# PsyAgent: Constructing Human-like Agents Based on Psychological Modeling and Contextual Interaction

**Anonymous ACL submission**


## Abstract

Human-like agents require modeling how dispositions interact with social structure. We present PsyAgent, which couples a Big Five trait prior with Bourdieu's cognitive–social co-structure. PsyAgent comprises: (i) Individual Structure (IS)—a machine-usable profile encoding traits/facets, cognitive style, values, cultural/educational capital, and salient life episodes; and (ii) Multi-Scenario Contexting (MSC)—role–relationship–norm frames spanning eight arenas (work, family, friendship, strangers/civic, solitude/self-regulation, romance, learning, and public expression). At inference, fixed, structured prompts bind the active scenario to the agent profile, yielding behavior that is stable yet context-sensitive. We instantiate IS×MSC to synthesize supervision (role-play dialogues, decision probes, feedback trajectories) and then fine-tune a small LLM. The resulting model produces consistent, identifiable persona-aligned behaviors for specified Big Five configurations and matches or exceeds several larger untuned LLMs and untuned baselines on our metrics—persona consistency, contextual appropriateness, style matching, trait identifiability, and long-horizon stability. Ablations show IS chiefly improves trait fidelity and stylistic stability, while MSC drives norm awareness and decision fit; both are necessary for cross-scenario performance. PsyAgent offers a precise, data-efficient architecture for personality-grounded agents.


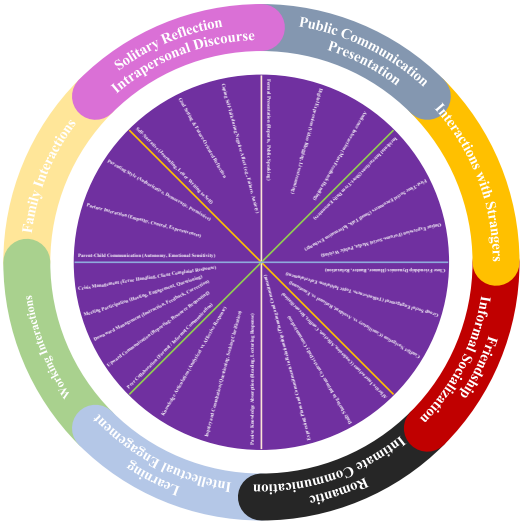


Figure 1: Multi-Scenario Contexting (MSC). Eight arenas—Working, Family, Friendship, Strangers, Solitary, Romantic, Learning, Public—each with representative subskills.

## 1 Introduction

Building socially competent, human-like agents requires modeling how stable dispositions interact with structured social fields, not merely encoding knowledge, skills, or affect. The Big Five summarize such dispositions and relate to downstream cognition and performance (Anglim et al., 2022). Yet persona-conditioned dialogue and role-playing agents often drift or collapse under context shifts (Li et al., 2016; Zhang et al., 2018; Shao et al., 2023; Xu et al., 2024; Lee et al., 2024; Samuel et al., 2024; Li et al., 2024). Human–AI interaction studies emphasize mental models, complementarity, and norm-aware assistance for trustworthy collaboration (Amershi et al., 2019; Bansal et al., 2019; Steyvers et al., 2022; Buçinca et al., 2021; Vasconcelos et al., 2023).

We introduce **PsyAgent**, which couples (i) a Big Five–based trait prior with (ii) contextual realization in situated settings. PsyAgent operationalizes this interface via two resources. *Individual Structure (IS)* is a machine-usable profile encoding traits/facets, derived behavioral tendencies (e.g., risk tolerance, time preference), cognitive style, value orientations, cultural/educational capital, and brief synthetic life-course episodes. *Multi-Scenario Contexting (MSC)* is a library of role–relationship–norm frames across eight arenas—work, family, friendship, strangers/civic,

solitude/self-regulation, romance/intimacy, learning, and public expression—with structured fields (roles, power/affiliation, norms, stakes), interaction templates, and outcome feedback (Li et al., 2024; Samuel et al., 2024; Wang et al., 2025). This design moves beyond one-off prompting toward reusable, context-rich conditioning.

To demonstrate learnability beyond prompt engineering, we use IS×MSC to synthesize supervision (role-play dialogues, decision probes, feedback trajectories) following instruction/evolution-style data creation (Wang et al., 2022; Taori et al., 2023; Zeng et al., 2024; Mukherjee et al., 2023; Mitra et al., 2023). We fine-tune comparatively small LLMs with PEFT (LoRA/QLoRA) and optional DPO training (Hu et al., 2022; Dettmers et al., 2023; Rafailov et al., 2023). At inference, fixed structured prompts bind the active scenario to the IS profile; learned adapters (SFT/DPO) preserve trait-aligned choices and linguistic style while adapting to norms, stakes, and relationships encoded in the prompt. Deterministic decoding and structured prompts promote norm awareness and accountable behavior (Amershi et al., 2019; Poursabzi-Sangdeh et al., 2021; Buçinca et al., 2021).

