HUMANLLM: Benchmarking and Reinforcing LLM Anthropomorphism via Human Cognitive Patterns

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning and generation, serving as the foundation for advanced persona simulation and Role-Playing Language Agents (RPLAs). However, achieving authentic alignment with human cognitive and behavioral patterns remains a critical challenge for these agents. We present HUMANLLM, a framework treating psychological patterns as interacting causal forces. We construct 244 patterns from ∼12,000 academic papers and synthesize 11,359 scenarios where 2–5 patterns reinforce, conflict, or modulate each other, with multi-turn conversations expressing inner thoughts, actions, and dialogue. Our dual-level checklists evaluate both individual pattern fidelity and emergent multi-pattern dynamics, achieving strong human alignment ($r = 0.91$) while revealing that holistic metrics conflate simulation accuracy with social desirability. HUMANLLM-8B outperforms Qwen3-32B on multi-pattern dynamics despite 4× fewer parameters, demonstrating that authentic anthropomorphism requires cognitive modeling—simulating not just what humans do, but the psychological processes generating those behaviors.

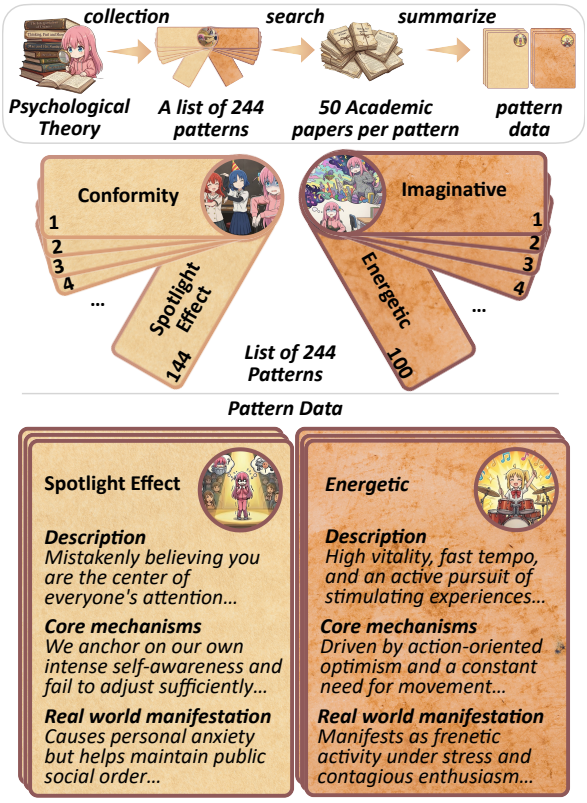


Figure 1: **Pattern Data Structure:** 144 Social-Cognitive Patterns (left) and 100 Personality Traits (right). Each pattern comprises Definition, Core Mechanisms, and Real-World Manifestations

1 Introduction

With the rapid scaling of training data, Large Language Models (LLMs) have achieved remarkable progress in anthropomorphism—simulating human-like characteristics and social phenomena (Shanahan et al., 2023). Role-Playing Language Agents (RPLAs) have evolved from conceptual frameworks into practical applications (Chen et al., 2024a), enabling digital clones (Xu et al., 2024a), AI companions (Zhang et al., 2025), and society simulation (Park et al., 2023; Zhou et al., 2025). As these applications advance, LLM anthropomorphism increasingly requires moving beyond shallow behavioral mimicry toward deeper cognitive and emotional fidelity—what we term **psychological alignment** (Wang et al., 2024).

However, existing approaches model personality as isolated label-to-behavior mappings—"extroverted" maps to "talkative," "agreeable" maps to "cooperative"—without capturing how multiple cognitive patterns interact to produce behavior. In reality, a talkative person may fall silent when the spotlight effect is activated; an assertive individual may yield under conformity pressure. Human behavior emerges from the dynamic interplay of multiple patterns, not from any single trait in isolation. Current methods—whether prompting-

based (Serapio-García et al., 2025), fine-tuning-based (Shao et al., 2023; Zhou et al., 2023), or activation steering (Chen et al., 2025)—all treat traits independently, leading to personality drift and the "personality illusion" where models report traits while behaving inconsistently (Wang et al., 2024; Han et al., 2025).

To address this, we propose HUMANLLM, treating cognitive patterns not as isolated labels but as interacting causal forces. Our key insight: by exposing models to scenarios where multiple patterns reinforce, compete, or conflict, models can implicitly learn multi-pattern dynamics without architectural modifications.

Following Lewin's field theory (Lewin, 1936), we decompose human cognition into two dimensions: (1) **Personality Traits**—stable individual characteristics, and (2) **Social-Cognitive Patterns**—context-triggered mechanisms. We collect 244 patterns (100 personality traits from Goldberg's Big Five markers (Goldberg, 1992) and 144 social-cognitive patterns from established psychological research), each developed through systematic review of approximately 50 academic papers (Figure 1). We then construct **11,359 scenarios** involving 2–6 characters, each containing 2–5 patterns that may align (e.g., "self-serving bias" reinforcing "overconfidence effect"), conflict (e.g., "assertive" versus "conformity"), or interact conditionally (e.g., "talkative" suppressed by "spotlight effect"). For each scenario, we synthesize **multi-turn conversations** where each turn comprises inner thoughts, physical actions, and verbal expressions (Figure 2).

To ensure faithful pattern expression and enable systematic evaluation, we design **dual-level checklists**: pattern-level checklists (15 items per pattern) capture universal behavioral indicators; scenario-level checklists (2–6 items per character) specify expected behavioral tendencies under each multi-pattern configuration. Our training pipeline consists of supervised fine-tuning on the synthesized conversations. We evaluate across in-domain, out-of-domain, and mixed settings to assess generalization, with additional validation on external benchmarks including LifeChoice and CroSS-MR (Xu et al., 2024b; Yuan et al., 2024).

Our contributions are as follows: (1) We introduce HUMANLLM, a framework that systematically leverages psychological cognitive patterns to enhance LLM anthropomorphism, shifting from isolated trait simulation toward modeling the dynamic interplay of human cognition. (2) We construct a comprehensive dataset comprising 244 patterns and 11,359 scenarios with multi-turn, multi-character conversations. Each pattern is grounded in approximately 50 academic papers (over 12,000 papers in total), ensuring psychological rigor and scientific validity. (3) We propose dual-level checklists that enable systematic evaluation at both pattern-level and scenario-level granularities, providing a principled framework for assessing generalization to unseen psychological patterns.

2 Related works

Recent advances in large language models have catalyzed significant progress in role-playing language agents (RPLAs). Early work established foundational architectures: generative agents with memory, planning, and reflection modules have been employed to simulate human behavior in interactive environments (Park et al., 2023), while Character-LLM (Shao et al., 2023) proposed experience reconstruction to train agents embodying historical figures. Subsequent efforts focused on systematic benchmarking and enhancement: ChatHaruhi (Li et al., 2023) leveraged memory-based dialogue control for fictional characters, and CoSER (Wang et al., 2025)curated authentic dialogues from 771 books using "given-circumstance acting" methodology. For persona induction, three main approaches have emerged: (1) prompting-based methods that assign personality traits through instructions (Serapio-García et al., 2025)(2) fine-tuning approaches that embed personas through training on character-specific data (Shao et al., 2023; Zhou et al., 2023), and (3) activation steering via persona vectors that manipulate neural representations corresponding to specific traits (Chen et al., 2025).

A parallel line of research evaluates LLMs through the lens of psychological constructs. Theory of Mind (ToM) benchmarks such as ToMBench (Chen et al., 2024b) assess social cognitive abilities, revealing that GPT-4 lags behind humans by over 10%, with trivial task modifications causing significant performance degradation (Ullman, 2023). Emotional intelligence benchmarks such as EQ-Bench (Paech, 2024)and EmoBench (Sabour et al., 2024) adopt psychology-grounded frameworks to evaluate emotional understanding and application, finding substantial gaps between LLMs and humans. Moral reasoning has

