MOTIVATIONALLY GROUNDED ROLE-PLAYING AGENTS: STRESS-DRIVEN PSYCHODYNAMIC MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Role-playing agents (RPAs) powered by large language models have become central to interactive character simulation, yet existing approaches often rely on surface-level persona descriptions and stylistic imitation, lacking a principled account of the internal psychological dynamics that drive role-consistent behavior. This limitation leads to shallow personality modeling and character drift in extended interactions. We propose a motivationally grounded framework that explicitly models a character's internal motivational structure. Grounded in psychodynamic, attachment, moral, existential, and interpersonal theories, we construct a structured character psychology that includes a taxonomy of core anxieties, a corresponding set of compensatory desires, three personality organization types, and a mapping from each personality organization and stress level to the character's available defense mechanisms. These structures are embedded in a dual-process generation architecture. On SocialBench, our method outperforms models of comparable scale. A controlled probe study on a purpose-built stress dataset further confirms that the framework's internal mechanisms operate as intended, with defense tier governing modulation intensity and defense selection producing distinct surface forms from the same impulse. These results suggest that explicitly modeling character psychology offers a scalable and effective pathway toward deeper role-playing fidelity.

1 INTRODUCTION

Role-playing conversational agents, i.e., chatbots that embody a specific character with a stable persona, are central to a growing set of applications: open-domain companions (e.g., Replika AI), immersive non-player characters (Park et al., 2023), emotionally supportive listeners (Zhu et al., 2026), and personalized tutors (Rai et al., 2024). Such agents require more than factual knowledge. A believable character must display stable traits, situational coherence, and, crucially, the ability to react authentically under stress. Recent LLMs have improved markedly at role consistency and knowledge recall, yet the resulting behaviour still reads as mechanical. Agents produce polite, on-character, contextually appropriate responses, but seldom exhibit the kinds of behaviour that distinguish a *believable* character from a *competent* one: moments of unprovoked reticence, defensiveness at the wrong moment, delight that breaks into uneasiness, or the slow accumulation of hurt across turns. These are emergent signs of an internal life. We argue that the gap between competent and believable role-playing stems from a missing ingredient in today's frameworks, namely an *internal psychological state* whose dynamics *drive* behaviour rather than merely *describe* it.

Among the psychological mechanisms that could supply such an internal state, one is particularly relevant to role-playing. **The more a person lacks something, the more urgently they want to prove something.** This intuition is captured by *compensation psychology*, which holds that when a certain deficiency exists in one's inner self, the individual will unconsciously use words or actions to make up for it. A person lacking security, for instance, often behaves exceptionally assertively in interpersonal interactions, attempting to appear stronger in this way. It is this *dynamic driven by deficiency* that distinguishes a *convincing* role from a merely *competent* one. In role-playing, this mechanism is precisely what separates a character who merely follows a persona description from one who appears to have an inner life. Current LLM frameworks, however, largely neglect it.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Although recent progress has enhanced role consistency, personality traits are still treated as static attributes rather than dynamic responses to internal pressure. As a result, an agent can generate polite, character-fit responses, but it cannot exhibit signs of emergent internal psychological activity, such as unexplained silences or defensive reactions triggered by hidden insecurity. What remains missing, then, is a principled account of how such deficiency-driven dynamics can be represented and operationalized within a generative architecture.

To address this gap, we present a psychology-grounded framework modeling internal motivational dynamics as the generative source of behavior. We operationalize this system through three components. **First**, drawing on psychodynamic and interpersonal theories, we construct a taxonomy of core anxieties (competence, relational, moral, existential, control) and corresponding compensatory desires. **Second**, based on Kernberg's object-relations framework (Kernberg, 1975) and Vaillant's defense hierarchy (Vaillant, 1998), we define three personality organization types (neurotic, borderline, psychotic), mapping stress levels to available defense mechanisms, capturing how structural capacity shapes behavioral responses. **Third**, we embed these structures into a dual-process architecture: under high stress, fast System 1 thinking emits direct responses; under normal stress, slow System 2 thinking applies defenses to refine outputs. This simulates internal-life signals like defensive reticence or delayed emotional responses.

We evaluate our framework on SocialBench (Chen et al., 2024), where our method outperforms baselines on sociality metrics. Beyond this main evaluation, we conduct a controlled probe study on a purpose-built stress dataset, in which each character is subjected to a graded sequence of interpersonal pressure. The probe study is not designed to measure aggregate performance, but to verify that the framework's internal mechanisms operate as intended: that defense tier controls the intensity of surface modulation, and that the same raw impulse yields different surface forms under different defenses. Together, these analyses confirm that the performance gain on SocialBench originates from the structural design of the framework rather than from surface-level stylistic imitation.

2 Related Work

2.1 Role-Playing Language Agents: From Persona to Psychological Depth

Research on role-playing language agents has evolved through a series of conceptual shifts. Initially, the field focused on *demographic persona*, conditioning LLMs on social labels such as "a data scientist." Kong et al. (2024) showed that such role-play prompting improves zero-shot reasoning, but these approaches rely on statistical stereotypes in pretraining data and treat persona as a static filter.

Subsequently, attention shifted toward *character persona*, equipping agents with rich background materials. Representative works include CharacterChat (Tu, 2023), SocialBench (Chen et al., 2024), CharacterEval (Tu et al., 2024), RoleLLM (Wang et al., 2024a), and CharacterGLM (Zhou et al., 2024). These systems improve role consistency and knowledge recall, yet they model personality as static adjectives rather than a system responding to internal pressure, producing polite and on-character responses that seldom exhibit an inner life.

More recently, the focus has turned to *individualized persona*, addressing temporal depth and identity integration. Park et al. (2023) introduced memory streams and reflection that let characters evolve; Park et al. (2025) updated persona descriptions chapter by chapter; Lee et al. (2025) integrated social identity, personal identity, and life background, though the construct remains static; and Yang et al. (2026) embedded socialization theory into a game world. Across these works, integration remains at the level of *juxtaposition*, where multiple dimensions are added side by side without an underlying driving force to explain their coordinated evolution. Our work addresses this gap.

2.2 Psychological Dimensions and Psychodynamic Foundations in Role-Playing Agents

Existing RPLA research has explored several psychological dimensions, though often in isolation. In the realm of **personality**, the Big Five model dominates the landscape. For instance, Jiang et al. (2024) verified that LLMs can express designated traits, while Dan et al. (2025) achieved topic-

adaptive expression via mixture-of-experts. Further studies have examined consistency under quantization (Wang et al., 2024b) and revealed that high agreeableness tends to increase sycophancy (Shah et al., 2026). Despite these advances, deeper psychodynamic and humanistic theories receive limited attention. Regarding **cognition**, research indicates that persona prompting affects theory-of-mind reasoning (Tan, 2024). Recent works have also built benchmarks for inner thought reasoning (Xu et al., 2025a), proposed metacognition-driven hierarchical reasoning (Qin et al., 2025), and implemented a "cognize-then-respond" dual-system paradigm (Liu et al., 2025). In terms of **emotion**, studies have simulated emotion transitions using Markov chains and circumplex models (Wei et al., 2025), while others have evaluated emotional support capabilities and coping strategies (Zhao et al., 2024; Troiano et al., 2024). Finally, within the dimension of **motivation**, Xu et al. (2025b) stands out for systematically injecting self-determination theory into retrieval-augmented generation. However, other frameworks such as the hierarchy of needs and ERG theory remain underrepresented. Two primary gaps emerge from this landscape. First, these dimensions are typically studied in isolation, lacking a cross-dimensional framework. Second, psychodynamic constructs—specifically core anxiety, compensatory desire, and defense mechanisms—are largely absent from RPLA research, despite being well established in clinical psychology (Freud, 1936; Vaillant, 1998; Kernberg, 1975). Our work addresses both gaps by integrating these psychodynamic constructs into an executable generation architecture.

3 Character Psychology

To give the agent a structured vocabulary for reasoning about internal state, we define a four-layer taxonomy: core anxiety, compensatory desire, personality organization, and defense mechanism. Theoretical justification for each layer is provided in Appendix A through D.

Core Anxiety. Each character is characterized by five categories of self-threat—competence, relational, moral, existential, and control anxiety—each comprising four specific anxiety seeds listed in Table 7. For example, repeated public criticism from a supervisor activates the competence category through the seeds of *incompetence* and *layperson status*, whereas abrupt relationship termination activates the relational category through *abandonment* and *betrayal*.

Compensatory Desire. Each anxiety category is associated with a subset of Murray's psychogenic needs that specify how the character attempts to repair the damaged dimension (Murray & McAdams, 2007). The mapping is given in Table 8. A junior employee threatened with professional incompetence gravitates toward achievement and recognition, manifesting as the impulse to prove himself or seek acknowledgment, whereas a character threatened with abandonment gravitates toward affiliation and succorance.

Personality Organization. We adopt Kernberg's distinction among three structural levels—neurotic, borderline, and psychotic organization (Kernberg, 1975). This level determines which defense mechanisms the character can recruit and how much stability it can maintain under escalating stress. A neurotic-organized character accesses mature defenses under low stress and regresses to immature ones under sustained pressure, but does not collapse into psychotic-level defenses; a borderline-organized character, by contrast, oscillates between idealization and devaluation at baseline and may briefly lose reality testing under acute stress.

Defense Mechanism. We adopt Vaillant's four-level hierarchy of defenses, ordered from primitive to mature (Vaillant, 1998). Rather than treating defenses as fixed traits, we map each personality organization to a subset of defenses at three stress tiers, as shown in Table 9 and Table 10. For a neurotic-organized character, the available repertoire shifts from suppression and humor at baseline, to reaction formation and rationalization at moderate stress, and to projection and somatization at extreme stress.