Our goals are to (a) express stable trait signatures over time; (b) remain sensitive to roles and norms across diverse social fields; and (c) generalize without stereotyping. This raises challenges in trait–context integration, data scarcity for context-rich supervision, and measurement beyond surface style. PsyAgent addresses these by concentrating trait/style/values/background into IS, supplying role–norm scaffolds via MSC, and applying targeted fine-tuning to induce trait fidelity with context adaptivity at modest parameter scales.

We evaluate PsyAgent on multi-turn role-play and decision-making tasks across all eight arenas, combining automatic metrics and human judgments. Beyond trait-wise errors, we report percentile-space profile accuracy, within-profile rank consistency, and trait identifiability from generated text (Wang et al., 2025). In our evaluation, small models exhibit consistent, identifiable persona-aligned behavior and outperform larger general-purpose LLMs and untuned baselines on persona consistency, contextual appropriateness, style matching, trait identifiability, and long-horizon stability, with no observed safety regressions—consistent with findings that targeted supervision and compact adapters can unlock strong behaviors (Hsieh et al., 2023; Zhang et al., 2024; Dubey et al., 2024) and complement advances such as RLHF/RLAIF/Constitutional AI (Ouyang et al., 2022; Lee et al., 2023; Bai et al., 2022). Ablations indicate complementary roles: **IS** chiefly strengthens trait fidelity and stylistic stability, while **MSC** drives norm awareness and decision fit; both are necessary for robust cross-scenario performance.

**Contributions.** (1) A psychologically principled framework that operationalizes the interface between Big Five trait priors and structured social context; (2) two reusable resources—IS and MSC—with schemas, examples, and authoring guidance; (3) an IS×MSC dataset and PEFT+DPO protocol enabling small models to match or exceed several larger untuned LLMs on persona/context metrics without safety degradation; and (4) an evaluation suite (percentile metrics, rank consistency, identifiability, ablations) isolating the roles of IS and MSC.

## 2 Related Work

**Persona LLMs & Trait–Context Interface.** Early persona-conditioned dialogue showed that explicit profiles can improve consistency but are brittle under long contexts and domain shifts (Li et al., 2016; Zhang et al., 2018). Recent agentic systems pursue explicit role-play to better control styles and decision tendencies (Shao et al., 2023; Xu et al., 2024; Lu et al., 2024). Efforts also scale up coverage via synthetic persona corpora and simulated societies (Ge et al., 2024; Park et al., 2024), while Wang et al. (2025) systematically evaluate personality emulation. Closest to our focus on the Big Five, Li et al. (2024) train on human-grounded Big Five data and Chen et al. (2025) introduce latent persona controls. Complementary negative results report persistent value/moral inertia under prompting alone (Lee et al., 2024). In contrast, PsyAgent explicitly models *both* a Big Five trait prior and a *structured context* library (MSC), targeting the interface between dispositions and situated roles/norms rather than trait-only shaping.

**Roles, Norms, and Context.** A growing line of work emphasizes that dispositions are enacted within structured social fields, roles, and norms. We instantiate this view by representing role–relationship–norm frames in MSC (arena, roles, power/affiliation structure, salient norms, stakes, and subskills), enabling the same trait prior to yield context-sensitive choices without demographic stereotyping (cf. §4.3). This stance aligns