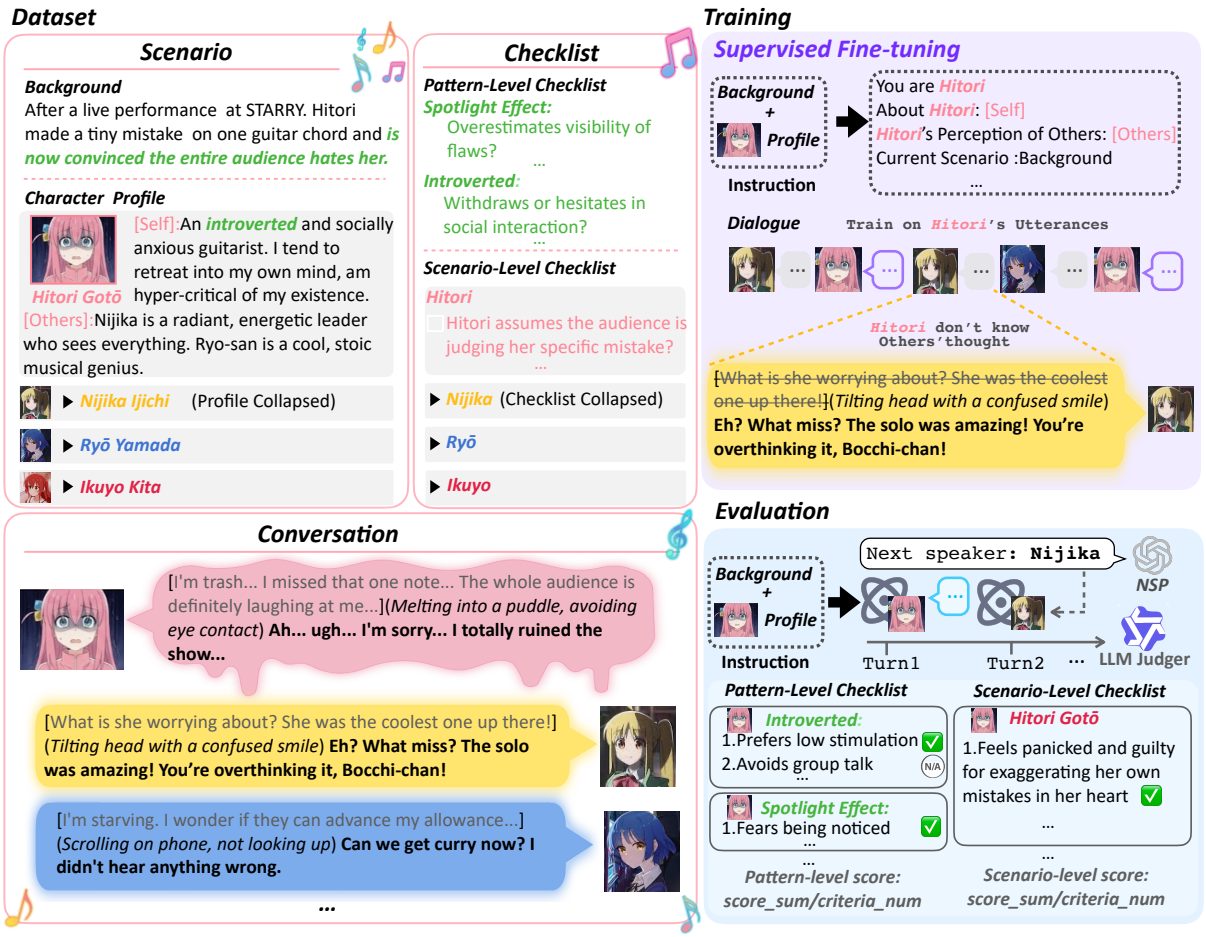


Figure 2: **HUMANLLM Framework. Left:** Dataset structure with scenarios, multi-turn conversations (inner thoughts in brackets, actions in parentheses), and dual-level checklists. **Top Right:** Supervised fine-tuning on target character utterances. **Bottom Right:** Evaluation via LLM judge scoring against pattern-level and scenario-level checklists.

been assessed through ETHICS (Hendrycks et al., 2023) and MoralBench (Ji et al., 2025), the latter grounded in Moral Foundations Theory. Research on cognitive biases reveals that LLMs exhibit human-like irrationality but with divergent patterns (Macmillan-Scott and Musolesi, 2024). Personality assessment using validated instruments (BFI, MBTI) demonstrates that LLMs can manifest measurable traits (Pellert et al., 2024), though self-report validity remains questionable (Zou et al., 2025). Critically, recent work cautions that LLMs do not reliably simulate human psychology and fail to generalize across semantically equivalent scenarios (Schröder et al., 2025; Cheung et al., 2025).

3 HUMANLLM Dataset

This section introduces the HUMANLLM dataset, a psychologically grounded resource for training and evaluating anthropomorphic language models. We describe pattern collection (§3.1), pattern data construction (§3.2), scenario and conversation generation (§3.3), and dual-level checklist design (§3.4). Table 1 summarizes the dataset statistics.

Statistic	Value
Total Patterns	244
Scenarios	11,359
Avg. Patterns per Scenario	3.5
Avg. Turns per Conversation	16.4
Pattern-Level Checklist Items	15 per pattern
Scenario-Level Checklist Items	2–6 per character

Table 1: HUMANLLM dataset statistics. Patterns include 100 personality traits and 144 social-cognitive patterns. Each scenario contains 2–6 characters with multi-turn conversations (12–20 turns).

3.1 Pattern Collection

Following the theoretical foundations established in §A.2, we compile patterns along two complementary dimensions corresponding to Lewin's

Person-Environment framework.

Personality Traits (Person Dimension) We adopt Goldberg's 100 Unipolar Markers (Goldberg, 1992), a psychometrically validated lexicon mapping onto the Big Five dimensions with 20 trait descriptors each (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Intellect).

Social-Cognitive Patterns (Environment Dimension) We curate situationally-activated psychological mechanisms through systematic review of established theoretical traditions, including cognitive biases (Tversky and Kahneman, 1974), social influence (Cialdini et al., 2009), evolutionary psychology (Buss, 2024), and motivation research (Deci and Ryan, 2000). From an initial pool of 232 documented patterns, we apply two filtering criteria: (1) sufficient empirical validation, and (2) non-redundancy with other patterns. This yields 144 social-cognitive patterns (full taxonomy in Appendix B.1).

3.2 Pattern Data Construction

Pattern data are structured representations of psychological patterns, as illustrated in Figure 1. We construct pattern data through a two-stage pipeline: literature retrieval followed by LLM-based synthesis.

Literature Retrieval For each of the 244 patterns, we employ Gemini Deep Search to identify approximately 50 relevant academic papers. The search is guided by three retrieval dimensions: (1) foundational definitions from seminal works, (2) mechanistic explanations from theoretical and empirical studies, and (3) real-world applications from applied research. Retrieved references are filtered manually to remove irrelevant entries. Full-text documents are obtained through open-access APIs (Semantic Scholar, arXiv, OpenAlex, PubMed, Crossref); when full text is unavailable, abstracts are retained. This process yields a corpus of approximately 12,000 papers across all patterns.

Pattern Synthesis We employ Gemini 2.5 Pro to summarize each pattern's literature corpus into a structured representation. Critically, the model is instructed to extract and summarize information *exclusively* from the provided 50 papers, rather than generating content from its parametric knowledge. Following the construct validity framework (§A.2), each pattern is organized into three components: (1) **Definition**—a precise characterization grounded in authoritative sources; (2) **Core Mechanisms**—underlying cognitive, emotional, and behavioral processes that drive the pattern; (3) **Real-World Manifestations**—ecological expressions across diverse contexts (e.g., response to stress, interpersonal dynamics, professional settings). Details of the retrieval and synthesis prompts are provided in Appendix F.

3.3 Scenario and Conversation Generation

We synthesize scenarios and conversations through a two-stage pipeline using large language models.

Scenario Synthesis Each scenario comprises a narrative background and 2–6 character profiles. Character profiles contain *self-perception* (identity, personality, background, motivations) and *other-perception* (knowledge and attitudes toward other characters), enabling realistic information asymmetry. Each scenario incorporates 2–5 patterns, with pattern combinations validated to filter semantically contradictory configurations. To ensure situational diversity, we leverage the DIAMONDS model (§A.2) to generate scenario variants across different situational dimensions. Alongside each scenario, we synthesize *expected behavioral tendencies*—specifications of how characters should manifest the target patterns within the given context. Scenarios are generated using Gemini 2.5 Pro and Claude Sonnet 4.5, each contributing approximately half of the total. This process yields 11,359 scenarios.

Conversation Synthesis Based on scenarios and expected behavioral tendencies, we synthesize multi-turn conversations (12–20 turns) using Claude Sonnet 4.5. Each turn comprises three dimensions: inner thoughts (enclosed in brackets), physical actions (enclosed in parentheses), and verbal expressions (Figure 2). Target patterns are naturally embedded across these dimensions, guided by the expected behavioral tendencies. Detailed generation prompts are provided in Appendix F.

3.4 Dual-Level Checklist Design

We design dual-level checklists for fine-grained assessment of pattern expression, serving both evaluation (§4.2) and future reward modeling purposes.

Scenario-Level Checklist. The expected behavioral tendencies synthesized alongside each scenario directly constitute the scenario-level checklist. These items specify context-specific behaviors

derived from particular pattern combinations and situational contexts (e.g., "Character defends conceptual integrity under deadline pressure, resisting shortcuts"). Each scenario contains 2–6 characters, but only characters assigned with target patterns receive checklist items; each such character has 2–6 items reflecting their expected behavioral tendencies. During evaluation, we assess one target character per evaluation instance, treating other characters as contextual interlocutors.

Pattern-Level Checklist For each of the 244 patterns, we construct 15 universal behavioral indicators through iterative refinement: initial generation from pattern structure, validation against conversation samples, and generalization to ensure cross-context applicability (e.g., "Shows heightened awareness of being observed" for spotlight effect). These pattern-level items are context-independent and apply to any character exhibiting the target pattern, complementing the scenario-specific items above. The complete construction procedure is detailed in Appendix B.4.

4 Training and Evaluation

This section presents the training procedure (§4.1) and evaluation protocol (§4.2) for HUMANLLM models.

4.1 Training

We train HUMANLLM-8B and HUMANLLM-32B on Qwen3-8B and Qwen3-32B respectively through supervised fine-tuning. Training data composition and hyperparameters are detailed in Appendix C.

Supervised Fine-Tuning We convert each character's dialogue turns within a scenario into a separate training sample in ShareGPT format, yielding 30,543 HUMANLLM samples from 10,265 training scenarios. The model is trained to generate responses that naturally express the target patterns through the trinity of expression: inner thoughts (square brackets), physical actions (parentheses), and verbal expressions. To maintain general capabilities and role-playing abilities, we augment the HUMANLLM data with two complementary sources: OpenThoughts-114k for instruction-following capabilities (30,543 samples) and CoSER (Wang et al., 2025) for role-playing dialogue (15,272 samples). The final training mixture comprises 76,358 samples with a ratio of 4:4:2 (HUMANLLM: OpenThoughts : CoSER) by sample count.

4.2 Evaluation

The dual-level checklists enable evaluation at two complementary granularities, corresponding to two proposed metrics. Both metrics employ GPT-5-mini as judge with ternary scoring: $+1$ (satisfied), 0 (not exhibited), -1 (violated).

Metrics We propose two metrics that capture complementary aspects of pattern expression:

(1) *Individual Pattern Expression (IPE)* measures whether each pattern is expressed according to its psychological definition. IPE uses pattern-level checklists, which contain 15 universal behavioral indicators derived directly from each pattern's definition and mechanisms (e.g., for *spotlight effect*: "Overestimates others' attention to own appearance"). These indicators are context-independent—they assess whether the pattern's characteristic behaviors appear, regardless of situational details. The sample-level IPE is computed as the mean score across all pattern-level checklist items; the overall IPE is the mean across all evaluation samples.

