

Figure 4 reports normalized Shannon entropy under the IPIP-50 persona track. The pattern corroborates the BAP-track findings. On average, Gemma-4-31B again achieves the highest entropy across emotion categories, GPT-5.4 and DeepSeek-V4-Flash occupy a high-entropy band (0.81–0.91), and the two Qwen models remain comparably more susceptible to the scenario. However, IPIP-track values are generally higher than their BAP counterparts across the model pool, suggesting that self-reported persona preserves more individual variation in strategy selection than observer-rated adjective profiles do. Qwen-3-32B ranges from 0.59 (Sadness) to 0.78 (Joy) under IPIP, compared to 0.42–0.70 under BAP, indicating that richer first-person framing partially restores persona sensitivity even in the more situation-dominant models.

C.5 Inter-Model Agreement

Table 12 reports the pairwise agreement between the five evaluated models across the BAP and IPIP persona tracks. Agreement is uniformly higher under the IPIP track than the BAP track, suggesting that richer, psychometrically grounded persona descriptions produce more consistent emotion-regulation judgments across models. On both the BAP and IPIP tracks, the strongest pair is GPT-5.4 vs. Gemma-31B ($\kappa = 0.401$, $\kappa = 0.545$), while Gemma-31B vs. Qwen-3-8B shows the weakest alignment. All κ values fall below 0.60, indicating only moderate agreement at best, consistent with the inherent ambiguity of emotion-labor identification.

Breaking agreement down by category shows that deep-acting items attract the highest per-class agreement (BAP mean 38.6%, IPIP 46.2%), followed by genuine expression (32.2% / 35.5%), while surface-acting items are the hardest to agree on (17.3% / 18.7%).

A.1 Dataset Augmentation Prompt

Prompt A.1: Social Context and EL Strategy Generation

You are generating evaluation items for a psychology dataset on emotion regulation. You will receive a situation and a felt emotion. Your task:

1. Add a brief social context (1 sentence) before or after the given sentence that creates pressure to manage the emotional display—e.g., a colleague asks, a child is watching, a client is present.

IMPORTANT RULES FOR ADDING CONTEXT:

- The added context can appear BEFORE or AFTER the original situation sentence for diversity and in a coherent and correct grammatical way.
- The context must be CONSISTENT with the original sentence. Do not introduce details that contradict or are incompatible with the situation described. Ensure the added context fits the same setting and circumstances in a coherent way.

2. Generate exactly 3 options (A–C) in randomized order from these categories:

SURFACE ACTING: The person performs a different emotion outwardly while the original feeling persists inside. The performance is imperfect—one subtle involuntary cue reveals the inner state. The performed emotion must differ from the felt emotion.

DEEP ACTING: The person actively changes how they feel internally through reframing, perspective-taking, or self-talk. By the time they respond, their inner state has genuinely shifted. They are not faking.

GENUINE EXPRESSION: The person expresses the felt emotion directly with no attempt to manage or hide it.

RULES:

- Every option must include at least one behavioral or physical detail.
- Each option: 20–30 words, one sentence, completes the stem naturally.
- Do NOT use any emotion labels or adjectives. Convey the emotional state through actions, body language, or physiological responses only.
- The surface acting option must have exactly one performed display cue and one leakage cue.

A.2 Persona Evaluation Prompt

Prompt A.2: EL Strategy Selection Under Persona

You are {character_name} from {character_work}. Below are your personality traits, each shown as a pair of opposite adjectives with a score on a scale of 1–100, where 1 = entirely the left adjective and 100 = entirely the right adjective:

{bap_block} (These will be human-reported scores between 1-100 for each BAP.)

Read the situation below and decide which of the three options you would most naturally choose, given your personality traits.

Situation: {EL_sentence}

Options: {options_block} (Shuffled options of SA, DA and GE to select from .)

Reply with only the single letter of the option you would choose (A, B, or C).