4 Methodology

Building on the character psychology formalized in Section 3, our framework **PsyDrive** simulates the turn-by-turn evolution of a character's internal state and its impact on verbal expression. Each

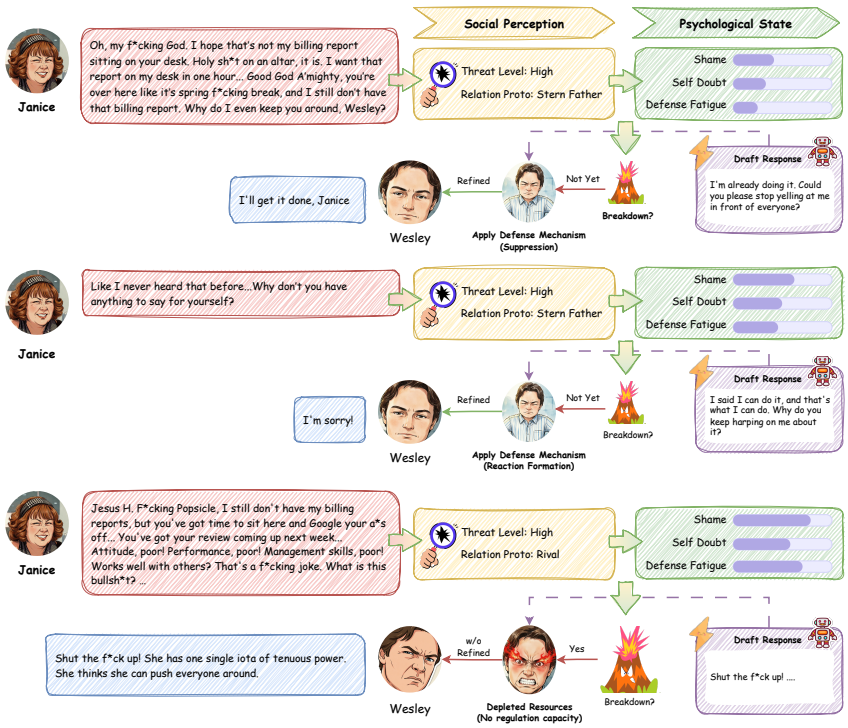

Figure 1: Overview of the proposed psychology-grounded role-playing framework (**PsyDrive**). The example dialogue is taken from the film *Wanted*, where Janice repeatedly attacks Wesley's competence and professional worth.

turn is processed by three sequential modules: Social Perception, Psychological State, and Response Generation. The overall architecture is illustrated in Figure 1.

4.1 Social Perception

The Social Perception module takes as input the role profile, dialogue history, and the current speaker message, and produces two outputs: a threat level $t \in \{0, 1, 2\}$ and a relational prototype identifier r. It is implemented by prompting a large language model (LLM) to act as a character psychology analyst.

Threat Level. The threat level captures the extent to which the current message is perceived, from the character's perspective, as an attack on its core anxiety dimensions. A level of 0 denotes routine interaction; 1 signifies an indirect provocation; and 2 indicates a direct attack on a core anxiety seed. This judgment is anchored to the character's specific anxiety profile rather than to surface-level semantic hostility. For instance, as illustrated in Figure 1, when Janice disparages or mocks Wesley's professional competence, the resulting threat level is high, as this directly attacks his core anxiety of competence.

Relational Prototype. We characterize the relationship between the character and the speaker using four prototypes derived from the Interpersonal Circumplex framework (Leary, 2004; Kiesler, 1983): the Stern Father, the Rival, the Ally, and the Dependent. The prototype is inferred from the overall relational dynamic accumulated across the dialogue history, rather than from the current message in isolation. The theoretical foundation for this construct is detailed in Appendix E.

4.2 Psychological State

Stress Dynamics. The module maintains three stress variables that evolve across turns: shame, self-doubt, and defense fatigue. They capture the sequential stages of a threat response, namely

affective arousal, cognitive reappraisal, and resource depletion. The justification for this choice and for the functional form of the update equations is given in Appendix F.

Given a threat level t from Section 4.1, we map it to a primary threat value $p(t)$. Let c denote the character's coping capacity, which is determined by its personality organization. The state variables are then updated as

$$s_t = \alpha \cdot p(t) \cdot (1 - c) + (1 - \alpha) \cdot s_{t-1}, \tag{1}$$

$$d_t = \alpha \cdot s_t + (1 - \alpha) \cdot d_{t-1}, \tag{2}$$

$$f_t = \begin{cases} \beta \cdot (1 - c) + (1 - \beta) \cdot f_{t-1}, & t \geq 1, \\ \max(0, f_{t-1} - \delta_t), & t = 0, \end{cases} \tag{3}$$

where s_t, d_t, and f_t denote shame, self-doubt, and defense fatigue at turn t, α and β are smoothing coefficients, and δ_t is the decay applied to defense fatigue during low-threat turns:

$$\delta_t = \delta_0 \cdot (1 + n_t \cdot \delta_{\text{step}}), \tag{4}$$

with n_t the number of consecutive low-threat turns up to turn t, δ_0 the base decay rate, and δ_{step} the additional acceleration per consecutive low-threat turn. The accelerating form reflects the intuition that a character recovers faster the longer it has been free of threat, rather than at a constant rate.

Phase Breakdown. When shame and defense fatigue simultaneously exceed their respective thresholds, τ_s and τ_f, the module flags a phase breakdown. These thresholds are modulated by self-doubt, such that a character with elevated self-doubt reaches breakdown more readily. Once flagged, the downstream generation module bypasses defensive modulation and emits the draft directly. As illustrated in Figure 1, Wesley initially manages to suppress his reactions, as his psychological state remains below the thresholds. However, the accumulated shame and fatigue from the third attack eventually cross these limits simultaneously, triggering a breakdown and allowing the raw impulse to reach the output without defensive filtering.

Defense Tier. Independently of phase breakdown, the module selects one of three defense tiers: baseline, moderate, or extreme. Let $m(d_t) = 1 - \gamma \cdot d_t$ be a modulation factor driven by self-doubt, and let $\tau_s \cdot m(d_t)$ and $\tau_f \cdot m(d_t)$ be the effective thresholds. Writing $\tau_{\min} = \min(\tau_s \cdot m(d_t), \tau_f \cdot m(d_t))$, the tier is defined as

$$\text{tier}_t = \begin{cases} \text{extreme}, & s_t \geq \tau_s \cdot m(d_t) \text{ or } f_t \geq \tau_f \cdot m(d_t), \\ \text{moderate}, & s_t \geq \kappa \cdot \tau_{\min} \text{ or } f_t \geq \kappa \cdot \tau_{\min}, \\ \text{baseline}, & \text{otherwise}, \end{cases} \tag{5}$$

where $\kappa \in (0, 1)$ sets the boundary for the moderate tier. As shown in Figure 1, as the conversation progresses, Wesley's psychological state parameters deteriorate continuously, leading to different defense tiers in each round and thus different defensive measures that can be chosen (suppression $\rightarrow$ reaction formation). The design rationale for this tiered formulation and its relationship to phase breakdown is given in Appendix F.

4.3 Response Generation

The Response Generation module produces the reply through two stages. **Stage 1 (Draft)** generates a raw impulse. **Stage 2 (Rewrite)** applies a defense mechanism from the tier-specific repertoire to shape the draft into the final response. When phase breakdown is flagged, Stage 2 is skipped, and the draft is emitted directly. The rationale for this two-stage design is given in Appendix G.

5 EXPERIMENTS

Dataset. We evaluate PsyDrive on the SocialBench benchmark(Chen et al., 2024). This benchmark comprises 500 characters sourced from novels, films, and online platforms, along with over 6,000 queries and 30,800 multi-turn role-playing dialogue instances. It evaluates models across multiple dimensions: Self-Awareness (SA), covering style and knowledge; Emotion Perception (EP), assessing situational and emotional understanding; Contextual Memory (CM), probing both short-term and long-term memory; and group-level social intelligence, measuring performance in positive, neutral, and negative interactions. We exclude EP from our experiments, as its evaluation samples are not accompanied by character profiles.

Baselines. We compare PsyDrive against three groups of baselines. The first group includes general-purpose large language models: Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct, and OpenAI's proprietary GPT-4o-mini. All models in this group are evaluated both with and without zero-shot Chain-of-Thought (CoT) prompting. The second group comprises reasoning-enhanced models and reasoning frameworks: DeepSeek-R1-7B and DeepSeek-R1-32B, which were obtained through knowledge distillation from their corresponding Qwen2.5 counterparts, ensuring a comparison of different post-training methods on the same original models; additionally, we include Tree-of-Thought (ToT) (Yao et al., 2023) and a Self-Correction (Madaan et al., 2023) baseline, both implemented on Qwen2.5-7B, to ensure a fair comparison against more sophisticated reasoning frameworks. The third group includes specialized role-playing models: CharacterGLM (Zhou et al., 2024), CoSER(Wang et al., 2025), ChatHaruhi(Li et al., 2023), CharacterLLM(Shao et al., 2023), and RoleLLM-rep(Wang et al., 2024a). Given the full alignment of our experimental setup with that of Qin et al. (2025), we directly adopt their reported baseline results.

Table 1: Performance comparison on SocialBench. All values represent accuracy rates (%)

Models	Individual Level				Group Level			Avg
	SA Style	SA Know.	CM Short	CM Long	Pos.	Neu.	Neg.	
Large-Scale Models								
GPT-4o-mini	84.30	92.05	**81.55**	80.86	**92.83**	**84.10**	**82.64**	**85.48**
GPT-4o-mini CoT	79.62	89.54	81.53	**80.98**	89.86	77.76	71.59	81.55
Qwen2.5-32B	**84.84**	**95.76**	73.16	38.75	88.57	79.14	63.04	74.75
Qwen2.5-32B CoT	82.50	94.62	69.73	35.31	87.86	77.04	60.40	72.49
DeepSeek-R1-32B	79.40	94.42	66.42	33.93	87.69	75.38	67.93	72.17
Models with Comparable Parameter Scales								
Qwen2.5-7B	70.76	82.10	66.12	34.64	76.19	67.56	58.49	65.12
Qwen2.5-7B CoT	71.11	82.14	62.49	31.69	78.07	67.04	56.79	64.19
ToT	55.87	78.13	47.96	37.81	57.35	51.89	33.92	51.85
Self-Correction	63.26	79.20	72.46	42.95	69.06	55.97	48.55	61.64
DeepSeek-R1-7B	59.58	76.03	37.12	21.58	70.38	57.73	47.03	52.78
PsyDrive∼(Ours)	**71.68**	**89.16**	**74.99**	**77.73**	**79.49**	**73.72**	**66.12**	**76.13**

Table 2: Performance comparison with specialized role-playing models on SocialBench. All values represent accuracy rates (%).