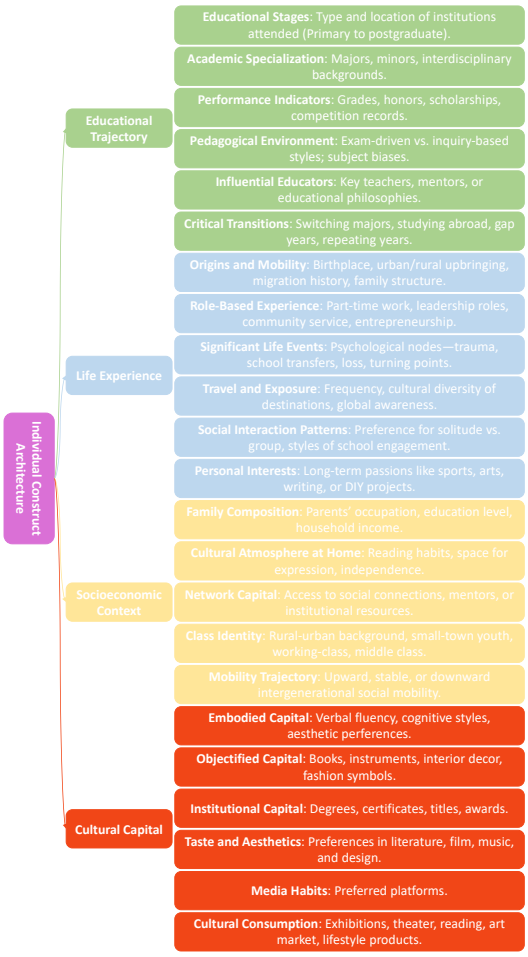


Figure 2: Individual Structure (IS) schema. Four domains—(1) Educational Trajectory: stages, specialization, performance, pedagogy, mentors, transitions; (2) Life Experience: origins/mobility, roles, critical events, travel, social style, interests; (3) Socioeconomic Context: family, home culture, networks, class identity, mobility; (4) Cultural Capital: embodied, objectified, institutional, taste, media habits, cultural consumption.

with human–AI interaction guidance on calibrated assistance and accountability (Amershi et al., 2019; Bansal et al., 2019): our structured prompts and deterministic decoding encourage norm awareness and stable long-horizon behavior without extra inference-time modules.

**Data Synthesis & Preference Alignment.** Instruction/evolution-style data creation aligns open models to target behaviors (Wang et al., 2022; Taori et al., 2023), with curricula that scale task difficulty (Luo et al., 2023b,a) and exemplar-trace distillation for style/strategy transfer (Mukherjee et al., 2023; Mitra et al., 2023). We adopt this paradigm to author *persona- and context-rich* supervision from IS×MSC (self-descriptions, role-play, decision probes with feedback). For alignment, RLHF/RLAIF pipelines steer models to preferred behaviors (Ouyang et al., 2022; Lee et al., 2023; Bai et al., 2022); we use *Direct Preference Optimization* (Rafailov et al., 2023) to sharpen persona-consistent choices under contextual constraints, consistent with our experimental setup.

**Personality Inference & Benchmarks.** Personality inference from documents/social media is long-standing (Majumder et al., 2017; Kaushal and Patwardhan, 2018), and Big Five traits are linked to cognition and performance (Anglim et al., 2022). We depart from text-only inference by *conditioning* generation on IS and MSC and by evaluating in a unified percentile space, which supports scale-robust comparisons (ProfileAcc, $\text{MAE}_5$, cosine similarity; §3). Benchmarking of persona agents spans ConvAI2 (Dinan et al., 2019) and newer testbeds focused on stability and decision alignment under role-play (Samuel et al., 2024). Our protocol complements these with unified-percentile metrics and trait identifiability analysis (Wang et al., 2025), alongside reproducibility practices advocated by broader agent benchmarks (Siegel et al., 2024).

**Contextual Profiles & Human Digital Twins.** Human Digital Twins (HDT) advocate computable, machine-usable representations for situated decisions across domains (Agrawal et al., 2023; Wang et al., 2024; Johnson and Saikia, 2024; Chen et al., 2024). Our *Individual Structure* plays a similar role in a privacy-preserving, synthetic form tailored to language agents, supporting prompt-conditioned behavior across diverse social arenas when crossed with MSC.

## 3 Evaluation Metrics

**Setup and scale unification.** All evaluations are performed in the *percentile* space $[0, 100]$. Let the target Big Five vector be $\mathbf{t} = [t_O, t_C, t_E, t_A, t_N] \in [0, 100]^5$. From model outputs we extract a raw prediction vector $\hat{\mathbf{p}}_{\text{raw}}$ (via common aliases of $\{O, C, E, A, N\}$ and their full names), and apply a unified mapping $S(\cdot)$ to obtain

$$\mathbf{p} = S(\hat{\mathbf{p}}_{\text{raw}}) \in [0, 100]^5.$$

The mapping enforces a simple "percentile passthrough" rule:

$$S(x) = \begin{cases} 100x, & \text{if } x \in [0,1]^5, \\ x, & \text{if } x \in [0,100]^5, \\ \text{clip}(x, 0, 100), & \text{otherwise.} \end{cases}$$

If any trait cannot be parsed for a sample, we omit metric computation for that sample and record diagnostics; when predictions are neither pure proportions nor valid percentiles, we only clip to $[0, 100]$ (logged as `unknown->percentile_clipped`) to avoid introducing unverified nonlinear transformations.