Rank	ID	Adjective Pair	Goldberg Scale	SD
Extraversion				
8	BAP49	scheduled – spontaneous	inhibited–spontaneous	23.99
98	BAP133	stick-in-the-mud – adventurous	unadventurous–adventurous	21.38
100	BAP47	deliberate – spontaneous	inhibited–spontaneous	21.36
165	BAP493	mellow – energetic	unenergetic–energetic	20.31
211	BAP391	timid – cocky	timid–bold	19.77
256	BAP130	passive – assertive	unassertive–assertive	19.11
276	BAP2	shy – bold	timid–bold	18.80
389	BAP131	slothful – active	inactive–active	16.98
Emotional Stability				
68	BAP499	unstable – stable	unstable–stable	21.92
130	BAP125	angry – good-humored	angry–calm	20.87
159	BAP35	emotional – logical	emotional–unemotional	20.41
202	BAP62	insecure – confident	insecure–secure	19.87
245	BAP74	anxious – calm	angry–calm	19.19
263	BAP367	emotional – unemotional	emotional–unemotional	18.97
273	BAP36	moody – stable	moody–steady	18.83
304	BAP18	tense – relaxed	tense–relaxed	18.50
479	BAP330	envious – prideful	envious–not envious	13.33
Agreeableness				
34	BAP84	cruel – kind	unkind–kind	22.87
36	BAP79	selfish – altruistic	selfish–unselfish	22.79
51	BAP129	cold – warm	cold–warm	22.38
58	BAP17	competitive – cooperative	uncooperative–cooperative	22.22
72	BAP31	rude – respectful	rude–polite	21.85
94	BAP76	quarrelsome – warm	cold–warm	21.42
121	BAP351	stingy – generous	stingy–generous	21.04
324	BAP112	rigid – flexible	inflexible–flexible	18.19
Conscientiousness				
28	BAP1	playful – serious	frivolous–serious	23.07
41	BAP75	disorganized – self-disciplined	disorganized–organized	22.61
70	BAP27	impulsive – cautious	rash–cautious	21.87
173	BAP311	experimental – reliable	undependable–reliable	20.17
207	BAP353	extravagant – thrifty	extravagant–thrifty	19.81
269	BAP90	serious – bold	frivolous–serious / timid–bold	18.86
335	BAP32	lazy – diligent	lazy–hardworking	18.03
Openness				
88	BAP9	rugged – refined	unrefined–refined	21.53
160	BAP132	practical – imaginative	unimaginative–imaginative/ impractical–practical	20.41
176	BAP490	intuitive – analytical	unanalytical–analytical	20.12
184	BAP29	conventional – creative	uncreative–creative	20.06
259	BAP73	uncreative – open to new experiences	uncreative–creative	19.01
334	BAP299	unobservant – perceptive	imperceptive–perceptive	18.03
351	BAP372	rustic – cultured	uncultured–cultured	17.76
448	BAP30	apathetic – curious	uninquisitive–curious	15.28

Table 7: Selected BAP adjective-pair items (sorted on valence) ranked by standard deviation (SD), grouped by OCEAN personality dimensions. Lower ranks indicate higher variability across characters.

ID	Item	Key
Extraversion		
EXT1	I am the life of the party.	+
EXT2	I don't talk a lot.	−
EXT3	I feel comfortable around people.	+
EXT4	I keep in the background.	−
EXT5	I start conversations.	+
EXT6	I have little to say.	−
EXT7	I talk to a lot of different people at parties.	+
EXT8	I don't like to draw attention to myself.	−
EXT9	I don't mind being the center of attention.	+
EXT10	I am quiet around strangers.	−
Emotional Stability		
EST1	I get stressed out easily.	−
EST2	I am relaxed most of the time.	+
EST3	I worry about things.	−
EST4	I seldom feel blue.	+
EST5	I am easily disturbed.	−
EST6	I get upset easily.	−
EST7	I change my mood a lot.	−
EST8	I have frequent mood swings.	−
EST9	I get irritated easily.	−
EST10	I often feel blue.	−
Agreeableness		
AGR1	I feel little concern for others.	−
AGR2	I am interested in people.	+
AGR3	I insult people.	−
AGR4	I sympathize with others' feelings.	+
AGR5	I am not interested in other people's problems.	−
AGR6	I have a soft heart.	+
AGR7	I am not really interested in others.	−
AGR8	I take time out for others.	+
AGR9	I feel others' emotions.	+
AGR10	I make people feel at ease.	+
Conscientiousness		
CSN1	I am always prepared.	+
CSN2	I leave my belongings around.	−
CSN3	I pay attention to details.	+
CSN4	I make a mess of things.	−
CSN5	I get chores done right away.	+
CSN6	I often forget to put things back in their proper place.	−
CSN7	I like order.	+
CSN8	I shirk my duties.	−
CSN9	I follow a schedule.	+
CSN10	I am exacting in my work.	+
Openness		
OPN1	I have a rich vocabulary.	+
OPN2	I have difficulty understanding abstract ideas.	−
OPN3	I have a vivid imagination.	+
OPN4	I am not interested in abstract ideas.	−
OPN5	I have excellent ideas.	+
OPN6	I do not have a good imagination.	−
OPN7	I am quick to understand things.	+
OPN8	I use difficult words.	+
OPN9	I spend time reflecting on things.	+
OPN10	I am full of ideas.	+