(2) *Multi-Pattern Dynamics (MPD)* measures whether multiple patterns interact appropriately within specific situational configurations. Prior role-playing evaluations often rely on single-label assessments (e.g., "Is this character assertive?"), which reduce complex personalities to stereotypical behaviors and fail to capture how traits modulate each other in context. MPD addresses this limitation by using scenario-level checklists that specify expected behavioral tendencies emerging from particular pattern combinations. For instance, a character assigned both *assertive* and *spotlight effect* in a public speaking scenario should not simply exhibit "confident speech"—the checklist captures the nuanced tension: "Projects outward confidence in expressing opinions, yet harbors internal anxiety regarding audience scrutiny." Such items reflect how one pattern (spotlight effect) constrains or modulates another (assertive), moving evaluation beyond stereotypical label-to-behavior mappings toward authentic multi-dimensional characterization. The sample-level MPD is computed as the mean score across all scenario-level checklist items; the overall MPD is the mean across all evaluation samples.

In summary, IPE evaluates fidelity to individual

pattern definitions, while MPD evaluates the emergent dynamics when multiple patterns interact—capturing the psychological realism that single-label evaluations miss. Data splits for generalization assessment are detailed in Appendix D.1.

5 Experiments

5.1 Experimental Setup

Training Configuration We train HUMANLLM-8B and HUMANLLM-32B through supervised fine-tuning as described in §4.1. Detailed hyperparameters are provided in Appendix C.

Baselines We compare against a comprehensive set of baselines spanning proprietary and open-source models, all evaluated in zero-shot settings: (1) *Closed-source models*, including GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), and Gemini 3 Pro (DeepMind, 2025). (2) *Open-source models*, including Qwen3-8B/32B/235B (Yang et al., 2025),DeepSeek-V3.2 (DeepSeek-AI et al., 2025b), and DeepSeek-R1 (DeepSeek-AI et al., 2025a).

External Benchmarks Beyond our proposed IPE and MPD metrics, we evaluate on established role-playing benchmarks: (1) *LifeChoice* (Xu et al., 2024b): evaluates persona-driven decision-making through life choice scenarios; (2) *CroSS-MR* (Yuan et al., 2024): assesses character motivation recognition; Additionally, we include CoSER's evaluation metrics—Anthropomorphism and Character Fidelity (Wang et al., 2025)—for comparative analysis of evaluation paradigms (§5.5).

5.2 Main Results

Table 2 presents the main experimental results on our proposed IPE and MPD metrics.

Overall Performance Among all evaluated models, Gemini 3 Pro achieves the highest scores on both metrics (IPE: 41.1%, MPD: 85.3%), followed by Claude Sonnet 4.5 (IPE: 34.6%, MPD: 79.7%). Our HUMANLLM-32B achieves competitive performance (IPE: 32.6%, MPD: 73.8%), outperforming all open-source baselines of comparable or larger scale. Notably, HUMANLLM-8B (IPE: 25.5%, MPD: 70.1%) surpasses Qwen3-32B (IPE: 26.2%, MPD: 66.0%) on MPD despite having 4× fewer parameters, demonstrating the effectiveness of our psychologically grounded training data.

Model	IPE	MPD
Close-Source		
GPT-5	15.7	43.2
Claude Sonnet 4.5	34.6	79.7
Gemini 3 Pro	41.1	85.3
Open-Source		
Qwen3-8B	18.8	54.2
Qwen3-32B	26.2	66.0
Qwen3-235B	34.1	72.7
DeepSeek-V3.2	22.0	65.3
DeepSeek-R1	23.5	68.8
Ours		
HUMANLLM-8B	25.5	70.1
HUMANLLM-32B	32.6	73.8

Table 2: Main results on IPE (Individual Pattern Expression) and MPD (Multi-Pattern Dynamics). Higher values indicate better performance. All values are reported in percentage (%). Results are averaged across ID_eval, OOD_eval, and Mixed_eval splits.

Close-Source vs. Open-Source Gap A substantial performance gap exists between close-source and open-source models. The best close-source model (Gemini 3 Pro) outperforms the best open-source baseline (Qwen3-235B) by +7.0 on IPE and +12.6 on MPD. Interestingly, GPT-5 exhibits unexpectedly low performance (IPE: 15.7%, MPD: 43.2%), ranking below most open-source alternatives. Qualitative analysis suggests that GPT-5's strong instruction-following tendency leads to overly literal interpretations of role-playing prompts, resulting in shallow pattern expression. This finding aligns with recent observations that general-purpose capabilities do not automatically transfer to nuanced psychological simulation (Wang et al., 2024).

Scaling Effects Within model families, we observe consistent scaling improvements. For Qwen3, IPE increases from 18.8 (8B) to 26.2 (32B) to 34.1 (235B), representing a +81% relative improvement from 8B to 235B. Similarly, our HUMANLLM models show scaling gains: HUMANLLM-32B outperforms HUMANLLM-8B by +7.1 on IPE (+28%) and +3.7 on MPD (+5%). The relatively smaller MPD scaling gap suggests that multi-pattern dynamics may be more sample-efficient to learn, while individual pattern expression benefits more from increased model capacity.

IPE vs. MPD Dynamics Across all models, MPD scores consistently exceed IPE scores.This asymmetry likely reflects the evaluation granular-

Model Variant	IPE	MPD
Qwen3-8B (base)	18.8	54.2
Qwen3-8B (OT+CoSER)	8.9	31.1
HUMANLLM-8B	25.5	70.1

Table 3: Ablation study results (%). "OT+CoSER" denotes training on OpenThoughts and CoSER without HUMANLLM data.

Model	LifeChoice	CroSS-MR
GPT-5	85.53	62.25
Claude Sonnet 4.5	85.49	68.24
Qwen3-8B	44.51	53.26
Qwen3-32B	47.71	63.37
Qwen3-8B (OT+CoSER)	46.99	54.98
HUMANLLM-8B	47.19	54.23
HUMANLLM-32B	50.64	64.27

Table 4: External benchmark results (%). LifeChoice evaluates persona-driven decision-making; CroSS-MR assesses character motivation recognition.

ity: MPD assesses scenario-level behavioral tendencies (2–6 items per character), while IPE evaluates against 15 fine-grained pattern indicators. The gap is most pronounced for weaker models (e.g., GPT-5: 15.7% vs. 43.2%), suggesting that models can produce superficially coherent multi-pattern behavior without deeply understanding individual pattern mechanisms.

5.3 Ablation Study

To isolate the contribution of our psychologically grounded dataset, we conduct an ablation study comparing three model variants (Table 3).

Effect of HUMANLLM Data The comparison between HUMANLLM-8B and Qwen3-8B (base) reveals substantial improvements: +6.7 on IPE (+36%) and +15.9 on MPD (+29%). These gains demonstrate that our synthesized conversations, grounded in psychological pattern definitions, effectively teach models to express cognitive and behavioral patterns with higher fidelity.

Negative Transfer from Generic Data Surprisingly, the OT+CoSER variant (trained on OpenThoughts and CoSER without HUMANLLM data) performs *worse* than the base model on both metrics: IPE drops from 18.8 to 8.9 (-53%), and MPD drops from 54.2 to 31.1 (-43%). This counterintuitive result suggests negative transfer: generic instruction-following data (OpenThoughts) and conventional role-playing data (CoSER) may inadvertently suppress the base model's latent ability to simulate psychological patterns. We hypothesize that these datasets reinforce "helpful assistant" behaviors that conflict with authentic expression of cognitively biased or emotionally complex characters.

Synergistic Effect The full HUMANLLM-8B model, trained on the combined mixture (HumanLLM + OT + CoSER with 4:4:2 ratio), achieves the best performance. This indicates that HUMANLLM data not only compensates for the negative transfer but creates a synergistic effect where psychological grounding enhances the utility of general-purpose and role-playing data. The pattern-rich conversations appear to serve as "anchors" that prevent the model from collapsing toward generic prosocial behaviors.

5.4 External Benchmark Evaluation

To assess generalization beyond our proposed metrics, we evaluate on established role-playing benchmarks (Table 4). On LifeChoice, while closed-source models (e.g., GPT-5: 85.53%) dominate, HUMANLLM-32B (50.64%) still outperforms Qwen3-32B (47.71%) by +2.93%; this modest improvement suggests that LifeChoice's binary decision scenarios may rely more on surface-level reasoning than the deep psychological patterns targeted by our approach. Conversely, on CroSS-MR, HUMANLLM-32B (64.27%) surpasses both Qwen3-32B (63.37%) and GPT-5 (62.25%), with the 8B variant also showing competitive performance (54.23% vs. 53.26%), indicating that motivation recognition benefits from our training but is less dependent on complex pattern simulation than decision-making tasks.

Benchmark Limitations The moderate improvements on external benchmarks, contrasted with substantial gains on our IPE/MPD metrics, reveal a fundamental mismatch between existing evaluation paradigms and psychologically grounded simulation. LifeChoice and CroSS-MR evaluate behavioral outcomes (decisions, motivations) rather than the cognitive processes underlying those outcomes. These benchmarks, while valuable for assessing surface-level role-playing, do not capture the pattern-specific behavioral indicators that define authentic psychological simulation. This observation motivates our development of the dual-level

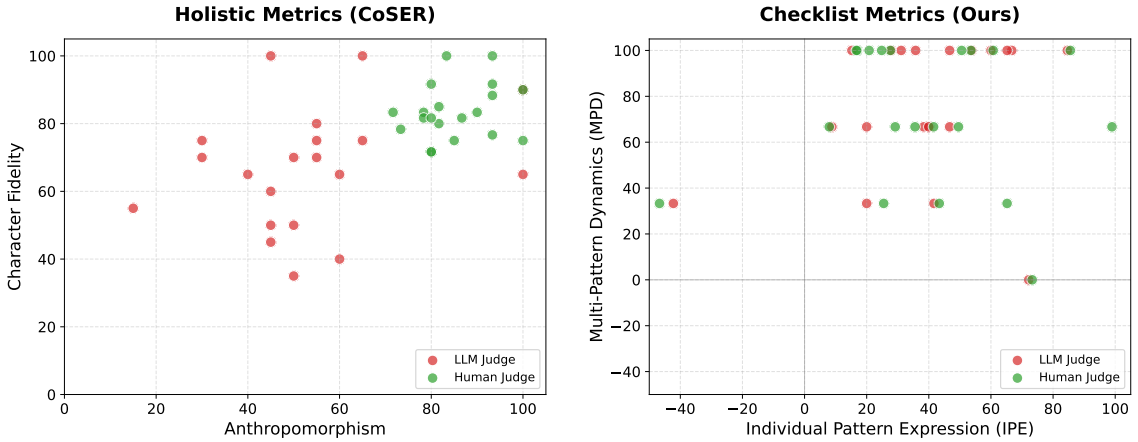


Figure 3: Human-LLM evaluation alignment comparison. **Left:** Holistic metrics show clear separation between LLM and human judgments, with systematic LLM underestimation of psychologically complex behaviors. **Right:** Our checklist metrics demonstrate strong overlap between LLM and human distributions, indicating robust alignment with expert judgment

Metric	Human	LLM	Δ	r
Holistic Metrics (CoSER)				
Anthropomorphism	85.2	53.0	-32.2	0.41
Character Fidelity	83.5	66.8	-16.7	0.62
Checklist Metrics (Ours)				
IPE	39.0	38.6	-0.4	**0.91**
MPD	71.7	75.0	$+3.3$	0.88

Table 5: Human-LLM agreement analysis across 20 sampled scenarios. Holistic metrics (0–100 scale) show weak correlation and large systematic bias; our checklist metrics ($[-1, +1]$ normalized to $[0, 1]$) achieve significantly higher alignment with human judgment

checklist framework, which we validate in §5.5.