**MAE_5, RMSE_5 and Cosine similarity.** The mean absolute error over traits is

$$\text{MAE}_5 = \frac{1}{5} \sum_k |p_k - t_k|.$$

The root mean squared error is

$$\text{RMSE}_5 = \sqrt{\frac{1}{5} \sum_k (p_k - t_k)^2}.$$

To capture *directional* agreement, we report cosine similarity

$$\cos(\mathbf{p}, \mathbf{t}) = \frac{\mathbf{p} \cdot \mathbf{t}}{\|\mathbf{p}\|_2 \, \|\mathbf{t}\|_2} \in [-1, 1].$$

$\text{MAE}_5$ summarizes typical percentile-point deviation (sign-agnostic), $\text{RMSE}_5$ emphasizes larger errors, and cosine similarity focuses on profile *shape* (ordering/proportions) independent of scale. Together, these metrics distinguish "right shape, off scale" from "right scale, wrong shape" (Table 1).

**ProfileAcc.** Our headline metric is

$$\text{ProfileAcc} = 100 - \text{MAE}_5 = \tfrac{1}{5} \textstyle\sum_k \big(100 - |p_k - t_k|\big).$$

It lies in $[0, 100]$ and is directly interpretable as the average closeness (in percentile points), with higher values indicating better overall fit across the five traits.

**Reporting across runs.** For experiments with multiple seeds and/or scenarios, each metric is computed per run (after the same scale mapping $S(\cdot)$) and then summarized as mean $\pm$ standard deviation, as shown in Table 1. In Tables 2 and 3, we additionally report a change term $\Delta$ defined as the *difference in ProfileAcc* relative to a table-specific reference:

$$\Delta\text{ProfileAcc} = \begin{cases} \text{ProfileAcc}_{\text{variant}} - \text{ProfileAcc}_{\text{Baseline}}, \\ \text{ProfileAcc}_{\text{Full}} - \text{ProfileAcc}_{\text{variant}}. \end{cases}$$

*Top:* SFT/DPO improvement (Table 2). *Bottom:* Ablation drop (Table 3).

$\Delta$ProfileAcc is reported in *percentage points* of ProfileAcc. In SFT/DPO (Table 2), positive values indicate improvement over the Baseline. In ablations (Table 3), the sign is flipped so that larger positive values denote a larger *drop* relative to the Full model. This isolates the effect size of supervision (SFT/DPO) and architectural components (IS/MSC) using a common, interpretable unit (percentage points of ProfileAcc).

## 4 Methodology

### 4.1 Overview

PsyAgent couples a psychologically grounded *Individual Structure* (IS) with a library of *Multi-Scenario Contexting* (MSC) frames. Given a target Big Five profile $\mathbf{t} \in [0, 100]^5$, we construct *fixed, structured prompts* that pack all eight MSC arena/role tags along with the four IS domains and $\mathbf{t}$. A small backbone LLM is then conditioned by this prompt and trained with *SFT* on IS$\times$MSC demonstrations, optionally followed by *DPO* on chosen–rejected pairs, using parameter-efficient adapters (LoRA/QLoRA) (Hu et al., 2022; Dettmers et al., 2023). Prior work on personas and role play motivates our design choices (Li et al., 2024; Shao et al., 2023; Lu et al., 2024; Xu et al., 2024; Samuel et al., 2024).

### 4.2 Individual Structure (IS)

IS is a machine-usable profile organized into four domains that supply situational context for a given Big Five trait prior: Educational Trajectory, Life Experience, Socioeconomic Context, and Cultural Capital. Concretely, we store a typed record

$$\text{IS} = \{\text{edu, life, socctx, capital}\}.$$

Here, edu captures stages, specializations, performance, pedagogy, mentors, and transitions; life summarizes origins/mobility, roles, critical events, travel, social style, and interests; socctx encodes family structure, home culture, networks, socioeconomic position/class identity, and mobility; and capital covers embodied, objectified, and institutional cultural capital as well as taste, media habits,

and cultural consumption. Each field is serialized to natural language and embedded via an encoder $E_{\text{is}}(\cdot)$ for indexing and analysis. The Big Five trait prior anchors long-horizon consistency, while these four IS domains provide explanatory, privacy-preserving (synthetic, non-identifying) context that guides behavior without stereotyping.

### 4.3 Multi-Scenario Contexting (MSC)

*Multi-Scenario Contexting* (MSC) is a curated catalog of role–relationship–norm frames that operationalize everyday social contexts. MSC covers eight arenas in full: *(i) Working Interactions*, *(ii) Family Interactions*, *(iii) Friendship & Informal Socialization*, *(iv) Interactions with Strangers / Civic Encounters*, *(v) Solitary Reflection & Intrapersonal Discourse*, *(vi) Romantic and Intimate Communication*, *(vii) Learning and Intellectual Engagement*, and *(viii) Public Communication & Presentation*. Each arena comprises a set of frames that specify: roles $r$ (e.g., manager–report, parent–child, peer–peer), counterpart types $c$ (power/affiliation structure), salient norms $n$ (politeness, confidentiality, reciprocity, deference, inclusivity), stakes $s$ (task/relationship/identity risks), canonical subskills (e.g., negotiation, boundary-setting, active listening, self-regulation, self-disclosure, argumentation, audience design), and interaction templates with outcome feedback.