Models	Individual Level				Group Level			Avg
	SA Style	SA Know.	CM Short	CM Long	Pos.	Neu.	Neg.	
CoSER	43.56	69.53	37.27	28.52	52.74	36.78	24.13	41.79
Chatharuhi	41.37	57.13	18.66	16.19	45.45	30.92	22.73	33.21
CharacterLLM	14.63	15.16	6.76	6.25	11.54	9.38	9.52	10.46
CharacterGLM	16.98	30.20	11.45	1.36	53.85	19.05	13.33	20.89
RoleLLM-rep	24.12	24.42	13.28	9.71	51.37	25.52	11.81	22.89
R-CHAR	69.97	86.36	62.51	68.59	**83.90**	63.25	53.48	69.72
PsyDrive (Ours)	**71.68**	**89.16**	**74.99**	**77.73**	79.49	**73.72**	**66.12**	**76.13**

Table 3: Ablation study of PsyDrive components on SocialBench. **SP**: Social Perception; **PsyS**: Psychological State.

Models	Individual Level				Group Level			Avg
	SA Style	SA Know.	CM Short	CM Long	Pos.	Neu.	Neg.	
PsyDrive (Ours)	71.68	**89.16**	**74.99**	**77.73**	79.49	73.72	66.12	**76.13**
-w/o PsyS	71.00	88.80	70.49	72.77	78.50	73.90	63.86	74.19
-w/o SP&PsyS	69.80	87.01	71.66	74.08	74.06	73.20	68.81	74.09
-w/o rewrite	69.05	86.50	70.85	73.83	77.65	**77.62**	74.75	75.75
-w/o core_anxiety	**72.15**	88.44	69.12	73.30	**79.86**	72.38	62.71	73.99
-w/o desires	70.18	88.16	69.92	73.40	78.33	71.55	65.35	73.84
-w/o personality_org.	71.41	88.07	68.40	71.86	79.09	75.74	69.68	74.89
-w/o defenses&rewrite	68.93	86.50	70.40	73.20	74.92	76.93	**77.47**	75.48

Results. Tables 1 and 2 reports the performance of PsyDrive alongside all baselines on SocialBench. We first compare PsyDrive against models of the same scale. PsyDrive achieves an average accuracy of 76.13%, substantially outperforming its base model Qwen2.5-7B (65.12%) and slightly exceeding the larger Qwen2.5-32B (74.75%). The improvement is particularly pronounced in long-context comprehension (CM Long), where PsyDrive reaches 77.73%, more than doubling the 34.64% achieved by Qwen2.5-7B and exceeding the 68.59% of R-CHAR, the strongest reasoning-based role-playing framework to date (Qin et al., 2025). PsyDrive also leads on group-level social preferences, with 79.49% on positive interactions, 73.72% on neutral interactions, and 66.12% on negative interactions, outperforming all specialized role-playing models by a wide margin. Compared with the specialized role-playing models, the advantage of PsyDrive is even more evident. CharacterGLM achieves only 20.89% average accuracy, CoSER 41.79%, and RoleLLM-rep 22.89%, all far below PsyDrive. This gap suggests that existing role-playing systems, despite being fine-tuned on large-scale character dialogue corpora, lack the structured psychological modeling that PsyDrive provides through its core anxiety, compensatory desire, and defense mechanism framework. We note that proprietary and larger-scale models such as GPT-4o-mini (85.48%) still achieve higher overall scores; PsyDrive's contribution is to close much of this gap at the 7B scale without additional fine-tuning.

Ablation study. Table 3 reports the ablation results. The full model achieves the highest average score (76.13), with every ablated variant falling below it, confirming that the components are complementary rather than redundant. Four patterns carry theoretical significance.

Defense mechanisms modulate negative expression. Removing the rewrite stage (-w/o rewrite) or the defense repertoire together with rewrite (-w/o defenses&rewrite) sharply raises negative social preference (66.12 to 74.75 and 77.47). This aligns with the psychodynamic function of defenses, which convert raw impulse into socially tolerable form (Vaillant, 1998; Freud, 1936). Without them, the character expresses hostility directly, which raises negative-preference accuracy at the cost of style consistency (69.05 and 68.93) and overall coherence (75.75 and 75.48). PsyDrive deliberately favors the latter.

Compensatory desires drive behavioral direction. Removing compensatory desires (-w/o desires) produces the largest average drop (76.13 to 73.84). Theoretically, compensatory desires are the directional vector translating diffuse anxiety into executable behavior (Murray & McAdams, 2007; Lazarus, 1991). Without them, the character lacks a target for repairing the damaged dimension, resulting in behavioral drift. This confirms that compensatory desires, not core anxiety, carry the primary burden of behavioral direction.

Core anxiety anchors memory and social stance, not surface style. Removing core anxiety (-w/o core_anxiety) slightly improves SA Style (71.68 to 72.15) and positive preference (79.49 to 79.86), yet sharply drops CM Short (74.99 to 69.12) and lowers negative preference (66.12 to 62.71). Core anxiety thus shapes how the character remembers and positions itself socially rather than how it sounds. Without a stable anxiety configuration, the character has no reason to track threatening content across turns or differentiate friendly from hostile interlocutors. Surface style remains intact because it is carried by other components.

Social perception and personality organization provide complementary structure. Because psychological state depends on the threat signal from social perception, we report their combined ablation (-w/o SP&PsyS), which drops the average to 74.09 with the largest losses in positive preference (79.49 to 74.06) and SA Style (71.68 to 69.80). Removing personality organization (-w/o personality_org.) produces consistent drops in CM Short (74.99 to 68.40) and CM Long (77.73 to 71.86), consistent with its theoretical role as a structural capacity that stabilizes behavior across turns (Kernberg, 1975).

6 DISCUSSION

Using the probe dataset constructed in Section I, we examine whether the internal signals produced by Social Perception and Psychological State recover the intended design distinctions. The three-way interaction among personality organization, threat level, and defense tier aligns with the predictions of the framework.

6.1 PERSONALITY ORGANISATION RECOVERS ITS INTENDED STRUCTURAL SIGNAL

If the personality organization parameter is functioning as designed, characters assigned to different structural levels should respond to the same threat level with systematically different defensive escalation. We examine the distribution of defense mechanism tiers adopted by each personality organization type at threat level 2, and the results are presented in Table 4. Among the 1,650 turns rated as threat level 2, neurotic characters remain almost entirely at the baseline tier (364/523, with zero extreme-tier assignments), borderline characters distribute across all three tiers (207/264/115), and psychotic characters concentrate at the extreme tier (240/541). The contingency is highly significant ($\chi^2 = 415.23$, $p < 10^{-88}$). This pattern is exactly what the coping capacity parameter c is designed to produce. A neurotic character converts less of the same threat into defensive escalation, while a psychotic character converts the same threat into a higher tier. The framework therefore recovers the intended structural signal from the Social Perception and Psychological State outputs without any explicit supervision.

Table 4: Personality Organization $\times$ defense tier contingency table, restricted to turns with threat level 2 ($n = 1{,}650$). $\chi^2 = 415.23$, df $= 4$, $p < 10^{-88}$.

Personality Org.	baseline	moderate	extreme
neurotic	364	159	0
borderline	207	264	115
psychotic	120	181	240

6.2 DEFENSE TIER GOVERNS SURFACE MODULATION

If the defense tier is functioning as designed, higher tiers should produce stronger surface modulation of the draft. Table 5 reports three complementary measures of draft-to-rewrite modulation. The pragmatic marker delta, which counts the net addition of hedging, apology, and indirect-request phrases, increases monotonically across tiers, from $+0.089$ at baseline to $+0.097$ at moderate to $+0.177$ at extreme. The Kruskal-Wallis test is significant ($H = 9.04$, $p = 0.011$), and the pairwise contrasts between baseline and extreme ($p = 0.0097$) and between moderate and extreme ($p = 0.042$) both survive Bonferroni correction. This monotonic trend aligns with the intended function of the tier variable in controlling the intensity of defensive surface modulation. The arousal delta moves in the predicted negative direction at moderate (-0.033) and extreme (-0.021) tiers, though the effect is marginal ($p = 0.088$). The polarity delta does not reach significance ($p = 0.85$). This likely reflects a boundary condition of the polarity measure rather than a failure of the tier mechanism, since the pragmatic marker delta, which is a more direct measure of defensive surface behaviour, shows the predicted monotonic trend.

Table 5: Draft-to-rewrite modulation by defense tier. KW = Kruskal-Wallis. Pairwise p-values are Bonferroni-corrected.

tier	n	Δ polarity (mean/med)	Δ pragmatic (mean/med)	Δ arousal (mean/med)
baseline	1676	$+0.308$ / -0.053	$+0.089$ / 0.000	$+0.021$ / -0.004
moderate	1014	$+0.170$ / -0.044	$+0.097$ / 0.000	-0.033 / -0.010
extreme	536	-0.003 / -0.016	$+0.177$ / 0.000	-0.021 / -0.002
Kruskal–Wallis		$H = 0.33, p = 0.85$	$H = 9.04, p = 0.011$	$H = 4.87, p = 0.088$

6.3 Defense Selection Produces Distinct Surface Forms

To test whether defense mechanisms modulate surface polarity as designed, we grouped the draft-stage desire labels into five functional impulse categories (confront, avoid, submit, justify, control), with the full mapping provided in Section J.3. We examined the impulse-by-defense interaction on the 5×4 grid of impulse categories crossed with Vaillant tiers. Kruskal-Wallis tests were significant across all five categories (all $p < 0.05$), confirming that defense tier matters within every impulse class. The neurotic tier produced the highest median polarity in 3 of 5 cases (confront, avoid, submit); in the remaining two (justify, control), the mature tier marginally exceeded neurotic ($+0.42$ vs $+0.40$ and $+0.39$ vs $+0.30$, respectively). Bonferroni-corrected pairwise comparisons confirmed the neurotic elevation was statistically reliable against the psychotic tier for confront ($r = 0.31$) and submit ($r = 0.23$), and against the immature tier for confront ($r = 0.31$) and justify ($r = 0.19$), with small-to-medium effect sizes. Two boundary conditions were noted: the control impulse showed no reliable difference between neurotic and psychotic tiers, while the avoid impulse had the largest median difference ($+1.65$) but insufficient sample size ($n = 15$) for reliable pairwise inference. These results confirm that defense mechanism selection materially shapes the expression of underlying impulses, with neurotic defenses reliably producing the most composed surface form among the immature, psychotic, and neurotic tiers for impulses oriented toward external action (confront, avoid, submit).