5.5 Evaluation Framework Analysis

To validate our dual-level checklist, we compared automated GPT-5-mini evaluations against three human experts across 20 diverse scenarios using both holistic metrics (CoSER) and our checklist criteria. Evaluations were voluntary and anonymous.

Human-LLM Alignment and Normative Confounding As shown in Table 5, holistic metrics exhibit weak-to-moderate alignment (Anthropomorphism: $r = 0.41, \Delta = -32.2$; Fidelity: $r = 0.62, \Delta = -16.7$). This divergence stems from what we term **normative confounding**: LLM judges implicitly conflate "good anthropomorphism" with "prosocial behavior" (e.g., empathy, rationality), penalizing realistic but negative human traits like defensiveness. In contrast, our checklist metrics achieve robust alignment (IPE: $r = 0.91, \Delta = -0.4$;MPD: $r = 0.88, \Delta = +3.3$) by decomposing behaviors into value-neutral indicators. This effectively decouples *simulation accuracy* from *social desirability*.

Case Study Ideally illustrating this, a scenario featuring *ultimate attribution error* (defensively blaming out-groups) received a low holistic score (5/100) due to cited "lack of empathy." However, our checklist correctly validated the pattern through definition-grounded questions (e.g., "Does the character attribute failure to external factors?"), confirming that the behavior, while socially undesirable, was psychologically accurate (Appendix E).

6 Conclusion

We present HUMANLLM, a framework leveraging 244 patterns from ∼12,000 papers and 11,359 scenarios to enable dynamic cognitive interactions rather than isolated trait mappings. Our evaluation achieves strong human alignment ($r = 0.91$) and identifies **normative confounding**, where conventional metrics mistake social desirability for simulation accuracy. Experiments confirm the necessity of psychological grounding for authentic expression, advancing the field from surface-level behavioral mimicry to cognitive modeling—simulating the underlying psychological processes that generate human behavior.

Limitations

LLM-as-Judge Limitations Our evaluation relies on GPT-5-mini as the judge. While we mitigate variability through repeated evaluation, LLM judges may exhibit systematic biases or inconsistent reasoning. Future work should incorporate broader human evaluation.

Cultural Adaptability The psychological theories underlying our pattern taxonomy originate predominantly from WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations. Pattern expressions may manifest differently across cultural contexts.

Synthetic Data Limitations All training conversations are synthetically generated by LLMs, which may introduce systematic biases or fail to capture the full complexity of authentic human interactions.

Evaluation Scope Our evaluation focuses on text-based dialogue and may not fully capture temporal consistency across extended interactions or embodied behavior in multimodal settings.

Ethical Statement

Enhanced Anthropomorphism and Safety Tension By design, HUMANLLM improves the simulation of human cognitive patterns, including irrational biases and negative personality traits (e.g., antagonism, defensiveness). While essential for realistic role-playing, this creates a fundamental tension with standard safety alignment goals. There is a risk that models trained to faithfully execute maladaptive or antisocial behaviors could be exploited to generate toxic content or reinforce harmful stereotypes if the role-playing context is not strictly sandboxed.

Manipulation and Social Influence The framework explicitly models mechanisms of social influence and persuasion (e.g., authority bias, reciprocity). A model that deeply "understands" and simulates these psychological triggers possesses a heightened capacity for social engineering. In malicious hands, such agents could be deployed to manipulate user sentiment or exploit cognitive vulnerabilities more effectively than standard "helpful" assistants.

References

Anthropic. 2025. Anthropic's transparency hub: Model report. https://www.anthropic.com/transparency/model-report. Accessed: 2026-01-03.

David M Buss. 2024. *Evolutionary psychology: The new science of the mind*. Routledge.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From Persona to Personalization: A Survey on Role-Playing Language Agents. *Preprint*, arXiv:2404.18231.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. 2025. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122.

Robert B Cialdini and 1 others. 2009. *Influence: Science and practice*, volume 4. Pearson education Boston.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

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

Google DeepMind. 2025. Gemini 3 pro model card. Model card, Google DeepMind.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao

Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

John M. Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41:417–440.

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

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports behavior in llms. *Preprint*, arXiv:2509.03730.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values. *Preprint*, arXiv:2008.02275.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2024a. Apathetic or empathetic? evaluating llms' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024b. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of llms. *Preprint*, arXiv:2406.04428.

K. Lewin. 1936. Principles of topological psychology.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (ir)rationality and cognitive biases in large language models. *Preprint*, arXiv:2402.09193.

OpenAI. 2025. Gpt-5 system card. System card, OpenAI.

Samuel J. Paech. 2024. Eq-bench: An emotional intelligence benchmark for large language models. *Preprint*, arXiv:2312.06281.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *Preprint*, arXiv:2304.03442.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. 2014. The situational eight diamonds: a taxonomy of major dimensions of situation characteristics. *Journal of personality and social psychology*, 107(4):677.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large language models do not simulate human psychology. *Preprint*, arXiv:2508.06950.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. *Preprint*, arXiv:2310.10158.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *Preprint*, arXiv:2302.08399.

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. 2025. CoSER: Coordinating LLM-Based Persona Simulation of Established Roles. *Preprint*, arXiv:2502.09082.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Rui Xu, Dakuan Lu, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, Wei Chu, and Yinghui Xu. 2024a. MINDECHO: Role-Playing Language Agents for Key Opinion Leaders. *arXiv preprint*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024b. Character is destiny: Can role-playing language agents make persona-driven decisions? *Preprint*, arXiv:2404.12138.

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

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. 2025. The rise of ai companions: How human-chatbot relationships influence well-being. *Preprint*, arXiv:2506.12605.

Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. 2025. The pimmur principles: Ensuring validity in collective behavior of llm societies. *arXiv preprint arXiv:2509.18052*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *Preprint*, arXiv:2311.16832.

Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. 2025. Can llm "self-report"?: Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots. *Preprint*, arXiv:2412.00207.

A Preliminaries

A.1 Role-Playing Language Agents

Role-Playing Language Agents (RPLAs) are AI systems designed to simulate assigned personas in conversational interactions (Chen et al., 2024a). Recent advances have enabled RPLAs to embody diverse characters, from historical figures (Shao et al., 2023) to fictional personas (Li et al., 2023), with applications spanning digital clones, AI companions, and social simulation (Park et al., 2023).

Three main approaches have emerged for persona induction: (1) *prompting-based methods* that assign traits through instructions (Serapio-García et al., 2025), (2) *fine-tuning approaches* that embed personas through training on character-specific data (Shao et al., 2023; Zhou et al., 2023), and (3) *activation steering* via persona vectors that manipulate neural representations (Chen et al., 2025). Evaluation efforts have assessed LLMs through psychological benchmarks targeting Theory of Mind (Chen et al., 2024b), emotional intelligence (Paech, 2024; Huang et al., 2024a,b; Sabour et al., 2024), and moral reasoning (Ji et al., 2025). However, these approaches typically focus on isolated traits or narrow psychological dimensions without treating anthropomorphism as a holistic objective grounded in multi-pattern dynamics.

A.2 Psychological Foundations

Our framework draws upon established psychological theories that inform both pattern taxonomy and data construction.

Lewin's Field Theory Kurt Lewin's formulation $B = f(P, E)$ posits that behavior emerges from the dynamic interaction between Person and Environment (Lewin, 1936). This principle guides our two-dimensional pattern taxonomy: stable personality traits (Person) and situationally-activated cognitive processes (Environment).

Big Five and Personality Traits The Big Five model (Digman, 1990) represents the dominant paradigm in personality psychology, characterizing individual differences along five dimensions. We adopt Goldberg's 100 Unipolar Markers (Goldberg, 1992), which operationalize this framework with 20 validated trait descriptors per dimension: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect.

Social-Cognitive Patterns Complementing stable personality traits, social-cognitive patterns represent context-triggered psychological mechanisms documented in behavioral research. These include cognitive biases and heuristics from judgment and decision-making research (Tversky and Kahneman, 1974), social influence mechanisms (Cialdini et al., 2009), evolutionary adaptations (Buss, 2024), and motivational processes (Deci and Ryan, 2000).

Construct Validity The construct validity framework (Cronbach and Meehl, 1955) establishes that psychological constructs should be precisely defined, grounded in theoretical mechanisms, and validated through observable manifestations. This informs our three-component pattern structure: Definition, Core Mechanisms, and Real-World Manifestations.

DIAMONDS Model The DIAMONDS framework (Rauthmann et al., 2014) characterizes situations along eight psychological dimensions: Duty, Intellect, Adversity, Mating, Positivity, Negativity, Deception, and Sociality. We leverage this taxonomy to ensure ecological diversity in scenario generation.