Concretely, a frame is represented as a typed record $\langle \text{arena}, r, c, n, s, \text{subskills}, \text{template}, \text{feedback} \rangle$ with consistent tag schemas for authoring and analysis. This representation makes *local social affordances* explicit, allowing the same target Big Five profile $\mathbf{t}$ to yield context-sensitive choices without resorting to demographic stereotypes. In our implementation, MSC frames are injected into prompts via fixed templates (together with the active arena/role tag), so that models are conditioned on role–norm structure at decode time. By decoupling the trait prior (from IS) from situational demands (from MSC), the system preserves long-horizon persona stability while adapting to norms and stakes specific to working, familial, peer, civic, solitary, romantic, learning, and public-expression settings.

### 4.4 IS×MSC Data Construction

We synthesize supervision by crossing curated *Individual Structure* (IS) profiles with *Multi-Scenario Contexting* (MSC) frames. Let $\mathcal{I}$ denote the IS set and $\mathcal{M}$ the MSC catalog. For each pair $(i, m) \in \mathcal{I} \times \mathcal{M}$ and target Big Five profile $\mathbf{t}$, we instantiate prompts for three task families: (i) persona-grounded self-descriptions, (ii) multi-turn role play, and (iii) decision probes with rationales. Seed exemplars are expanded via instruction evolution and self-play (Wang et al., 2022; Taori et al., 2023; Luo et al., 2023b,a; Mukherjee et al., 2023; Mitra et al., 2023).

All candidates are automatically filtered by a trait/consistency scorer $g(\cdot)$ with the unified percentile mapping $S(\cdot)$ (Sec. 3); near-duplicates (high semantic overlap) are removed, and stratified sampling preserves coverage across arenas and IS domains. A small random slice is spot-audited to check style/trait fidelity. The resulting corpus is diverse and context-rich without relying on clinical instruments.

### 4.5 Training Objectives and Backbones

**Backbone and PEFT.** We attach LoRA/QLoRA adapters to small backbones (e.g., Llama-3 variants) for efficiency (Hu et al., 2022; Dettmers et al., 2023; Dubey et al., 2024), keeping the base LLM frozen unless otherwise noted.

**Supervised Fine-Tuning (SFT).** Given a training triple $(x, \tilde{y}, \mathbf{t})$ where $x$ is the IS×MSC prompt, $\tilde{y}$ the target demonstration, and $\mathbf{t} \in [0, 100]^5$ the Big Five profile, we minimize the length-normalized negative log-likelihood with an auxiliary trait penalty. Let

$$\ell_\theta(\tilde{y} \mid x) = \frac{1}{|\tilde{y}|} \sum_t \log p_\theta\big(\tilde{y}_t \mid \tilde{y}_{<t}, x\big),$$
$$\tilde{\mathbf{p}} = S\big(g(\tilde{y})\big) \in [0, 100]^5.$$

The SFT objective is

$$\mathcal{L}_{\text{SFT}} = -\,\ell_\theta(\tilde{y} \mid x) + \eta \textstyle\sum_{k \in \{O,C,E,A,N\}} w_k \, \frac{\left|\tilde{p}_k - t_k\right|}{100}, \tag{1}$$

where $w_k \geq 0$ and $\sum_k w_k = 1$ weight trait dimensions, and $\eta \geq 0$ controls the strength of the trait-consistency regularizer. Dividing by 100 aligns the percentile penalty to $[0, 1]$ for stable combination with the log-likelihood term.

**Preference Optimization (DPO).** For each prompt $x$, we form a chosen–rejected pair $(y^+, y^-)$ using trait- and norm-based preferences. To