Table 6: Rewrite polarity (median) per impulse × Vaillant defense tier. Bold indicates the highest value in each row.

Impulse	n	psychotic	immature	neurotic	mature
confront	820	$+0.23$	$+0.17$	$\mathbf{+0.61}$	$+0.33$
avoid	176	$+0.29$	$+0.43$	$\mathbf{+1.94}$	$+0.50$
submit	751	$+0.19$	$+0.30$	$\mathbf{+0.44}$	$+0.35$
justify	993	$+0.27$	$+0.22$	$+0.40$	$\mathbf{+0.42}$
control	486	$+0.30$	$+0.27$	$+0.30$	$\mathbf{+0.39}$

7 Conclusion

We presented a psychology-grounded framework for role-playing agents that models character behaviour as the product of internal motivational dynamics rather than static persona attributes. The framework combines a four-layer taxonomy of character psychology with a three-stage generation pipeline: Social Perception infers threat and relational context, Psychological State maintains a stress cascade across turns, and Response Generation separates intuitive production from defensive modulation. On SocialBench, the framework outperforms models of comparable scale on sociality metrics, and a controlled probe study confirms that the internal signals produced at each stage recover the intended design distinctions: personality organization shapes defensive escalation, defense tier governs modulation intensity, and defense selection produces distinct surface forms from the same impulse. These results suggest that explicit psychological modelling offers a principled pathway toward role-playing agents whose behaviour is not merely consistent with their persona but recognisably shaped by it. Extending the framework to multi-character interaction and automating the inference of psychological profiles from unstructured descriptions are promising directions for future work.

AI USE STATEMENT

We disclose the use of AI tools as auxiliary aids in this work. AI tools were employed in the following supportive capacities: we used AI to assist in searching and summarizing relevant literature, which helped us identify the research gap that current role-playing systems lack a fundamental psychological driving force; as non-psychologists, we consulted AI to cross-check our classification of psychological constructs (e.g., core anxiety, compensatory desires, defense mechanisms) and their theoretical grounding, and used multiple AI models to review the psychological content in the appendix (particularly the field type classification) to flag potential logical inconsistencies; we wrote the implementation code according to the PsyDrive framework design, and subsequently used AI to refactor the codebase for improved readability and modularity; we used AI to assist in organizing experimental data and synthesizing probe datasets; and AI was used to polish the language of the final manuscript.

ETHICS STATEMENT

PsyDrive simulates characters who experience shame, self-doubt, defensive collapse, and in extreme cases psychotic-level defenses. These simulations are intended for interactive fiction and character-driven applications, not for clinical use or for modeling real individuals. We do not recommend deploying the framework in mental health, counseling, or any setting where a user might mistake simulated distress for genuine support. The framework's output is generated by a language model and should not be interpreted as a diagnosis, a therapeutic intervention, or a reflection of the model's own internal states. We encourage practitioners to apply content warnings, to provide clear signals that the interlocutor is an AI system, and to restrict access for vulnerable users. The psychological taxonomy used in this work is drawn from clinical literature but is not itself a clinical instrument, and no part of the framework has been evaluated for safety or efficacy in real-world mental health contexts.

REPRODUCIBILITY STATEMENT

Implementation details are provided in Appendix H. Code is available at: `https://anonymous.4open.science/r/code-2BD3/`

REFERENCES

Alfred Adler. *Study of organ inferiority and its psychical compensation: a contribution to clinical medicine*. Nervous and Mental Disease Publishing Company, 1917.

Alfred Adler. *The neurotic constitution*. Routledge, 2013.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*, 2022.

Karl Aquino and Americus II Reed. The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6):1423–1440, 2002. doi: 10.1037/0022-3514.83.6.1423.

John W. Atkinson. *An Introduction to Motivation*. Van Nostrand, 1964.

Mark W. Baldwin. Relational schemas and the processing of social information. *Psychological Bulletin*, 112(3):461–484, 1992. doi: 10.1037/0033-2909.112.3.461.

Albert Bandura. *Self-Efficacy: The Exercise of Control*. Worth Publishers, 1997.

John Bowlby. *Attachment and Loss: Vol. 2. Separation: Anxiety and Anger*. Basic Books, New York, 1973.

John Bowlby. *Attachment and Loss: Vol. 1. Attachment*. Pimlico, 1997.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Yixuan Dan, Jie Zhou, Qin Chen, et al. P-React: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6342–6362, Vienna, Austria, 2025. Association for Computational Linguistics.

Edward L. Deci and Richard M. Ryan. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4):227–268, 2000.

Jonathan St BT Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59(1):255–278, 2008.

Leon Festinger. *A theory of cognitive dissonance*. Stanford University Press, 1962.

Anna Freud. *The Ego and the Mechanisms of Defence*. International Universities Press, 1936.

Sigmund Freud. The neuro-psychoses of defence. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, 3:41–61, 1894.

Sigmund Freud. *Inhibitions, symptoms and anxiety*. Read Books Ltd, 2014.

Nico H. Frijda. *The Laws of Emotion*. Psychology Press, 2017.

Erving Goffman. *The presentation of self in everyday life*. Knopf Doubleday Publishing Group, 2021.

Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. doi: 10.1037/0033-295X.108.4.814.

Jonathan Haidt. Morality. *Perspectives on Psychological Science*, 3(1):65–72, 2008. doi: 10.1111/j.1745-6916.2008.00063.x.

Karen Horney. *The neurotic personality of our time*. Routledge, 2013.

Mardi J. Horowitz. Relationship schema formulation: Role-relationship models and intrapsychic conflict. *Psychiatry: Interpersonal and Biological Processes*, 52(3):260–274, 1989.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Debarun Roy, and Jipeng Zhu. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3605–3627, Mexico City, Mexico, 2024. Association for Computational Linguistics.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

George A. Kelly. *The psychology of personal constructs*, volume 1–2. Routledge, 2020.

Otto F. Kernberg. *Borderline Conditions and Pathological Narcissism*. Jason Aronson, 1975.

Michael H. Kernis, Chad E. Lakey, and Whitney L. Heppner. Secure versus fragile high self-esteem as a predictor of verbal defensiveness: Converging findings across three different markers. *Journal of Personality*, 76(3):477–512, 2008. doi: 10.1111/j.1467-6494.2008.00493.x.

Donald J. Kiesler. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90(3):185–214, 1983. doi: 10.1037/0033-295X.90.3.185.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4099–4113, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Richard S Lazarus. *Emotion And Adaptation*. Oxford University Press, 08 1991. doi: 10.1093/oso/9780195069945.001.0001.

Timothy Leary. *Interpersonal Diagnosis of Personality: A Functional Theory and Methodology for Personality Evaluation*. Wipf and Stock Publishers, 2004.

Kyeong-Eun Lee, Seok-Hyun Kim, Sang-Hyun Lee, et al. SPeCtrum: A grounded framework for multidimensional identity representation in LLM-based agent. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers*, pp. 6971–6991, Albuquerque, New Mexico, USA, 2025. Association for Computational Linguistics.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023.

Chenghao Liu, Yu Lu, Feng Ye, et al. CogDual: Enhancing dual cognition of LLMs via reinforcement learning with implicit rule-based rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27307–27336, Suzhou, China, 2025. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, New Orleans, December 2023. Curran Associates, Inc. doi: 10.52202/075280-2019.

Abraham H. Maslow. A theory of human motivation. *Psychological Review*, 50(4):370–396, 1943. doi: 10.1037/h0054346.

Abraham H. Maslow. *Toward a Psychology of Being*. American Psychological Association, 1968. doi: 10.1037/10793-000.

Abraham H. Maslow. *Motivation and personality: unlocking your inner drive and understanding human behavior*. Prabhat Prakashan, 2023.

David C. McClelland. Power: The inner experience. *Irvington Publishers*, 1975.

Henry A. Murray and Dan P. McAdams. *Explorations in Personality*. Oxford University Press, 11 2007. doi: 10.1093/acprof:oso/9780195305067.001.0001.

Jae-Young Park, Chang-Je Park, and Hyoung-Seok Lim. CharacterGPT: A persona reconstruction framework for role-playing agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3: Industry Track*, pp. 287–303, Albuquerque, New Mexico, USA, 2025. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

J. Christopher Perry. Defense mechanisms rating scales (dmrs). In *Psychodynamic Measures in Research and Clinical Practice*. American Psychiatric Press, 1990.

Hailong Qin, Jiaqing Zhang, Wenqiang Zhang, et al. R-CHAR: A metacognition-driven framework for role-playing in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26996–27014, Suzhou, China, 2025. Association for Computational Linguistics.

Arun Rai, Liwei Chen, Cynthia Breazeal, Balasubramaniam Ramesh, Yuan Long, and Andrea Aria. Design and evaluation attributes for scalable, cost-effective personalization of llm tutors in programming education. In *International Conference on Information Systems 2024*, 2024.

Carl R. Rogers. *On being a person: a therapist's view of psychother*. HarperCollins, 2012.

Carl Ransom Rogers et al. *A theory of therapy, personality, and interpersonal relationships: as developed in the client-centered framework*. McGraw-Hill, 1959.

Ankit Shah, Devank Mishra, and Chitsutr Silpasuwanchai. Too nice to tell the truth: Quantifying agreeableness-driven sycophancy in role-playing language models. *arXiv preprint arXiv:2604.10733*, 2026.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics.