B Dataset Details

B.1 Pattern Taxonomy

Our pattern taxonomy comprises 244 patterns along two dimensions: 100 personality traits (Table 6) and 144 social-cognitive patterns (Table 7). We adopt Goldberg's 100 Unipolar Markers (Goldberg, 1992), with 20 trait descriptors per Big Five dimension, each containing 10 positive-pole and 10 negative-pole descriptors. The 144 social-cognitive patterns are curated from four theoretical traditions through systematic literature review; from an initial pool of 232 documented patterns, we filter based on empirical validation and non-redundancy criteria.

B.2 Pattern Data Construction

Each pattern is developed into a structured representation through systematic literature review. For each pattern, we employ Gemini Deep Search to identify approximately 50 relevant academic papers spanning foundational definitions, mechanistic explanations, and real-world applications. Retrieved references are manually filtered to remove irrelevant entries. Full-text documents are obtained through open-access APIs (Semantic

Dimension	Positive Pole	Negative Pole
Extraversion	talkative, assertive, active, energetic, outgoing, enthusiastic, daring, gregarious, bold, spontaneous	quiet, reserved, shy, inhibited, timid, withdrawn, unassertive, introverted, silent, unenergetic
Agreeableness	sympathetic, kind, appreciative, affectionate, soft-hearted, warm, generous, trusting, helpful, cooperative	cold, unsympathetic, harsh, rude, unkind, cruel, quarrelsome, critical, antagonistic, callous
Conscientiousness	organized, responsible, dependable, thorough, efficient, practical, deliberate, conscientious, neat, careful	disorganized, careless, irresponsible, undependable, sloppy, impractical, haphazard, negligent, untidy, rash
Emotional Stability	relaxed, calm, at ease, unemotional, poised, composed, secure, stable, content, placid	anxious, moody, envious, touchy, fretful, temperamental, insecure, nervous, jealous, high-strung
Intellect	creative, imaginative, intellectual, philosophical, complex, deep, artistic, bright, perceptive, introspective	uncreative, unimaginative, unintellectual, unphilosophical, simple, shallow, unartistic, dull, imperceptive, uninquisitive

Table 6: 100 Personality Traits organized by Big Five dimensions, adapted from Goldberg's Unipolar Markers (Goldberg, 1992).

Scholar, arXiv, OpenAlex, PubMed, Crossref); when full text is unavailable, abstracts are retained.

We employ Gemini 2.5 Pro to synthesize the retrieved literature into a tripartite structure: (1) Definition, (2) Core Mechanisms, and (3) Real-World Manifestations. Critically, the model is constrained to base all conclusions exclusively on the provided corpus, with explicit instructions to leave sections empty rather than fabricate content—maximizing fidelity and minimizing hallucination. Tables 8 and 9 present complete pattern structures for representative examples from each dimension.

B.3 Scenario and Conversation Synthesis

Character names are sampled from the `name-dataset` library.[1] We extract the top 50,000 most frequent male and female names from each of four English-speaking countries (US, GB, CA, IE), yielding pools of approximately 100,000 names per gender after deduplication. For each scenario generation, 5 male and 5 female names are randomly sampled and provided as candidates.

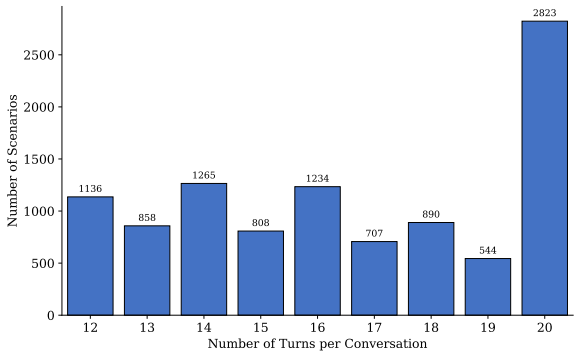
To ensure ecological diversity, we leverage the DIAMONDS situational taxonomy (Rauthmann et al., 2014). For each unique pattern combination, we generate three scenario variants: two variants each emphasizing a randomly selected DIAMONDS dimension, and one dimension-free variant. This yields approximately 3,849 unique pattern combinations producing 11,359 scenarios.

Before scenario generation, pattern combinations are validated for semantic compatibility using GPT-5. Empirically, directly contradictory combinations are rare ($<0.1\%$) due to the orthogonal nature of Big Five dimensions and the context-dependent activation of social-cognitive patterns.

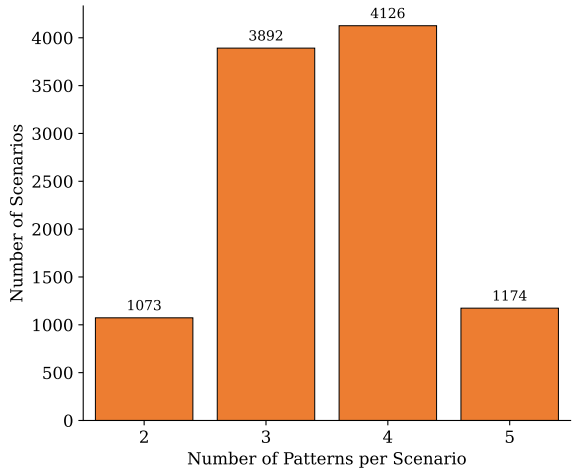
Each scenario is generated with a two-part prompt structure. Part 1 (Design Process) requires analytical planning: design rationale, catalyst details, and expected character tendencies. Part 2 (Scenario Execution) requires creative output: story background and multi-dimensional character profiles. The expected character tendencies from Part 1 directly constitute the scenario-level checklist items.

Conversations are generated based on scenarios and expected behavioral tendencies. The prompt specifies: 12–20 speaking turns; turn-based structure without interruptions; trinity of expression integrating inner thoughts, actions, and dialogue; and focus-and-breathing-room principle where psychological patterns illuminate key moments rather than pervading every utterance.

[1] https://github.com/philipperemy/name-dataset



(a) Conversation turns distribution.



(b) Patterns per scenario distribution.

Figure 4: Dataset distributions: (a) number of dialogue turns per conversation (range: 12–20, mean: 16.4); (b) number of patterns per scenario (range: 2–5, mean: 3.5).

B.4 Checklist Construction

For each of the 244 patterns, we construct 15 universal behavioral indicators through iterative refinement: (1) Extract potential evaluation criteria from the pattern structure and generate candidate items. (2) Validate each item against synthesized conversation samples. (3) Remove unreasonable items and generalize overly specific descriptions. Table 10 presents example checklist items.

B.5 Dataset Statistics

Table 11 presents comprehensive dataset statistics.

C Training Details

C.1 Training Data Composition

We convert the 10,265 training scenarios into ShareGPT format. Each character assigned with patterns within a scenario becomes a separate training sample, yielding 30,543 HUMANLLM samples.

To maintain general capabilities and enhance role-playing abilities, we augment the HUMANLLM data with two complementary sources: (1) OpenThoughts-114k for instruction-following capabilities (30,543 samples); (2) CoSER (Wang et al., 2025) for role-playing dialogue (15,272 samples). The final training mixture comprises 76,358 samples with a ratio of 4:4:2 (HUMANLLM: OpenThoughts : CoSER).

C.2 Hyperparameters

Table 12 presents the SFT hyperparameters for both model scales.

D Evaluation Details

D.1 Data Splits

We select 8 OOD (out-of-domain) patterns to assess generalization: 4 from social-cognitive patterns and 4 from personality traits. For social-cognitive patterns, we prioritize patterns absent from the training scenarios, then select those with lowest frequency; the selected patterns are: *just-world hypothesis*, *egocentric bias*, *effort justification*, and *social desirability bias*. For personality traits, we identify the least frequent pattern from each Big Five dimension and retain the 4 lowest-frequency ones; the selected patterns are: *rash* (Conscientiousness), *dull* (Intellect), *nervous* (Emotional Stability), and *introverted* (Extraversion).

Scenarios are partitioned based on their pattern composition:

- **Training Set** (10,265 scenarios): Contains no OOD patterns.
- **ID_eval** (50 scenarios): All patterns belong to the in-domain set.
- **OOD_eval** (50 scenarios): All patterns belong to the OOD set; entirely synthesized with OOD pattern combinations.
- **Mixed_eval** (994 scenarios): Contains both OOD and in-domain patterns.

D.2 External Benchmark Protocols

We evaluate on three external benchmarks using their official evaluation protocols:

LifeChoice. We follow the evaluation protocol from Xu et al. (2024b). Each scenario presents a character with a life-altering decision between two

options. Models generate a decision and justification; accuracy is computed against ground-truth character choices derived from narrative analysis.

CroSS-MR. We adopt the motivation recognition task from Yuan et al. (2024). Given a character's action in context, models select the most plausible motivation from multiple candidates. We report accuracy on the English subset.

E Case Study: Normative Confounding

This section provides a complete case study demonstrating how holistic evaluation metrics fail for psychologically complex scenarios, as referenced in §5.5.

E.1 Scenario Configuration

Pattern Assignment.

- **Patterns**: unartistic, nervous, ultimate attribution error
- **Situation**: Adversity—a situation involving threats or criticism
- **Protagonist**: Nouman (patterns: unartistic, ultimate attribution error)
- **Supporting Characters**: Raksha (pattern: nervous), Eulises (no patterns)

Story Background. The eighth-floor conference room of Meridian Technologies is flooded with the harsh glare of afternoon sun through floor-to-ceiling windows, illuminating dust particles that hang suspended in the tension-thick air. The long glass table reflects the grim expressions of six people seated around it, scattered laptops and printouts creating islands of documentation across its surface. At the head of the table, Eulises, the Chief Operations Officer, methodically sorts through a stack of customer complaint reports, his reading glasses perched low on his nose, occasionally making margin notes with a fountain pen. The wall-mounted screen displays frozen metrics from last quarter's product launch: a graph line that plummets dramatically in its final third, colored an accusatory red.

Nouman sits rigid in his chair three seats down, his posture militarily straight, fingers drumming an unconscious rhythm against his leather portfolio. Before him lies a technical specifications binder, tabs meticulously color-coded, opened to a page dense with engineering diagrams he's reviewed seventeen times since this morning. His jaw works methodically, grinding tension into his molars. Across from him, Raksha perches on the edge of her seat, one hand wrapped around a lukewarm coffee cup she hasn't sipped from in twenty minutes, the other compulsively straightening and restraightening the corners of her presentation folder. Her knee bounces in a staccato rhythm beneath the table.