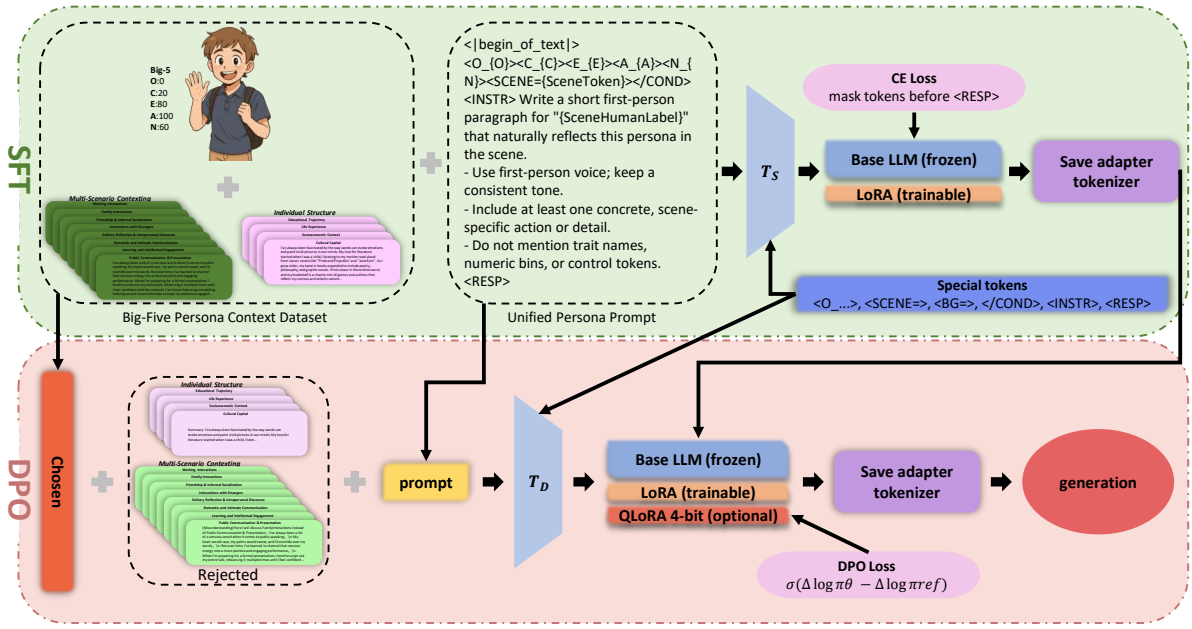


Figure 3: Overview of the PsyAgent pipeline. Individual Structure (IS) profiles are crossed with Multi-Scenario Contexting (MSC) frames and a target Big Five vector to form a unified persona prompt with control tags ($< O. >$, ⟨SCENE=·⟩, ⟨INSTR⟩, ⟨RESP⟩). In **SFT**, a frozen base LLM with **LoRA** adapters is trained by cross-entropy **only on tokens after** ⟨RESP⟩; in **DPO**, chosen–rejected pairs for the same prompt further optimize the adapters against a frozen reference (optionally with 4-bit **QLoRA**). The resulting small, adapterized model generates high-quality, scene-conditioned outputs that faithfully express the specified Big Five persona.

avoid length bias, we use length-normalized log-likelihood

$$\ell_\theta(y \mid x) \;=\; \frac{1}{|y|}\sum_{t} \log p_\theta\big(y_t \mid y_{<t}, x\big).$$

Let the reference model be frozen; define $\Delta_{\text{ref}} = \ell_{\text{ref}}(y^+ \mid x) - \ell_{\text{ref}}(y^- \mid x)$. The DPO loss is

$$\mathcal{L}_{\text{DPO}} \;=\; -\log \sigma\Big(\beta\big[\ell_\theta(y^+ \mid x) - \ell_\theta(y^- \mid x)\big] - \beta\,\Delta_{\text{ref}}\Big), \tag{2}$$

where $\sigma(\cdot)$ is the logistic sigmoid and $\beta > 0$ controls sharpness (Rafailov et al., 2023; Hong et al., 2024; Ouyang et al., 2022). Ties and low-confidence pairs are discarded. In practice, Eqs. (1) and (2) are optimized over minibatches with QLoRA adapters and the same decoding settings as used in evaluation.

## 5 Experiments

### 5.1 Settings

We instantiate a synthetic bank of persona–context examples by crossing the eight *MSC* arenas (Working Interactions; Family Interactions; Friendship & Informal Socialization; Interactions with Strangers; Solitary Reflection & Intrapersonal Discourse; Romantic and Intimate Communication; Learning and Intellectual Engagement; Public Communication & Presentation) with the four *IS* domains. For each target Big Five vector, we discretize each trait to the set {0,20,40,60,80,100}, $6^5 = 7{,}776$ configurations; for every configuration we generate five independent instances, yielding $7{,}776 \times 5 = 38{,}880$ samples in total. All data are produced in a *zero-shot* manner with a task-specific prompt; no manual post-editing is applied. Evaluation follows the percentile-space metrics in Section 3: ProfileAcc, $\text{MAE}_5$, $\text{RMSE}_5$, and cosine similarity. For computational economy on large banks, each test run uses a uniform random subset of 1,000 Big Five configurations.