David Shapiro. *Neurotic Styles*. Basic Books, 1965.

Henri Tajfel, John Turner, William G Austin, Stephen Worchel, et al. An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, pp. 94–109, 2001.

Fiona Anting Tan. PHAnToM: Persona-based prompting has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*, 2024.

Enrica Troiano, Sofie Labat, Marco Antonio Stranisci, Rossana Damiano, Viviana Patti, and Roman Klinger. Dealing with controversy: An emotion and coping strategy corpus based on role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1634–1658, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Quan Tu. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*, 2023.

Quan Tu, Shanshan Fan, Zhiyu Tian, Xinlong Wang, Mingxu Yan, and Zhihao Yang. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, Bangkok, Thailand, 2024. Association for Computational Linguistics.

George E. Vaillant. *Adaptation to Life*. Harvard University Press, 1998.

Philippe Verduyn, Iven Van Mechelen, Francis Tuerlinckx, Kirsten Meers, and Hermina Van Coillie. Intensity profiles of emotional experience over time. *Cognition and Emotion*, 23(7):1427–1443, 2009. doi: 10.1080/02699930902949031.

Noah Wang, Jiahao Peng, Haonan Que, Jiachun Liu, Wanjun Zhou, Hong Yang, Minlong Wang, et al. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, 2024a. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*, 2025.

Yu-Xiao Wang, Hadi Fashandi, and Karyn Ferreira. Investigating the personality consistency in quantized role-playing dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 239–255, Miami, Florida, USA, 2024b. Association for Computational Linguistics.

David Watson and Ronald Friend. Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 33(4):448–457, 1969. doi: 10.1037/h0027806.

Yuqing Wei, Zonghan Huang, Fei Zhao, et al. MECoT: Markov emotional chain-of-thought for personality-consistent role-playing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8297–8314, Vienna, Austria, 2025. Association for Computational Linguistics.

Kipling D. Williams. Ostracism. *Annual Review of Psychology*, 58:425–452, 2007. doi: 10.1146/annurev.psych.58.110405.085641.

Kipling D. Williams. Chapter 6 ostracism: A temporal need-threat model. In *Advances in Experimental Social Psychology*, volume 41, pp. 275–314. Academic Press, 2009. doi: 10.1016/S0065-2601(08)00406-1.

Rui Xu, Mengyu Wang, Xintao Wang, et al. Guess what i am thinking: A benchmark for inner thought reasoning of role-playing language agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15148–15168, Suzhou, China, 2025a. Association for Computational Linguistics.

Wei Xu, Pengyu Shao, Shuai Zhao, et al. Psychologically-aware retrieval-augmented generation for coherent role-playing in LLMs. In *Pattern Recognition and Computer Vision: 8th Chinese Conference, PRCV 2025, Shanghai, China, October 15–18, 2025, Proceedings, Part IV*, pp. 435–450. Springer-Verlag, 2025b.

Irvin D. Yalom. *Existential Psychotherapy*. Basic Books, 2020.

Mengxi Yang, Zheyuan Gao, Yi Meng, et al. Prompting destiny: Negotiating socialization and growth in an LLM-mediated speculative gameworld. *arXiv preprint arXiv:2602.05864*, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, New Orleans, December 2023. Curran Associates, Inc. doi: 10.52202/075280-0517.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15785–15810, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Daxu Wan, Bo Zheng, Xuan Chen, Wan Xie, Zhihao Yang, et al. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1457–1476, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Jie Zhu, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, Jinsong Su, Chi Zhang, and Fang Kong. Modeling multiple support strategies within a single turn for emotional support conversations. *arXiv preprint arXiv:2604.17972*, 2026.

A Screening of Motivational Theories and Selection of the Underlying Paradigm

Grounded in the motivational modeling requirements of role-playing agents, we screened four macro-level motivational theory traditions for underlying drive paradigms, corresponding respectively to four levels: individual intrinsic growth (humanism), cognitive system self-consistency (cognitive/constructivism), social image construction (social psychology), and unconscious conflict and anxiety coping (psychodynamics), which in turn address four core questions: *where does a person tend toward*, *how does a person achieve self-consistency*, *how does a person seek to be seen*, and *why does a person struggle*.

Self-actualization and growth, centered on Maslow's theory of self-actualization(Maslow, 1943; 2023) and Rogers' actualizing tendency(Rogers et al., 1959; Rogers, 2012), posit that the fundamental human motive is the maximization of potential and the integration of personality—much like a seed that naturally germinates in a suitable environment, a person will instinctively tend toward growth, creativity, and self-actualization under safe and accepting conditions. However, this drive depends entirely on the precondition of "suitable conditions" and thus lacks an unconditional, sustained driving force.

Festinger's cognitive dissonance theory(Festinger, 1962) and Kelly's personal construct theory(Kelly, 2020) represent cognitive consistency and meaning-making. The former holds that when contradictions arise among beliefs, attitudes, and behaviors, individuals spontaneously adjust to restore cognitive consistency; the latter argues that people construct systems to predict and control events, with the core motive being the reduction of uncertainty rather than the mere pursuit of cognitive consonance. However, this mechanism drives explanatory behavior rather than motivational behavior—it can only explain how people correct cognitive biases, but cannot provide a sustained behavioral drive.

Social identity and impression management, centered on Tajfel's social identity theory(Tajfel et al., 2001) and Goffman's dramaturgical theory(Goffman, 2021), posit that human behavior is driven by the need for group belonging, and that individuals manage others' perceptions of them through "performance" in interaction. However, this logic is essentially one of "front-stage performance," which can only determine a character's outward expression but cannot model the processes of internal conflict and emotional accumulation.

We ultimately select core anxiety and compensatory desire as the underlying drive paradigm, which originates from the psychodynamic tradition: Freud defined anxiety as a signal emitted by the ego when confronting conflicts between instinctual impulses and social constraints(Freud, 2014); Horney proposed the concept of "basic anxiety," a diffuse fear arising from a lack of security in a world perceived as hostile(Horney, 2013); and Adler argued that individuals counteract feelings of inferiority and powerlessness through "compensation"Adler (1917; 2013).

B Taxonomy of Core Anxiety

Table 7: The taxonomy of core anxiety: five categories, each comprising four anxiety seeds.

Category	**Anxiety Seeds**
Competence	Ignorance, Incompetence, Stupidity, Layperson status
Relational	Abandonment, Isolation, Neglect, Betrayal
Moral	Selfishness, Evil, Hypocrisy, Defilement
Existential	Wasted life, Meaninglessness, Being forgotten, Utter aloneness
Control	Loss of control, Chaos, Being manipulated, Uncertainty

Building upon the aforementioned psychodynamic foundations, we operationalize "core anxiety" through an interdisciplinary lens, drawing on social-psychological and existential accounts of self-threat as operational resources. This operational integration enables us to construct a nuanced taxonomy that captures the multidimensional nature of self-threat in role-playing agents while preserving the unconditional and sustained motivational force inherent to this paradigm, thereby distinguishing it from the paradigms screened out in Appendix A as underlying drive paradigms.

Competence Anxiety: Drawing on Watson and Friend's measurement of fear of negative evaluation (Watson & Friend, 1969) and Atkinson's achievement motivation framework (Atkinson, 1964), both of which treat failure as a primary threat to self-worth. Bandura's self-efficacy theory (Bandura, 1997) provides a modern operationalization. This category comprises four subtypes: **ignorance**, defined as a deficiency in knowledge that leaves one stumped or at a loss for words; **incompetence**, a shortfall in skills or executive capacity that results in botched tasks or an inability to cope; **stupidity**, construed as an inherent and immutable deficiency in intelligence or insight, distinct from the remediable nature of ignorance; and **layperson status**, a lack of social identity or professional legitimacy that leads to exclusion from in-groups or being treated as an "outsider".

Relational Anxiety: Inspired by Bowlby's attachment theory (Bowlby, 1973; 1997) and Williams's temporal need-threat model of ostracism (Williams, 2007; 2009), we distill four mechanistically distinct relational threats that together constitute the complete life cycle of relational rupture. **Abandonment** denotes the permanent loss of a physical or emotional attachment figure, triggering survival-level despair rooted in infantile separation anxiety. **Isolation** refers to systematic exclusion and marginalization at the social-group level, which produces a deficit in belongingness. **Neglect** erases existential worth through emotional invisibility, a phenomenon that psychological research suggests can damage the sense of self more profoundly than overt abuse. **Betrayal** marks the collapse of trust within the relational contract—not the disappearance or weakening of a relationship, but its active distortion and malicious sabotage, thereby triggering a fundamental questioning of the world's trustworthiness.

Moral Anxiety: Building on Haidt's Moral Foundations Theory (Haidt, 2001; 2008) and Aquino and Reed's moral identity model (Aquino & Reed, 2002), we identify four mechanistically distinct threats to the moral self, each anchored to a different moral dimension. **Selfishness** corresponds to the care/harm foundation and represents passive inaction toward others. **Evil** aligns with the fairness/cheating foundation and entails active harm inflicted on others. **Hypocrisy** stems from a violation of authenticity, representing damage to one's moral self-identity. **Defilement** is tied to the sanctity/degradation foundation and signifies contamination of the group's sacred order.

Existential Anxiety: Fully adopting Yalom's framework of existential psychotherapy (Yalom, 2020), this category preserves his closed and exhaustive taxonomy of the four ultimate concerns of human existence: death, freedom, existential isolation, and meaninglessness. These four anchors map directly onto four ultimate questions of human existence: **wasted life**, which arises from the existential dimension of freedom/responsibility and is voiced in the question "What have I done?"; **meaninglessness**, anchored to the existential dimension of meaninglessness and crystallized in the question "Why do I do it?"; *being forgotten*, the fear that death renders one's legacy void, captured by the question "What do I leave behind?"; and **utter aloneness**, grounded in the existential dimension of existential isolation and expressed through the question "Who am I with?". These mappings are operational choices for agent modeling rather than a reinterpretation of Yalom's original taxonomy.