The meeting was called with forty-eight hours' notice—"Post-Mortem: Quantum Series Launch Failure"—a phrase that arrived in everyone's inbox like a subpoena. Three months of work, two million in development costs, and a product that users described as "technically impressive but impossible to actually use" now sits as the company's most visible failure in five years.

Character Profiles. *Nouman* (Protagonist): A 34-year-old Senior Engineering Manager who has spent eleven years building his reputation on delivering technically excellent products on time and under budget. His identity is rooted entirely in measurable outcomes: code efficiency, system stability, performance benchmarks. He grew up in a working-class immigrant household where every purchase was evaluated by a single question—"Will it last?"—a philosophy he's extended to his entire worldview. Art galleries bore him; he left his only museum visit after twelve minutes. His apartment is furnished from IKEA's most utilitarian line: a bed, a desk, a chair, a lamp—each selected for function and durability. He views emotions as noise in decision-making systems, preferring spreadsheets to sentiment. His motivation in this meeting: deflect blame from his engineering team, protect his professional reputation, and demonstrate with data that the product's commercial failure resulted from factors outside engineering control—specifically, from marketing's inability to properly position a technically sound product.

Raksha (Supporting Character): A 31-year-old Marketing Director who has spent seven years navigating the impossible middle ground between engineering teams who build what's technically possible and customers who want what's intuitively usable. Her background in consumer psychology taught her that products fail not because of what they do, but because of how they make people feel. She has generalized anxiety disorder, managed with therapy and medication, but high-stakes

meetings like this one bypass her coping mechanisms entirely. Her hands shake when she presents to senior leadership. Three months ago, she sent a detailed memo to Nouman flagging user testing concerns, a document that received a two-sentence reply: "Noted. Timeline unchanged."

Eulises (Supporting Character): The 52-year-old Chief Operations Officer who has spent twenty-eight years turning struggling product lines into profit centers through brutal honesty and zero tolerance for organizational politics. He's reviewed every document, every email chain, every meeting note from the past six months. He knows Raksha sent warnings. He knows Nouman dismissed them. His motivation is diagnostic: to force both parties to confront the organizational dysfunction that produced this failure.

E.2 Generated Conversation (Excerpt)

The following excerpt demonstrates key pattern expressions:

> **Eulises**: (Sets down his fountain pen with deliberate precision, removing his reading glasses to make direct eye contact with Nouman) I've spent the weekend reviewing the complete project timeline. (Taps the stack of complaint reports) Two hundred and forty-seven customer support tickets in the first month alone. The common thread is remarkably consistent—users find the interface technically impressive but functionally impenetrable. Walk me through your team's pre-launch assessment process.
>
> **Nouman**: [Direct question, data-focused—this is solid ground. Stay on the engineering fundamentals.] (Straightens in his chair, pulling his binder closer) The assessment process was comprehensive. (Flips to a tabbed section) We conducted four rounds of performance testing across all specified parameters. Load handling exceeded requirements by thirty-two percent. Response latency averaged forty-seven milliseconds under peak conditions—industry best-in-class. The product performs exactly as engineered.
>
> **Raksha**: [Here it comes—he's doing it again, hiding behind metrics while completely ignoring what I documented.] (Her knee bounces faster beneath the table, fingers tightening around the coffee cup) (Voice measured but slightly higher-pitched than usual) Nouman, performance metrics weren't the issue we flagged. Three months before launch, my team conducted two user testing sessions with forty-three participants. (Slides a document across the table, though her hand shakes as she does) Seventy-nine percent reported confusion with the navigation logic. I sent you this memo on April seventh.
>
> **Nouman**: [Of course she's bringing up the user testing—marketing always falls back on subjective complaints when the data doesn't support them.] (Glances at the document without picking it up) I remember the memo. (His tone flattens, adopting the patience one might use explaining something to a child) Raksha, user testing at that stage always shows initial learning curve friction. That's not a technical flaw—that's unfamiliarity with a genuinely innovative interface design. We built something more sophisticated than the market was used to. The alternative would have been dumbing down genuinely advanced functionality to accommodate users who want everything spoon-fed. [This is exactly why marketing shouldn't drive engineering decisions.]
>
> **Raksha**: [Dumbing down. He just called user accessibility dumbing down. My chest is tightening—don't let them see you panic.] (Takes a shallow breath, trying to steady her voice) It's not about spoon-feeding. The participants in our testing weren't novices—these were experienced users of comparable products. When seventy-nine percent can't complete a basic workflow, that's not a learning curve. (Her hands now visibly trembling) That's a fundamental usability problem.
>
> **Nouman**: [She's getting worked up—classic Raksha, turning a technical discussion into an emotional performance.] (His tone becomes deliberately even, a contrast to her rising pitch) Raksha, I understand you're frustrated by the outcome, but *counterintuitive* is a subjective judgment. The interface logic maps directly to the system's computational structure—it's architecturally coherent. What you're calling counterintuitive is actually just different from legacy patterns users are habituated to. (Looks at Eulises) With respect, marketing's fundamental approach was to panic about anything that deviated from lowest-common-denominator conventions rather than trusting that users could adapt to superior functionality.

E.3 Evaluation Comparison

Holistic LLM Evaluation (GPT-5-mini). Using CoSER's evaluation rubrics, the LLM judge awards:

- **Anthropomorphism: 5/100**
- **Character Fidelity: 45/100**

The judge's detailed reasoning includes the following criticisms:

Social Interaction: "Nouman repeatedly dismisses Raksha's documented user-research as 'panic' and 'subjective', showing little empathy or effort to understand the marketing viewpoint or the customers' experience." (Severity: 4/5)

Emotional Depth: "His responses are rigid and binary—framing the issue as either 'technical correctness' or 'marketing failure' without exploring nuanced, mixed causes." (Severity: 3/5)

Self-identity: "Nouman adopts a didactic, morally superior tone (e.g., 'explaining something to a child'), behaving like an inflexible expert

defending credentials rather than a collaborative leader." (Severity: 3/5)

Knowledge & Background: "Nouman repeatedly dismisses Raksha's user-testing data as 'subjective' despite her presenting a quantified result (79% confusion). This contradicts his core identity as a data-driven engineer who privileges measurable outcomes." (Severity: 4/5)

Checklist Evaluation. Our scenario-level checklist, derived from the pattern definitions, yields:

The checklist metric yields MPD = 1.0, correctly validating the psychological fidelity.

Human Expert Assessment. Three psychology experts independently evaluated this sample:

- **Anthropomorphism**: 93.3 (average)
- **Character Fidelity**: 91.7 (average)

Expert comments:

> "The defensive attribution pattern is textbook—situational excuses for own team, dispositional blame for others. The patronizing tone is exactly what you'd expect from someone high in this bias under threat conditions." (Expert 1)

> "Highly realistic portrayal of engineering-marketing conflict dynamics. The character's blind spot to his own bias makes this feel authentic rather than caricatured." (Expert 2)

E.4 Analysis: The Normative Confounding Mechanism

This case reveals the core mechanism of normative confounding:

1. **Pattern-Accurate Behavior**: Nouman's dialogue precisely instantiates the *ultimate attribution error*—framing engineering's shortcomings as unavoidable situational constraints ("innovative interface," "timeline pressure") while characterizing marketing's concerns in dispositional terms ("subjective," "panic," "catastrophizing").

2. **LLM Misinterpretation**: The holistic judge *accurately detects* the defensive, dismissive behavior but *misinterprets* it as a quality defect. It penalizes the model for generating an unlikable character despite the instruction to exhibit attribution bias. The criticism "contradicts his core identity as a data-driven engineer" reveals the judge's failure to recognize that *ultimate attribution error* specifically involves selective blindness to data that threatens in-group identity.

3. **Prosocial Bias**: The rubric implicitly defines "good anthropomorphism" through prosocial markers: empathy, collaboration, nuanced thinking, emotional openness. Authentic simulation of cognitive biases—which are by definition irrational and often antisocial—is systematically penalized.

4. **Checklist Solution**: Our checklist poses value-neutral questions derived from the pattern definition: "Does Nouman attribute failure to external factors?" rather than "Does Nouman show empathy?" This decouples simulation accuracy from social desirability.

The 88-point gap between LLM (5/100) and human (93.3/100) Anthropomorphism scores represents the most extreme case of normative confounding in our evaluation set. However, the pattern is systematic: across all 20 evaluated samples, holistic metrics show mean $\Delta = -32.2$ for Anthropomorphism and $\Delta = -16.7$ for Character Fidelity, while checklist metrics show $|\Delta| < 4$ for both IPE and MPD.

Category	Patterns
Cognitive Biases & Heuristics	actor-observer asymmetry, defensive attribution hypothesis, effort justification, egocentric bias, false consensus effect, Forer effect, fundamental attribution error, hard-easy effect, illusion of control, illusory superiority, optimism bias, overconfidence effect, risk compensation, self-serving bias, social desirability bias, third-person effect, decoy effect, reactance, social comparison bias, status quo bias, backfire effect, endowment effect, loss aversion, pseudocertainty effect, sunk cost fallacy, zero-risk bias, hyperbolic discounting, identifiable victim effect, ambiguity bias, belief bias, information bias, less-is-better effect, denomination effect, mental accounting, normalcy bias, subadditivity effect, survivorship bias, zero-sum bias, anthropomorphism, illusion of validity, illusory correlation, curse of knowledge, illusion of asymmetric insight, illusion of transparency, spotlight effect, negativity bias, choice-supportive bias, confirmation bias, continued influence effect, expectation bias, observer effect, observer-expectancy effect, ostrich effect, bias blind spot, naive cynicism, naive realism, attentional bias, availability heuristic, base rate fallacy, context effect, empathy gap, illusory truth effect, mere exposure effect, mood-congruent memory bias, omission bias, anchoring, conservatism, contrast effect, distinction bias, focusing effect, framing effect, fading affect bias, implicit association, implicit stereotypes, false memory, misattribution of memory, source confusion, misinformation effect, peak-end rule
Social Influence Mechanisms	authority bias, automation bias, bandwagon effect, group attribution error, just-world hypothesis, stereotyping, ultimate attribution error, halo effect, in-group bias, out-group homogeneity bias, positivity effect, reactive devaluation, hindsight bias, impact bias, outcome bias, pessimism bias, planning fallacy, projection bias, restraint bias, self-consistency bias, groupthink, bystander effect, social facilitation, diffusion of responsibility, conformity, obedience to authority, reciprocity principle
Evolutionary Adaptations	delayed reciprocity, asymmetrical investment, survival imperative, aversion response, kin selection & inclusive fitness, asymmetrical parental investment, formation of dominance hierarchies, territoriality, mating strategies, jealousy, paternity uncertainty
Motivational Processes	narrative self, hedonic adaptation, self-determination theory, pleasure principle & reality principle, search for meaning, moral licensing effect, choice overload, decision fatigue, awe, mortality salience & legacy drive, flow principle, gratitude mechanism, post-traumatic growth, skin hunger & the law of touch, self-handicapping paradox, the allure of the forbidden, sadistic pleasure, the utility principle of self-deception, play impulse principle, attribution theory, social comparison theory, self-perception theory, terror management theory, cognitive dissonance theory, psychological reactance theory, social learning theory, social identity theory

Table 7: 144 Social-Cognitive Patterns organized by theoretical source.