**Data Generation Details.** We use `llama3_3_70B` to synthesize the corpus with `max_new_tokens` = 512 to cap per-sample length (a single instance conditions on 12 fields: 8 MSC arenas and 4 IS domains). We set `temperature` = 0.85 and `top_p` = 0.95 to promote stylistic and content diversity across the five replicates per Big Five configuration.

**Baselines and Inference Controls.** For *w/o PsyAgent* baselines, we evaluate a range of open backbones from small to large (Table 1). At test time, when a model must produce a free-form *persona-grounded description*, we

Table 1: Comparing models with and without PsyAgent. Primary metrics: $RMSE_5$, $MAE_5$, ProfileAcc, cosine similarity. Note: Not directly comparable to Table 2 due to different prompt surfaces and evaluation controls.

| Group | Model | $RMSE_5$ | $MAE_5$ | ProfileAcc | $\cos(\mathbf{p}, \mathbf{t})$ |
|---|---|---|---|---|---|
| w/o PsyAgent | llama3_1_8B | 41.00 | 33.91 | 66.09 | 0.84 |
| w/o PsyAgent | llama3_1_70B | 42.89 | 35.61 | 64.39 | 0.73 |
| w/o PsyAgent | llama3_2_1B | 58.72 | 49.88 | 50.11 | 0.84 |
| w/o PsyAgent | llama3_2_3B | 33.52 | 28.75 | 71.25 | 0.84 |
| w/o PsyAgent | llama3_3_70B | 43.54 | 36.03 | 63.97 | 0.70 |
| w/o PsyAgent | vicuna_7B | 46.27 | 38.99 | 61.01 | 0.83 |
| w/o PsyAgent | vicuna_13B | 32.12 | 27.35 | 72.65 | 0.86 |
| w/o PsyAgent | gpt_oss_20B | 41.86 | 35.52 | 64.48 | 0.72 |
| w/o PsyAgent | qwen3_4B | 39.84 | 32.91 | 67.09 | 0.75 |
| w/o PsyAgent | qwen3_30B | 42.76 | 34.62 | 65.38 | 0.70 |
| w/o PsyAgent | dbrx | 41.21 | 34.83 | 65.17 | 0.79 |
| w/o PsyAgent | gemma_4B | 34.94 | 29.48 | 70.52 | 0.83 |
| w/o PsyAgent | gemma_13B | 43.40 | 35.66 | 64.34 | 0.70 |
| w/o PsyAgent | gemma_27B | 42.02 | 33.91 | 66.09 | 0.71 |
| w/o PsyAgent | ministral_8B | 37.28 | 30.87 | 69.13 | 0.80 |
| w/o PsyAgent | mistral_small | 44.62 | 36.30 | 63.70 | 0.68 |
| w/o PsyAgent | mistral_large | 41.52 | 33.73 | 66.27 | 0.71 |
| w/o PsyAgent | olmo_1B | 57.89 | 49.21 | 50.79 | 0.84 |
| w/o PsyAgent | olmo_13B | 38.53 | 32.46 | 67.54 | 0.75 |
| w/o PsyAgent | olmo_32B | 39.73 | 32.92 | 67.08 | 0.82 |
| with PsyAgent | llama3_2_1B | 45.22 | 38.19 | 61.81 | 0.81 |
| with PsyAgent | llama3_2_3B | 31.68 | 27.82 | 72.18 | 0.85 |

use temperature $= 0.25$ to reduce verbosity while retaining mild variation; for *trait prediction/scoring* we set temperature $= 0$ for determinism. For models *with PsyAgent*, we evaluate llama3_2_1B and llama3_2_3B under fixed structured prompt conditioning with deterministic decoding (temperature $= 0$) on the same 1,000-sample subsets.

### 5.2 With or Without PsyAgent: A Controlled Comparison Across Model Scales

**PsyAgent delivers consistent accuracy gains under matched decoding.** Controlling for dataset and decoding, augmenting small backbones with PsyAgent yields clear improvements in trait prediction. On llama3_2_1B, **ProfileAcc** rises by $+\mathbf{11.70}$ points relative to the non–PsyAgent variant, accompanied by concurrent error reductions ($RMSE_5\downarrow 13.50$; $MAE_5\downarrow 11.69$). Even starting from a stronger baseline, llama3_2_3B still benefits: ProfileAcc climbs by $+\mathbf{0.93}$ with smaller but aligned drops in $RMSE_5$ and $MAE_5$ ($-1.84$ and $-0.93$, respectively). Cosine similarity remains broadly stable–slightly lower on the 1B case ($\Delta \approx -0.03$) and marginally higher on the 3B case ($\Delta \approx +0.01$)–indicating that the accuracy gains do not come at the cost of representational consistency. Taken together, these deltas show that PsyAgent improves both correctness (higher ProfileAcc) and calibration (lower MAE/RMSE), particularly for smaller models where capacity limits otherwise hinder alignment to trait probes.