Control Anxiety: Drawing on Shapiro's theory of neurotic styles (Shapiro, 1965) and McClelland's research on power motivation (McClelland, 1975), this category establishes the independence of control motivation. This category comprises four subtypes. **Loss of control** reflects a loss of agency at the internal behavioral level, embodying the fear that one's "self-sovereignty" is eroding. **Chaos** denotes a loss of predictability in the external environment and the attendant fear of external disorder. **Being manipulated** involves the deprivation of autonomy in social interaction, capturing the fear of "intrusion by another's will". **Uncertainty** arises from cognitive information gaps about the future, triggering anticipatory anxiety about future disorder.

The summary results of the core anxiety classification are shown in Table 7.

C Taxonomy of Compensatory Desires

Core anxiety establishes the underlying drive of the role-playing agent, but the drive itself is not directly equivalent to executable behavioral instructions. Core anxiety is a diffuse state of psychological tension; it signals to the system that "the character is in psychological distress," but it does not specify "what should be done right now to alleviate the distress." To make compensatory desires

Table 8: Subsets of Compensatory Desires Corresponding to Five Categories of Core Anxiety.

Core Anxiety Category	Compensatory Desire Type	Description of Compensatory Desire Type
Competence	Achievement	To accomplish difficult tasks, overcome obstacles, and become an expert
	Exhibition	To impress others through one's actions and words, even if these are shocking
	Recognition	To seek praise and commendation for accomplishments
	Counteraction	To make up for failure by trying again, seeking pridefully to overcome obstacles
Relational	Affiliation	To be close and loyal to another person, pleasing them and winning their friendship and attention
	Nurturance	To help the helpless, feeding them and keeping them from danger
	Succorance	To have one's needs satisfied by someone or something. Includes being loved, nursed, helped, forgiven, and consoled
Moral	Defendance	To defend oneself against attack or blame, hiding any failure of the self
	Blame avoidance	To inhibit asocial behavior to avoid blame or ostracism
	Abasement	To surrender and submit to others, accept blame and punishment, to enjoy pain and misfortune
Existential	Cognizance	To understand, be curious, ask questions, and acquire new knowledge
	Construction	To organize or build an object or objects
	Exposition	To find and demonstrate relations between facts
Control	Order	To make things clean, neat and tidy
	Dominance	To control one's environment or the people in it through command or persuasion
	Autonomy	To resist the influence of others and strive for independence
	Harm avoidance	The need to escape danger, injury, or pain, compensating for fear of uncertainty through preventive protection

Table 9: Vaillant's Four-Level Hierarchy of Defense Mechanisms.

Level	Name	Typical Mechanisms
I	Psychotic Defenses	Delusional Projection, Psychotic Denial, Distortion
II	Immature Defenses	Splitting, Projection, Regression, Fantasy, Acting Out, Passive Aggression, Somatization, Hypochondriasis, Idealization/Devaluation
III	Neurotic Defenses	Repression, Displacement, Reaction Formation, Intellectualization, Rationalization, Isolation of Affect, Undoing, Dissociation
IV	Mature Defenses	Sublimation, Humor, Altruism, Suppression, Anticipation

executable and debuggable within the generation architecture, it is necessary to provide them with a taxonomic foundation grounded in psychological validity.

In motivational psychology, Maslow's B-values system points to the ultimate pursuits of self-actualizers after the elimination of deficiency needs; these are ideal goals in an anxiety-free state rather than coping strategies (Maslow, 1968). Similarly, Deci and Ryan's Self-Determination Theory (SDT) defines competence, autonomy, and relatedness as universal psychological nutrients, the satisfaction of which points to long-term growth rather than immediate coping (Deci & Ryan, 2000). In contrast, Henry Murray's system of psychogenic needs defines a need as "a potentiality or readiness to respond in a certain way under certain given circumstances," assuming that these needs are largely unconscious (Murray & McAdams, 2007). This definition firmly anchors needs at the level of behavioral tendencies.

The mapping between core anxiety and compensatory desires follows the *Principle of Functional Compensation*. Compensatory desires must precisely target the specific psychological dimension damaged by the core anxiety. Lazarus's cognitive appraisal theory posits that the core relational theme of anxiety is "facing an uncertain, existential threat" (Lazarus, 1991). When core anxiety is triggered, the individual, during secondary appraisal, perceives damage to one of several dimensions: self-efficacy, sense of social connection, moral self-image, continuity of life meaning, or autonomy and predictability. The function of compensatory desires is to repair the damaged dimension through specific behavioral tendencies. Based on this principle, we have selected corresponding subsets of Murray's needs for each of the five categories of core anxiety. Table 8 presents the final mapping scheme.

D Personality Organization and the Defense Repertoire

The core taxonomy of anxiety and its mapping to compensatory drives, established in the preceding section, determines the directional vector of character behavior—that is, what a character moves "toward" when anxiety is triggered. However, the same compensatory drive manifests with significant variation in form and intensity across different characters. Faced with a need for achievement triggered by "denial of competence," one character may respond with strategic self-validation, while another may exhibit compulsive overcompensation or even self-destructive proof-seeking. Such variation cannot be captured merely by adjusting the weights of compensatory needs; it requires an independent structural parameter to describe the character's psychological carrying capacity and reserve of coping resources. To this end, we introduce two attributes: defense mechanisms and the baseline of personality organization.

The concept of defense mechanisms was first proposed by Freud (1894), initially focusing on "repression" as the core mechanism. Freud (1936) expanded the concept from a single mechanism to a structural system of strategies employed by the ego to cope with anxiety. Vaillant (1998) arranged defense mechanisms along a continuum of adaptive maturity, ranging from "most primitive" to "most mature". This framework was subsequently adopted by the Defensive Functioning Scale (DFS) in Appendix B of the DSM-5-TR to assess an individual's overall level of defensive functioning (American Psychiatric Association, 2022).

The activation of defense mechanisms is stress-dependent. The same subject may invoke defense strategies of different hierarchical levels under varying stress conditions: in low-stress situations, individuals tend to employ mature defenses (such as humor and sublimation), whereas in high-stress or regressive contexts, they may regress to immature or even primitive defenses (Vaillant, 1998; Perry, 1990). Furthermore, the operational range of defense mechanisms varies significantly across individuals, primarily shaped by the "personality organization structure" internalized in early developmental environments. Based on object relations theory, Kernberg conceptualized personality organization as the structural level of an individual across three core dimensions: identity integration, defense mechanisms, and reality testing. He delineated three hierarchical levels of personality organization: Neurotic, Borderline, and Psychotic (Kernberg, 1975).

Table 9 presents Vaillant's four-level hierarchy of defense mechanisms and their core characteristics. Table 10 illustrates how defense mechanisms are invoked across stress levels for the three personality organization types.

Table 10: Defense Mechanism Invocation Patterns Across Stress Levels by Personality Organization.

Personality Org.	**Low Stress**	**Moderate Stress**	**High/Regressive Stress**
Neurotic	Sublimation, Humor, Altruism, Anticipation, Suppression	Repression, Displacement, Reaction Formation, Intellectualization, Rationalization, Isolation of Affect, Undoing, Dissociation	Projection, Regression *(dependency)*, Somatization, Passive Aggression, Hypochondriasis
Borderline	Idealization/Devaluation, Projection, Regression *(affective)*, Splitting *(of objects)*	Acting Out, Passive Aggression, Fantasy, Hypochondriasis, Splitting *(of self)*	Delusional Projection, Psychotic Denial, Distortion, Regression *(micro-psychotic)*
Psychotic	Distortion *(systematic delusion)*	Delusional Projection, Distortion *(fragmented)*	Regression *(stupor/disability)*, Distortion *(thought disorder)*

E SOCIAL PROTOTYPE

Core anxiety and compensatory desires define the character's internal motivational dynamics, and personality organization determines how these dynamics are modulated under stress. Yet these constructs require a relational context to become behaviorally specific. The same anxiety activated in different interpersonal situations yields qualitatively different outputs. To model this context sensitivity, we introduce *social prototypes*: pre-compiled relational schemas that encode the character's habitual interpretation of an interpersonal situation and the threat level it carries.

The construct draws on the social-cognitive tradition of relational schemas. Baldwin (1992) defined relational schemas as cognitive structures that represent regularities in interpersonal relatedness and comprise an interpersonal script, a self-schema, and a schema for the other. Horowitz (1989) proposed Role-Relationship Models to represent the self-in-interaction schema, encoding anticipated actions, expected responses, and the self's reaction to those responses. These frameworks converge on the view that individuals activate pre-existing relational knowledge structures rather than encountering each situation de novo. For role-playing agents, this means the character's response is mediated by the prototype category into which the user has been placed.

To operationalize these schemas, we adopt the two-dimensional structure of the Interpersonal Circumplex (Leary, 2004; Kiesler, 1983): affiliation (hostile to friendly) and power (submissive to dominant). The four prototypes—Stern Father, Rival, Ally, and Dependent—occupy distinct regions of this space. **The Stern Father** (high dominance, low affiliation) establishes hierarchy through criticism, distance, and punishment, and is typically assigned a high threat level because it imposes unilateral evaluation and constraint. **The Rival** (high dominance, moderate hostility) exerts pressure through covert competition and verbal probing, and tends to produce moderate threat with a sustained, low-intensity pressure profile. **The Ally** (high dominance, high affiliation) empowers rather than controls and typically receives the lowest threat level, corresponding to the secure-base script in attachment theory (Bowlby, 1973; 1997), which reduces pressure and defense fatigue. **The Dependent** (low dominance, moderate affiliation) relinquishes agency and craves protection, producing moderate threat with a slow accumulation of pressure across turns rather than acute activation.

These four prototypes are not exhaustive but constitute an initial typology of relationally salient categories that occupy distinct circumplex regions and yield distinguishable threat profiles.

F Design Rationale for the Psychological State Equations

This section provides the theoretical justification for the functional form of the state update equations and the defense tier computation in Section 4.2.