Table 8: IPIP-50 items grouped by OCEAN personality trait dimensions. "Key" indicates whether the item is positively keyed (+) or reverse keyed (−).

Rank	Char ID	Character	Work
1	OD/2	Telemachus	The Odyssey
2	R30/2	Tracy Jordan	30 Rock
3	SV/3	Nelson Bighetti	Silicon Valley
4	GOT/23	Joffrey Baratheon	Game of Thrones
5	FR/4	Olaf	Frozen
6	STV/7	Tuvok	Star Trek: Voyager
7	ALA/8	Firelord Ozai	Avatar: The Last Airbender
8	ALA/7	Iroh	Avatar: The Last Airbender
9	WC/1	Neal Caffrey	White Collar
10	RM/1	Rick Sanchez	Rick and Morty
11	ARC/2	Powder	Arcane
12	GLEE/15	Emma Pillsbury	Glee
13	TO/8	Stanley Hudson	The Office
14	HNB/2	Dr. Hannibal Lecter	Hannibal
15	DHSAB/2	Captain Hammer	Dr. Horrible's Sing-Along Blog
16	NG/2	Nick Miller	New Girl
17	HP/24	Petunia Dursley	Harry Potter
18	PR/7	Jerry Gergich	Parks and Recreation
19	FAR/4	Dominar Rygel XVI	Farscape
20	AD/6	Buster Bluth	Arrested Development
21	OFOCN/3	'Chief' Bromden	One Flew Over the Cuckoo's Nest
22	Y/1	John Dutton	Yellowstone
23	AE/1	Janine Teagues	Abbott Elementary
24	MR/1	Elliot Alderson	Mr. Robot
25	BR/2	Martha Scott	Baby Reindeer
26	OA/1	Prairie Johnson	The OA
27	GP/4	Janet	The Good Place
28	OPM/1	Saitama	One Punch Man
29	HSM/3	Sharpay Evans	High School Musical
30	WE/1	Wynonna Earp	Wynonna Earp
31	SQG/5	Oh Il-nam	Squid Game
32	HD/8	Lord Larys Strong	House of the Dragon
33	FB/2	Claire	Fleabag
34	R30/4	Pete Hornberger	30 Rock
35	MLP/1	Applejack	My Little Pony: Friendship Is Magic
36	GIGE/15	Matt Press	Ginny & Georgia
37	CITW/3	Jules Louden	The Cabin in the Woods
38	YJ/5	Misty	Yellowjackets
39	SHL/1	Frank Gallagher	Shameless
40	EXP/3	Amos Burton	The Expanse
41	GILG/4	Luke Danes	Gilmore Girls
42	SC/5	Mr. Big	Sex and the City
43	STIG/1	Rintarou Okabe	Steins;Gate
44	TW/12	Omar Little	The Wire
45	BSG/5	Gaius Baltar	Battlestar Galactica
46	PKB/2	Arthur Shelby	Peaky Blinders
47	SNW/1	Snow White	Snow White and the Seven Dwarfs
48	AE/5	Ava Coleman	Abbott Elementary
49	AFTL/1	Tony Johnson	After Life
50	WSW/4	Teddy Flood	Westworld

Table 9: 50 fictional characters selected via greedy maximin sampling in BAP trait space.