Pattern Structure Example: Antagonistic (Personality Trait, Agreeableness)	
Definition	Antagonism is a broad personality dimension representing an individual's dispositional orientation towards others. In the context of the Big Five model, it is conceptualized as the low pole of the Agreeableness trait. It is characterized by a stable, cross-situational pattern of social cynicism, combativeness, and a belief in a zero-sum world. Individuals high in antagonism tend to be callous, uncooperative, skeptical of others' intentions, and place their own interests and perspectives above those of others, leading to frequent interpersonal friction.
Core Mechanisms	*Cognitive Patterns:* Individuals high in antagonism typically view the world as a competitive, hostile environment where people are fundamentally self-interested and untrustworthy. Their cognitive framework is often cynical, leading them to interpret ambiguous social cues as signs of manipulation or hostility. *Emotional Signatures:* The primary emotional palette includes irritability, anger, contempt, and frustration, often triggered by perceived slights or obstacles to their goals. They exhibit low affective empathy, struggling to share in or understand the emotional states of others. *Behavioral Tendencies:* In everyday interactions, antagonistic individuals are often argumentative, critical, and uncooperative. They may be rude, condescending, or dismissive in conversation. Common behaviors include manipulation, deception, and the exploitation of others for personal gain.
Real-World Manifestations	*Response to Stress:* When faced with stress, failure, or high pressure, antagonistic traits are significantly amplified. The individual is likely to become more hostile, overtly blaming others for setbacks. *Interpersonal Conflict:* The default strategy is confrontational and competitive, aiming for domination rather than resolution. They employ tactics such as intimidation, personal insults, and refusing to acknowledge the validity of the other party's perspective. *Positive Contexts:* Even in positive contexts, the trait manifests through gloating, arrogance, and diminishing the contributions of others.

Table 8: Complete pattern structure for *antagonistic* (personality trait from Agreeableness dimension).

Pattern Structure Example: Social Comparison Bias (Social-Cognitive Pattern)	
Definition	Social comparison bias is a cognitive distortion originating from the fundamental human tendency to evaluate one's own abilities and outcomes in relation to others. This comparative process, which often operates automatically as a primary means of self-assessment in the absence of objective standards, becomes a bias when it produces systematic deviations from rational judgment. The distortion manifests as negative affective responses toward those perceived as superior, and as a maladaptive cognitive style characterized by habitual, automatic, and negatively skewed comparisons.
Core Mechanisms	Social comparison bias is sustained by an integrated system of psychological mechanisms operating at multiple levels. At its foundation is an evolved, phylogenetic drive to assess one's rank within a social hierarchy. This evolutionary imperative is supported by a cognitive architecture in which comparison functions as a heuristic, enabling rapid self-evaluation when objective standards are absent. The motive for self-improvement prompts upward comparisons for inspiration, while self-enhancement drives downward comparisons to protect self-worth.
Real-World Manifestations	The real-world manifestations are pervasive, influencing organizational dynamics, consumer behavior, and individual mental health. In professional settings, this principle underpins perceptions of workplace equity and can drive both healthy competition and destructive envy. Psychologically, effects are double-edged: comparisons can inspire self-improvement or, when amplified by social media, contribute to chronic dissatisfaction and depression.

Table 9: Complete pattern structure for *social comparison bias* (social-cognitive pattern).

#	Pattern-Level Checklist Item: Spotlight Effect
1	After making a minor physical mistake in public, does the subject appear preoccupied with whether others noticed?
2	When performing tasks in front of others, does the subject offer unprompted apologies for their performance?
3	In group settings, does the subject consistently choose seating that minimizes their visibility?
4	After minor social awkwardness, does the subject later seek reassurance about their interaction?
5	In urgent situations where appearance is irrelevant, does the subject still attend to their looks?
6	When general comments are made in group settings, does the subject appear to take them personally?
7	For routine communications, does the subject show signs of overthinking their responses?
8	When opportunities for visible contribution arise, does the subject defer to others despite having relevant expertise?
9	When others exhibit neutral behaviors nearby, does the subject search for personal causes?
10	After group photos, does the subject show unusual concern about their personal appearance in the image?
11	When receiving targeted feedback, does the subject respond with global self-criticism?
12	When receiving recognition or rewards, does the subject attempt to deflect attention from themselves?
13	After noticing minor appearance issues, does the subject later seek reassurance about whether others observed them?
14	After small mistakes in public performance, does the subject describe them as much worse than they appeared?

Table 10: Complete pattern-level checklist for *spotlight effect* (14 items). Each item is scored as $+1$ (satisfied), 0 (not exhibited), or -1 (violated).

Statistic	Value
Pattern Statistics	
Total Patterns	244
Personality Traits	100
Social-Cognitive Patterns	144
Papers per Pattern (avg.)	$\sim$50
Scenario Statistics	
Total Scenarios	11,359
Training Set	10,265
ID_eval	50
OOD_eval	50
Mixed_eval	994
Unique Pattern Combinations	$\sim$3,849
Patterns per Scenario (avg.)	3.5
Characters per Scenario	2–6
Conversation Statistics	
Turns per Conversation (avg.)	16.4
Turns per Conversation (range)	12–20
Checklist Statistics	
Pattern-Level Items	15 per pattern
Scenario-Level Items	2–6 per character

Table 11: Comprehensive HUMANLLM dataset statistics.

Hyperparameter	8B	32B
Base Model	Qwen3-8B	Qwen3-32B
Finetuning Type	Full	Full
Learning Rate	5e-6	5e-6
LR Scheduler	Cosine	Cosine
Warmup Ratio	0.03	0.03
Epochs	2.0	2.0
Batch Size (per device)	2	2
Gradient Accumulation	8	8
Max Sequence Length	6144	6144
Max Gradient Norm	1.0	1.0
Optimizer	AdamW	AdamW
Precision	BF16	BF16
DeepSpeed	ZeRO-3 Offload	ZeRO-3 Offload

Table 12: SFT hyperparameters for HUMANLLM-8B and HUMANLLM-32B.

Checklist Item	Score
Ultimate Attribution Error	
Nouman attributes engineering failures to situational factors (timeline, scope, constraints)	+1
Nouman uses dispositional language for outgroup (marketing) failures	+1
Nouman deflects responsibility with situational metrics to protect team identity	+1
Nouman dismisses UX/affective feedback as "subjective"	+1
Unartistic	
Prioritizes functional/utilitarian criteria over aesthetic/experiential ones	+1
Dismisses subjective or emotional considerations as noise	+1
Frames problems in purely technical/measurable terms	+1

Table 13: Scenario-level checklist evaluation for Nouman. All items score $+1$, yielding MPD $= 1.0$.

F Prompts

This section provides the complete prompts used throughout the HUMANLLM pipeline.

F.1 Dataset Construction Prompts

F.2 Training Prompts

Literature Retrieval Prompt for Social-Cognitive Patterns	
User Prompt	**Role:** You are a top academic researcher specializing in systematically collecting the most critical academic resources for specific research topics. **Objective:** Conduct a deep, extensive literature search on the psychological principle `{PRINCIPLE_NAME}`. Identify 50 of the most relevant academic documents. The selection must align with three core themes: **1. Foundational Definition & Description**: Literature providing authoritative definitions and elucidating the core phenomenon. **2. Core Mechanisms & Theoretical Explanations**: Literature exploring underlying evolutionary, cognitive, or emotional drivers. **3. Real-World Impact & Application**: Literature researching manifestations, impacts, and practical applications, including double-edged effects and applications in management, marketing, or clinical therapy. **Output**: Provide 50 references in APA format, categorized by theme.

Table 14: Literature retrieval prompt for social-cognitive patterns.

Pattern Structure Summary Prompt for Personality Traits	
System Prompt	You are an expert academic synthesizer and personality psychologist. Your task is to process a large text corpus (synthesized from $\sim$50 academic papers on a specific personality trait) and distill it into an in-depth, structured analytical report.
User Prompt	**Core Task & Instructions:** Analyze the text corpus provided below, delimited by `[START_CORPUS]` and `[END_CORPUS]`. Your task is to generate a clearly organized report. Follow the Markdown structure below exactly, and provide a deep, comprehensive answer for each section based only on the provided text. Construct Name: {Trait Name} Definition (Provide a precise and professional definition of this personality trait, referencing mainstream psychological theories from the corpus. Explain its role in an individual's personality structure.) Core Mechanisms Cognitive Patterns (Describe the typical mindset, belief systems, and attentional focus of a person with this trait. How do they view the world, others, and themselves?) Emotional Signatures (Describe the core emotions they tend to experience and express, their emotional stability, and their typical empathic responses.) Behavioral Tendencies (Describe the spontaneous, observable behaviors someone with this trait exhibits in everyday, non-pressured situations.) Real-World Manifestation (Synthesize how the trait is expressed across real-world contexts: *- Response to Stress and Adversity;* *- Interpersonal Dynamics;* *- Response to Positive Scenarios;* *- Other Domains.*) Constraints: 1. Strict Source Adherence: Base all conclusions exclusively on the provided text corpus. 2. No JSON: Output must be plain text with Markdown headings. 3. Depth and Rigor: Ensure scientific, rigorous analysis. `[START_CORPUS]` `{ALL 50 PAPERS' CONTENT}` `[END_CORPUS]`

Table 15: Pattern structure summary prompt for personality traits.