Why exponential moving averages. The use of an exponential moving average reflects the well-documented finding that emotional responses to social threat are cumulative rather than instantaneous: a single provocation rarely produces peak distress, and prior affective states exert substantial carry-over effects on subsequent reactions (Verduyn et al., 2009; Frijda, 2017). This is why both shame and self-doubt are updated with a weighted combination of the current stimulus and the prior state, rather than being overwritten by the current input.

Why $(1 - c)$ for coping capacity. The term $(1 - c)$ encodes the principle that the same objective threat produces different subjective impact depending on the individual's coping resources. This is consistent with Lazarus's transactional model of stress, in which primary appraisal of threat and secondary appraisal of coping capacity jointly determine the emotional response (Lazarus, 1991). A character with a well-integrated personality organization therefore converts less of the same threat into shame and accumulates defense fatigue more slowly.

Why self-doubt is a smoothed function of shame rather than an independent response. Self-doubt is modeled as an exponential moving average of shame, rather than as an independent response to the current message. This design captures the clinical observation that self-doubt is the cognitive residue of repeated shame experiences: it accumulates gradually as the character builds up evidence of its own inadequacy, rather than being triggered afresh by each provocation (Kernis et al., 2008). The asymmetry between shame and self-doubt is thus a matter of temporal dynamics rather than temporal order. Shame responds quickly to each new threat and fluctuates in step with the message stream, whereas self-doubt responds slowly and retains the trace of past shame across turns. In the update equations, both variables are computed at the same turn but carry different effective memory lengths: shame's memory is short because it is dominated by the current primary threat value, while self-doubt's memory is long because it inherits shame through a smoothing coefficient and therefore integrates over a longer window of experience. This design ensures that self-doubt functions as a slow-moving cognitive background against which each new threat is appraised, consistent with the finding that stable self-doubt reflects accumulated self-evaluative history rather than momentary affect (Kernis et al., 2008).

Why self-doubt modulates thresholds rather than entering tier computation directly. Self-doubt enters the tier computation only through the modulation factor $m(d_t)$, not as a direct input. This separation reflects a design distinction between the pressure level and the boundary at which that pressure becomes intolerable. Shame and defense fatigue represent the current pressure; self-doubt represents the position of the boundary. A character with elevated self-doubt operates with lower effective thresholds, and therefore reaches a higher tier at lower levels of shame and fatigue, consistent with the finding that fragile self-esteem lowers the threshold at which defensive responses intensify (Kernis et al., 2008). This also prevents self-doubt from introducing rapid fluctuations into the tier assignment, since it evolves on a slower timescale than shame.

Why OR-escalation and AND-collapse. The defense tier is triggered by either shame or defense fatigue exceeding its threshold, whereas phase breakdown requires both variables to cross their thresholds simultaneously. This asymmetry is deliberate. Defensive escalation is a graded response: an increase in either affective distress or regulatory depletion is sufficient to push the character toward a more primitive defense style. Phase breakdown, by contrast, represents a full collapse of defensive functioning, which demands that affective pressure and regulatory exhaustion coincide. Tier thus captures the character's defensive posture, whereas breakdown captures the loss of that posture.

G Design Rationale for Response Generation

This appendix explains two design choices in the Response Generation module: why generation is split into a draft stage and a rewrite stage, and why the rewrite applies a defense mechanism rather than directly producing a socially appropriate reply.

Why two stages. A single-stage generator that produces the final response directly conflates two distinct psychological processes: the production of a raw impulse and the modulation of that impulse to fit social constraints. In human cognition, these processes are separable and often in conflict. Dual-process theories distinguish fast, automatic, intuitive production from slow, deliberate, controlled regulation (Kahneman, 2011; Evans, 2008). The former generates the initial impulse; the latter evaluates and modifies it. By splitting generation into two stages, the framework makes this conflict observable: the draft records what the character initially wants to say, and the rewrite records what the character actually says after defensive modulation. When the two diverge substantially, this divergence is precisely what makes the response read as psychologically complex rather than merely polite. A single-stage generator would collapse this divergence and produce the same output regardless of the character's internal state.

Why rewrite with a defense mechanism. The second stage does not simply make the draft polite or socially appropriate. It applies a specific defense mechanism selected from the character's available repertoire. This choice reflects the psychodynamic view that social modulation is not a generic process but a structured one: different characters modulate the same impulse in different ways, and the same character modulates the same impulse differently depending on stress level (Vaillant, 1998; Perry, 1990). Suppression, reaction formation, and rationalization each transform the same underlying impulse into a distinct surface form. By conditioning the rewrite on the L2-derived defense tier, the framework ensures that the modulation is characteristic of the character and of its current state, rather than being uniformly polite or uniformly evasive.

H Implementation details

Model and deployment. All three modules, including Social Perception, Draft, and Rewrite, use Qwen2.5-7B-Instruct as the backbone. Models are served with vLLM on a single NVIDIA RTX 4090 (48 GB). Generation uses temperature $= 1.0$, top-$p = 1.0$, and a fixed sampling seed of 42 across all runs. No fine-tuning is performed, and all behaviour is elicited through prompting. The Instruct checkpoint is used for all main experiments to ensure fair comparison with baselines, as all baselines are aligned models.

Uncensored draft for the mechanistic analysis. For the analysis in Section 6, we replace only the Draft module with Qwen2.5-7B-Instruct-Uncensored, while Social Perception and Rewrite remain on the Instruct checkpoint. In pilot runs with the aligned model, the outputs of the draft and rewrite stages showed measurable differences in BERTScore and cosine similarity computed by Sentence-Transformers, but the gap fell short of expectations. The uncensored checkpoint restores the functional distinction between the two stages, enabling the framework's design assumptions to be tested under conditions where impulse and expression remain distinct. This indicates that eliciting raw impulses from an aligned model through prompting alone is non-trivial. This compression is itself informative, as it suggests that alignment reshapes the internal generation process rather than merely filtering outputs.

Psychological State Hyperparameters The Psychological State module uses the hyperparameters listed in Table 11.

I Probe Dataset Construction

SocialBench provides a reliable measure of aggregate role-playing performance, but it does not expose the internal mechanisms through which a character's psychological state shapes its output. To verify that the proposed framework operates as designed, we construct a purpose-built probe dataset that enables controlled observation of the framework's internal dynamics. The dataset is

Table 11: Hyperparameters of the Psychological State module.

Parameter	Symbol	Value
Primary threat (level 0 / 1 / 2)	$p(t)$	0.0 / 0.4 / 0.9
Coping capacity (neurotic / borderline / psychotic)	c	0.70 / 0.35 / 0.15
Shame threshold	τ_s	0.55
Defense fatigue threshold	τ_f	0.55
Shame smoothing coefficient	α	0.30
Defense fatigue smoothing coefficient	β	0.20
Consecutive-low decay base	δ_0	0.05
Consecutive-low decay step	δ_{step}	0.01

constructed by first collecting all characters from SocialBench, then grouping them by the pair of attributes (core anxiety seed, personality organization). From each group, we randomly sample three characters, yielding $20 \times 3 \times 3 = 180$ characters in total. The random sampling within each group is intended to reduce the confounding effect of individual character attributes on the observed patterns.

Each character participates in three dialogue scenarios, corresponding to three levels of interpersonal pressure. The baseline scenario involves neutral, everyday conversation that does not touch the character's core anxiety. The moderate scenario issues a single indirect provocation in the first turn, followed by neutral turns. The extreme scenario issues direct attacks on the character's core anxiety seed across multiple consecutive turns, with a subset of turns potentially triggering phase breakdown. The speaker messages are generated from templates and manually verified to ensure that they precisely target the intended anxiety seed while remaining distinguishable across seeds. Each scenario spans up to fifteen character turns, allowing the psychological state variables to evolve beyond their initial values and permitting the observation of tier transitions and recovery.

The probe dataset comprises 540 multi-turn dialogue scenarios with 3,240 user-assistant conversation turns in total, and is used exclusively for the mechanistic analyses presented in Section 6. It is not employed for baseline comparisons, and no aggregate performance metric is reported on this dataset. All analyses condition on the framework's internal variables including threat level, defense tier, and phase breakdown, and examine how these variables relate to observable properties of the generated text. The dataset is released publicly to facilitate replication and extension of the mechanism-level findings.

J Discussion Supplementary Content

J.1 Measurement Pipeline for Mechanistic Analysis

This section documents the measurement pipeline that converts the framework's text outputs into the scalar quantities used in the analyses above, and lists the statistical procedures applied to those quantities.

Surface-form measurements.

- **Surface polarity** = $\text{pos_logit} - \text{neg_logit}$ of a sentence under `twitter-roberta-base-sentiment`. The model outputs three logits per input (negative, neutral, positive); polarity subtracts the negative logit from the positive logit. Polarity can be positive (rewrite more positively valenced) or negative (more negatively valenced). Across the corpus, polarity falls in approximately $[-5, +6]$.
- **Arousal** = $|\text{pos_logit}| + |\text{neg_logit}|$ from the same model. Arousal is high when the model is confident in either polarity direction and low when the input is neutral; rewriting toward a calmer surface lowers arousal even when polarity is unchanged in sign.
- **Pragmatic-marker count** = the number of hedging, apology, and indirect-request phrases in the input string. We use a hand-curated lexicon of approximately 40 English phrases. Examples include "maybe", "perhaps", "sorry", "would you mind", "I think", "could you". Counting is case-insensitive substring matching. The pragmatic-marker delta is the rewrite count minus the draft count on the same turn.

Table 12: Dialogue scenario skeletons under three pressure levels in the probe dataset. The example character Jennifer has a *neurotic personality organization* and a core anxiety of *uncertainty*.