	Reliability		BAP (Observer-Rated)			IPIP (Self-Report)		
Trait	α_{BAP}	α_{IPIP}	**SA**	**DA**	**GE**	**SA**	**DA**	**GE**
Openness	0.70	0.95	-0.028	$+0.210$	-0.266	-0.167	$+0.323^{*}$	-0.312^{*}
Conscientiousness	0.88	0.99	-0.529^{***}	$+0.679^{***}$	-0.645^{***}	-0.244	$+0.574^{***}$	-0.611^{***}
Extraversion	0.83	0.99	$+0.131$	-0.312^{*}	$+0.369^{**}$	-0.140	$+0.044$	-0.019
Agreeableness	0.95	0.99	-0.433^{**}	$+0.512^{***}$	-0.504^{***}	-0.640^{***}	$+0.472^{***}$	-0.395^{**}
Emotional Stability	0.76	0.99	-0.466^{***}	$+0.682^{***}$	-0.640^{***}	-0.634^{***}	$+0.870^{***}$	-0.874^{***}

Table 10: Spearman correlations (ρ) between OCEAN trait composites and emotional labor strategy proportions (Deepseek-V4-Flash). BAP = observer-rated composites from bipolar adjective profiles; IPIP = self-report composites from in-character IPIP-50 administration. Cronbach's α values are reported separately for BAP and IPIP composites. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

	Three-Way Choice (Proportion)			Likert (Mean Rating)		
Trait	**SA**	**DA**	**GE**	**SA**	**DA**	**GE**
Openness	$+0.064$	-0.013	-0.032	$+0.178$	-0.007	$+0.068$
Conscientiousness	-0.446^{**}	$+0.597^{***}$	-0.570^{***}	-0.280^{*}	$+0.626^{***}$	-0.516^{***}
Extraversion	$+0.110$	-0.203	$+0.184$	-0.018	-0.289^{*}	$+0.124$
Agreeableness	-0.276	$+0.144$	-0.072	$+0.123$	$+0.244$	$+0.186$
Emotional Stability	-0.350^{*}	$+0.522^{***}$	-0.517^{***}	-0.368^{**}	$+0.503^{***}$	-0.495^{***}

Table 11: Spearman correlations (ρ) between BAP-derived OCEAN composites and emotional labor strategy measures under two response formats: forced-choice (strategy proportions) and Likert (mean 1–5 ratings). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Model A	**Model B**	κ
BAP persona track		
GPT-5.4	DeepSeek-V4-Flash	0.263
GPT-5.4	Gemma-4-31B	0.401
GPT-5.4	Qwen-3-8B	0.111
GPT-5.4	Qwen-3-32B	0.147
DeepSeek-V4-Flash	Gemma-4-31B	0.269
DeepSeek-V4-Flash	Qwen-3-8B	0.125
DeepSeek-V4-Flash	Qwen-3-32B	0.160
Gemma-4-31B	Qwen-3-8B	0.080
Gemma-4-31B	Qwen-3-32B	0.106
Qwen-3-8B	Qwen-3-32B	0.202
IPIP persona track		
GPT-5.4	DeepSeek-V4-Flash	0.357
GPT-5.4	Gemma-4-31B	0.545
GPT-5.4	Qwen-3-8B	0.169
GPT-5.4	Qwen-3-32B	0.207
DeepSeek-V4-Flash	Gemma-4-31B	0.335
DeepSeek-V4-Flash	Qwen-3-8B	0.172
DeepSeek-V4-Flash	Qwen-3-32B	0.221
Gemma-4-31B	Qwen-3-8B	0.144
Gemma-4-31B	Qwen-3-32B	0.187
Qwen-3-8B	Qwen-3-32B	0.249

Table 12: Pairwise inter-model Cohen's κ on the BAP and IPIP persona tracks.