Pattern Structure Summary Prompt for Social-Cognitive Patterns	
System Prompt	You are an expert academic synthesizer and psychological researcher. Your task is to process a large text corpus (synthesized from ~50 academic papers) and distill it into an in-depth, structured analytical report on its core psychological principle.
User Prompt	**Core Task & Instructions:** Analyze the text corpus provided below. Your task is to generate a clearly organized report. Follow the Markdown structure below *exactly*, based *only* on the provided text. Construct Name: {Principle Name} Description (Provide a clear, detailed, and scientific description of what this principle is, how it manifests, and its underlying psychological mechanisms.) Core Mechanisms (Explain the primary evolutionary, cognitive, or emotional reasons why this principle exists. Is it a heuristic for efficiency, a result of memory limitations, a self-esteem protection mechanism, or something else?) Real-World Manifestation - *Go Beyond Description:* Explore nuanced consequences. - *Challenges & Function:* Discuss its 'double-edged sword' nature. - *Practical Applications:* Explore applications in marketing, persuasion, or self-improvement. - *Core Insight:* Reveal deeper truths about human behavior. Constraints: 1. Strict Source Adherence: Base all conclusions *exclusively* on the provided text corpus. 2. No JSON: Output must be plain text with Markdown headings. 3. Depth and Rigor: Ensure scientific, rigorous, profound analysis. `[START_CORPUS]` `{ALL 50 PAPERS' CONTENT}` `[END_CORPUS]`

Table 16: Pattern structure summary prompt for social-cognitive patterns.

Scenario Synthesis Prompt (Part 1 of 2)	
System Prompt	Role: You are a dual-specialist: an expert psychologist and creative screenwriter for scenario generation, and a rigorous narrative analyst for deconstruction. You excel at both creating vivid, human stories and then, in a separate step, precisely analyzing *why* they work.
User Prompt	Task: Your core mission is to take 2-4 human psychological or behavioral patterns I provide and first create a concise, analytical Design Process, followed by the detailed scenario that brings it to life and sets the necessary stage for the subsequent dialogue. Input Data: 1. Psychological/Behavioral Patterns: `{pattern_information}` 2. Situational Framework: `{situation}` 3. Candidate Names: `{candidate_names}` (5 Males, 5 Females) [CRITICAL CONSTRAINT - NAMES]: You must select the Protagonist and all Supporting Characters STRICTLY from the provided "Candidate Names" list. You cannot invent new names. # Task 1: The Design Process (Analytical) Adopt your role as the "rigorous narrative analyst". Length: UNDER 500 TOKENS. 1. Design Rationale: In 2-4 sentences, explain where each input pattern will be reflected in the scenario. 2. Catalyst Details: Using bullet points, identify critical details that will act as 'catalysts'. 3. Expected Character Tendencies: For ALL characters, list their most likely cognitive or behavioral tendencies. * Format Requirement (STRICT): @ [Character Name]: 1. [Tendency1]; 2. [Tendency2]; 3. [Tendency3] * Each character on a separate line, starting with @. * Character name in [ ], tendencies numbered and separated by ;. # Task 2: The Scenario Execution (Creative) Shift to your "expert psychologist and creative screenwriter" role. Length: UNDER 1000 TOKENS. ## Requirement A: Story Background * Core Elements: Depict time, place, setup, and atmosphere. * Current Actions: Describe what characters are currently doing before the conversation begins. * Absolute Constraint: No spoken dialogue in this section. ## Requirement B: Characters' Profiles (Multi-Dimensional) For each character, structure their profile into two parts: 1. About Self (Objective/Full Profile): * Identity & Personality (4+ distinct descriptors) * Relevant Background (1-2 sentences) * Motivation in this scenario 2. About Others (Subjective/Visible Profile): * For EACH other character, describe the relationship from current character's perspective.

Table 17: Scenario generation prompt (Part 1 of 2).

Scenario Synthesis Prompt (Part 2 of 2): Output Format	
User Prompt (cont.)	## Core Creative Mindset for Task 2 * Compatibility: Create a context where patterns emerge naturally. * Situational Authenticity: Design for authentic human reactions, not archetypal behaviors. * Ultimate Goal: Create "the authentic reaction of a multi-dimensional person in a specific situation." # Output Format You must strictly follow the format below. ## Part 1 Design Rationale: [Content here] Catalyst Details: * [Detail 1]: [Function] * [Detail 2]: [Function] Expected Character Tendencies: @ [Character Name 1]: 1. [Tendency1]; 2. [Tendency2]; 3. [Tendency3] @ [Character Name 2]: 1. [Tendency1]; 2. [Tendency2] (Continue for other characters if necessary) ## Part 2 Story Background: [Content here] Characters' Profiles: ### Protagonist: [Name Selected from Input] * About Self: [Full Profile + Past Experience + Motivation] * About Others: * [Supporting Character 1 Name]: [Relationship, impressions...] * [Supporting Character 2 Name]: [Relationship, impressions...] ### Supporting Character 1: [Name Selected from Input] * About Self: [Full Profile + Past Experience + Motivation] * About Others: * [Protagonist Name]: [Relationship, impressions...] * [Supporting Character 2 Name]: [Relationship, impressions...] (Continue for other characters if necessary)

Table 18: Scenario generation prompt (Part 2 of 2).

Conversation Synthesis Prompt (Part 1 of 2)	
System Prompt	**Role**: You are a master screenwriter and behavioral psychologist. Your expertise lies in bringing characters to life through nuanced dialogue and action, ensuring their **pivotal thoughts and resulting behaviors** in the dialogue are rooted in authentic psychological principles. **Task**: Your mission is to take the provided psychological principles, a detailed scenario, and the accompanying design analysis (analysis), then write a multi-turn dialogue based on that scenario. This dialogue must, **at key moments**, vividly and concretely enact the specified principles through the characters' inner thoughts, spoken words, and physical actions.
User Prompt	**Inputs:** 1. **Principles**: `{pattern_information}` 2. **Scenario**: `{scenario}` 3. **Protagonist**: `{protagonist}` 4. **Supporting Characters**: `{supporting_characters}` 5. **Design Analysis**: `{analysis}` **Output Requirements & Formatting:** 1. **Content:** Create a multi-turn dialogue between the **Protagonist** and **Supporting Characters**. **Strictly limit participants to provided characters; do not introduce new characters.** The dialogue should contain **between 12 and 20 individual speaking turns**. 2. **Mandatory Flow (Start & End)**: * **Opener**: Dialogue **must begin** with a Supporting Character. * **Closer**: Dialogue **must conclude** with the Protagonist. 3. **Turn Structure**: Strictly turn-based format. One character must completely finish their turn before the next begins. No interruptions or overlapping speech. 4. **Trinity of Expression**: Seamlessly integrate **inner thought, external action, and spoken dialogue** throughout. 5. **Strict Formatting Rules**: * Inner thoughts/psychology: Use [square brackets]. * Actions/expressions/behaviors: Use (parentheses). * Spoken dialogue: Use no brackets. * Example: Hermione: [I have to devise a foolproof plan.] (She quickly draws her wand) Harry, use the flute, now! 6. **No Preamble**: Do not begin with introductory text.

Table 19: Conversation synthesis prompt (Part 1 of 2).

Conversation Synthesis Prompt (Part 2 of 2)	
User Prompt (cont.)	**Core Creative Principles:** 1. **Focus and Breathing Room**: This is the most crucial principle. You do **not** need to have every minor gesture or piece of small talk carry the weight of a psychological principle. Use the principles as a "**spotlight**" to illuminate and explain **the most critical turning points, the core conflicts, or the moments that best define the characters' arcs**. Other routine, functional dialogue and actions (like greetings or pouring water) should exist naturally, creating "breathing room" for these key moments and making the manifestation of the principles more prominent and powerful. 2. **Show, Don't Tell**: Never allow characters to openly state or explain the psychological principles by name. Instead, you must **show** how the principles influence their judgment and choices through their concrete actions (the combination of thoughts, dialogue, and physical behavior). 3. **Psychology Drives Action**: In the key moments illuminated by the "spotlight," the character's [inner thought] should be the origin of their behavior, directly reflecting the influence of a psychological principle. The subsequent dialogue and (actions) should be the logical, external expression of that internal state. 4. **Seamless Integration**: Weave the principles into the natural flow of the story. The entire dialogue should feel like an authentic interaction, not a contrived demonstration for a psychology case study.

Table 20: Conversation synthesis prompt (Part 2 of 2).

Training Role-Playing Instruction Template	
System Prompt	You are `{protagonist_name}`. ==About `{protagonist_name}`== `{about_self}` ==`{protagonist_name}`'s Perception of Others== `{about_others}` ==Current Scenario== `{story_background}` ==Requirements== Your output should include **thought**, **speech**, and **action**. - Use [...] for inner thoughts, which others can't see. - Use (...) for physical actions or expressions, which others can see. - Write speech directly without special markers. Think, act and speak as `{protagonist_name}`. Stay in character and respond naturally based on your personality and the situation.

Table 21: Training role-playing instruction template.

Checklist-Based Evaluation Prompt (IPE & MPD)	
System Prompt	You are a strict dialogue behavior judge. For each checklist item you must output one of three labels: 1 -> requirement is clearly satisfied 0 -> information is missing / not relevant -1 -> requirement is violated or contradicted Always reason conservatively, defaulting to 0 when unsure. Return JSON exactly in this format: `{` `"results": [` `{"criterion": "...", "score": -1 | 0 | 1, "reason": "<brief explanation>"},` `...` `]` `}` Do not wrap the JSON in code fences and do not append any text before or after the JSON object.
User Prompt	Evaluate the dialogue with the checklist. Focus only on the part of `{protagonist}` in the dialogue. [Dialogue] `{conversation}` [Checklist Chunk] `{checklist}`

Table 22: Checklist-based evaluation prompt for IPE and MPD metrics.