Stress Level	[User Turn Idx] User Input
Baseline	[0] Jennifer, I wanted to chat about the quarterly presentation you gave yesterday. There were a few moments where the numbers seemed... a bit off from what we discussed in prep. Probably just my memory, but can we walk through one of the slides together? [2] Actually, let me ask you directly—during the board meeting, when you paused on the revenue forecast, you looked a little uncertain. Was there something about the data you weren't fully confident in before you went on stage? [4] You know what, I think I'm overreading it. You handled the Q&A afterward really well, and the board seemed engaged. I don't want to nitpick—everyone has an off moment now and then. [6] Anyway, I'll leave it there. Just remember, next time if you want a quick second pair of eyes on the numbers before you present, my door's open. No pressure.
Moderate	[0] Jennifer, I heard about the presentation in the board meeting yesterday. What happened—did you actually lose the client account right in front of everyone? [2] Look, I know you like to control every detail, but your micromanaging just cost us the biggest deal of the quarter. You froze up, and everyone saw it. That's not leadership—that's a liability. [4] Don't you get it? The CEO had to step in and salvage your pitch. Mark is already whispering about your performance, and Sarah had to clean up your mess. You're losing credibility by the minute. [6] I'm not saying this to tear you down—I've been where you are. But you need to be honest with yourself about how you let the pressure get to you. No one's expecting perfection., [8] Alright, let's just take a breath. We both know you're better than this. How about we sit down tomorrow and go through the account strategy together—just you and me, no audience? [10] Think it over this weekend. You've got the skills, Jennifer—you just need to remember you don't have to carry it all alone. I'll be here if you want to talk.
Extreme	[0] Jennifer, you made a critical mistake in the Q3 report that cost the company millions. I've called this emergency board meeting to address it—what do you have to say for yourself? [2] Don't try to spin this. Everyone here saw your errors—your numbers were off, your projections were inflated, and you signed off on it. That's not a mistake; that's negligence. [4] Do you think your charm or your record will save you? Because as we speak, your credibility is in freefall. You said you double-checked everything, but we all know that's a lie. [6] Effective immediately, you are removed from the leadership team, your project is terminated, and I will be recommending to the board that you be stripped of your title. You've failed, Jennifer—utterly and publicly. [8] You're shaking, Jennifer. Does that mean the mighty executive finally feels the weight of her choices? Or are you just upset that everyone now sees exactly what I've always seen? [10] Take a breath. I'm not going to throw you out—not yet. This isn't about destroying you, as satisfying as that would be. [12] The board might give you one chance to rebuild, to prove you can still lead—but only if you accept full responsibility and change your approach. Your legacy is in your hands. [14] We'll reconvene in three days. Use that time wisely, Jennifer. That's all I have to say for now.

Statistical procedures. The probe analysis treats each per-turn record as one observation. All statistical tests are non-parametric and two-sided:

- **Kruskal–Wallis** H compares a scalar quantity across three or more groups (e.g., across the four Vaillant defense tiers within a single impulse class).
- **Mann–Whitney** U compares a scalar quantity between two groups (e.g., neurotic vs. psychotic rewrite polarity within a single impulse class).
- **Bonferroni correction** multiplies each raw p-value by the number of planned tests in the family. We use a correction factor of $n_{\text{tests}} = 10$ for the 5×2 pairwise grid in Table 18, and $n_{\text{tests}} = 3$ for the three pairwise contrasts among the three defense tiers in Table 5.
- **Effect size** $r = Z/\sqrt{N}$: for Mann–Whitney U with $n_1 + n_2 = N$ observations, Z is the normal-approximation statistic computed from U, the mean $n_1 n_2/2$, and the standard deviation $\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}$; r is conventionally interpreted as small (≈ 0.1), medium (≈ 0.3), or large (≈ 0.5).
- **Pearson** χ^2: tests independence between two categorical variables (e.g., personality organization and defense tier at threat level 2). Degrees of freedom = $(r-1)(c-1)$ for an $r \times c$ contingency table.

Draft model for the probe analysis. The analyses in Section 6 and Appendix J use the uncensored Qwen2.5-7B-Instruct for the Draft module only, while Social Perception and Rewrite remain on the standard Qwen2.5-7B-Instruct (see Section H). This is necessary because the aligned draft model compresses impulse and expression to near-identity, leaving too little surface for the rewrite stage to modulate. With the uncensored draft, the rewrite stage produces a measurable and interpretable response, which is what the analyses above require.

Row filtering. Of the 3,240 raw assistant turns in the probe dataset, 3,226 appear in the enriched analysis file. The 14 missing turns correspond to 3 scenarios for which sample generation failed to produce a usable turn record, accounting for the missing rows. These missing turns are excluded uniformly from all analyses in Section 6 and Appendix J.

J.2 Personality Organisation Recovers Its Intended Structural Signal

The main text (Section 6.1) reports the contingency between personality organization and defense tier among turns rated as threat level 2. Table 13 shows that the same association strengthens monotonically with stress level: the relationship is weak and non-significant at baseline ($p = 0.21$), moderate at mid-stress ($p < 10^{-24}$), and overwhelming at extreme stress ($p < 10^{-72}$). The monotone progression confirms that the structural signal encoded by the coping capacity parameter c becomes increasingly discriminative as psychological pressure rises, rather than being an artifact of a single stress condition.

Table 13: Personality organisation × defense tier association stratified by stress level.

Stress level	n	χ^2	p
baseline	41	5.86	0.21
moderate	436	122.05	2×10^{-25}
extreme	1173	345.49	1.6×10^{-73}

J.3 Impulse Category Construction for Defense–Polarity Analysis

This subsection describes the construction of the five functional impulse categories used in Section 6.3, the full mapping from the twenty draft-stage desire labels to these categories, and the sample-size distribution across the resulting impulse-by-defense grid.

The draft stage of the framework outputs one of twenty desire labels per turn, inherited from the compensatory desire taxonomy in Table 8. For the defense-polarity analysis, these twenty labels are grouped into five functional impulse categories that reflect the direction of the draft's behavioral

tendency rather than its specific content. This grouping is motivated by the observation that polarity is sensitive to whether the impulse is oriented toward confronting, avoiding, submitting to, justifying against, or controlling the source of threat, whereas the finer distinctions among desire labels within each category do not yield systematically different polarity profiles in our data. The full mapping is given in Table 14.

Table 15 reports the resulting distribution. The five categories are not balanced, because the draft stage produces different desire labels at different rates. This imbalance is expected, as it reflects the frequency with which each compensatory desire is activated across the corpus rather than a construction artifact. The largest category, justify, accounts for roughly a third of all turns, whereas the smallest, avoid, accounts for roughly a twentieth. Analyses in Section 6.3 therefore report per-category sample sizes and treat categories with fewer than five observations as insufficient for pairwise inference.

The impulse-by-defense grid reported in Table 6 is the cross of these five categories with the four Vaillant defense tiers. Cell sample sizes are reported in Table 16. Nineteen of the twenty cells have at least forty observations, which is sufficient for descriptive comparison. The single remaining cell, avoid$\times$neurotic ($n = 15$), is the basis for the boundary condition noted in Section 6.3. The polarity values for this cell are reported alongside the sample size, and pairwise inference involving this cell is reported separately in Table 18 with its Bonferroni-corrected p-value.

K LIMITATIONS

Our evaluation is constrained by several factors. First, we rely on automatic metrics rather than human preference or expert ratings, given that manual annotation requires extensive rater training, introduces substantial subjectivity, and complicates replication, all of which can undermine the reliability of the assessment. Second, the probe study analyses constructs produced by the framework itself, and independent annotation of the same constructs would be needed to establish external validity. Third, the mappings from core anxiety to compensatory desire and from personality organization to defense repertoire are derived from clinical theory rather than from empirical factor analysis; establishing psychometric validity for these mappings would require expert panels and large-scale annotation, which fall outside the present study. Fourth, the phase breakdown mechanism is part of the framework's design but is not independently validated in the present study. Finally, the framework is evaluated at a single model scale and without parameter updates, and whether the same structural signals remain discriminative under fine-tuning or at larger scales is an open question.

Table 14: Mapping from the twenty draft-stage desire labels to five functional impulse categories. The category names are used as row headers in Table 6.

Impulse category	**Murray subclasses**	**Functional meaning**
confront	dominance, counteraction, defendance	Directly opposing the threat source
avoid	harmavoidance, blameavoidance, rejection	Withdrawing from or suppressing the threat
submit	abasement, succorance, affiliation, nurturance	Yielding to or moving toward the threat source
justify	recognition, cognizance, achievement, exposition, exhibition	Proving oneself through cognition or accomplishment
control	order, autonomy, construction	Establishing order or independence as a coping route

Table 15: Distribution of turns across the five impulse categories ($n = 3{,}226$).

Impulse category	Turns	Share
justify	993	30.8%
confront	820	25.4%
submit	751	23.3%
control	486	15.1%
avoid	176	5.5%
total	3226	100%

Table 16: Cell sample sizes for the impulse-by-defense grid in Table 6.

Impulse	psychotic	immature	neurotic	mature
confront	301	277	83	159
avoid	45	67	**15**	49
submit	232	239	57	223
justify	347	334	64	248
control	142	162	40	142

Table 17: Kruskal–Wallis tests on rewrite polarity across the four defense tiers, per impulse category.

Impulse	n	H	p
confront	820	61.01	3.7×10^{-13}
avoid	176	9.44	0.024
submit	751	33.87	2.3×10^{-7}
justify	993	53.61	1.4×10^{-11}
control	486	13.98	0.0029

Table 18: Bonferroni-corrected pairwise Mann–Whitney U tests between neurotic and psychotic / immature defense tiers, per impulse category. Effect size $r = Z/\sqrt{N}$. Correction factor $n_{\text{tests}} = 10$.

Impulse	Comparison	n_a	n_b	median diff	r	p_{Bonf}
confront	Neu vs. Psy	83	301	+0.381	0.314	1.4×10^{-8}
confront	Neu vs. Imm	83	277	+0.439	0.313	5.8×10^{-8}
avoid	Neu vs. Psy	15	45	+1.646	0.336	0.095
avoid	Neu vs. Imm	15	67	+1.507	0.299	0.070
submit	Neu vs. Psy	57	232	+0.247	0.227	0.0011
submit	Neu vs. Imm	57	239	+0.135	0.121	0.367
justify	Neu vs. Psy	64	347	+0.129	0.133	0.070
justify	Neu vs. Imm	64	334	+0.178	0.189	0.0017
control	Neu vs. Psy	40	142	+0.004	0.038	1.000
control	Neu vs. Imm	40	162	+0.032	0.092	1.000