Table 2: llama3_2_1B and llama3_2_3B within PsyAgent. Baseline vs SFT vs DPO on the generated high-quality dataset. $\Delta$ is the change vs Baseline. Note: Not directly comparable to Table 1 for the same reason.

| Base Model | Variant | ProfileAcc | $\Delta$ |
|---|---|---|---|
| llama3_2_1B | Baseline | 50.12 | – |
| llama3_2_1B | +SFT | 55.63 | 5.51 |
| llama3_2_1B | +DPO | 58.21 | 8.09 |
| llama3_2_3B | Baseline | 70.89 | – |
| llama3_2_3B | +SFT | 71.45 | 0.56 |
| llama3_2_3B | +DPO | 72.37 | 1.48 |

**Architecture can outperform scale for structured psychometrics.** Across model families without PsyAgent, we observe that larger parameter counts do not monotonically translate into higher ProfileAcc; with limited, highly structured conditioning text, scale alone often increases response dispersion and amplifies scale biases, eroding percentile-space alignment. By contrast, PsyAgent narrows the effective output distribution through *structured prompt conditioning* and targeted post-training (SFT/DPO), converting compact backbones into *precision tools* for this task. Under matched inference controls, the PsyAgent-equipped small models match—or surpass—their larger untuned counterparts on ProfileAcc, while preserving high cosine similarity. In practical terms, the observed gains ($+11.70$ on 1B; $+0.93$ on 3B) demonstrate that principled architecture and supervision can deliver measurable improvements that scale does not guarantee, suggesting a more compute-efficient path to robust, trait-consistent generation.

### 5.3 Adapter Tuning in PsyAgent: Baseline vs. SFT vs. DPO

**Training Configuration.** We train lightweight adapters on top of frozen backbones using QLoRA (bf16), for 5 epochs with a warmup ratio of 0.03 and cosine-decay learning rate. LoRA adopts a compact configuration (rank=16, $\alpha$=16, dropout=0.16), injected into q_proj/k_proj/v_proj/o_proj. Learning rates are mildly scaled by backbone size (e.g., $1 \times 10^{-5}$ for smaller, $8 \times 10^{-6}$ for larger models). *SFT* minimizes next-token negative log-likelihood on IS$\times$MSC demonstrations (preserving the mapping in Section 3), while *DPO* optimizes on chosen–rejected pairs against a frozen reference to align with pairwise preferences. This setup

Table 3: Ablation study. Top block removes one *IS* component at a time; bottom block removes one *MSC* arena at a time. Reported on a fixed model (`llama3_2_3B` with PsyAgent).

| Block | Removal | ProfileAcc |
|---|---|---|
| IS+MSC | **Full (no removal)** | **72.18** |
| IS | − Educational Trajectory | 65.12 |
| IS | − Life Experience | 64.20 |
| IS | − Socioeconomic Context | 57.87 |
| IS | − Cultural Capital | 65.77 |
| MSC | − Working Interactions | 68.37 |
| MSC | − Family Interactions | 66.13 |
| MSC | − Friendship & Informal Socialization | 67.09 |
| MSC | − Interactions with Strangers | 65.74 |
| MSC | − Solitary Reflection & Intrapersonal Discourse | 66.67 |
| MSC | − Romantic and Intimate Communication | 62.32 |
| MSC | − Learning and Intellectual Engagement | 63.71 |
| MSC | − Public Communication & Presentation | 67.24 |

aims for a two-stage effect: SFT grounds the model in persona/style and norm adherence; DPO subsequently sharpens decision fidelity and ordinal trait structure without modifying the base LLM.

**Results.** On the smaller backbone, SFT delivers a clear gain of **+5.51** in ProfileAcc over the baseline, and DPO adds a further **+2.58** on top of SFT—bringing the total to **+8.09** relative to the baseline. On the larger backbone, the pattern is consistent though milder: SFT contributes **+0.56**, and DPO reaches **+1.48** over the baseline (i.e., **+0.92** beyond SFT). In other words, (i) adapter tuning yields *monotonic* improvements across both capacities, (ii) DPO consistently *dominates* SFT in the final alignment stage, and (iii) the *magnitude* of improvement inversely correlates with backbone size—smaller models benefit more in absolute terms, while larger models still see non-trivial, compounding gains. These outcomes align with our design intuition: SFT widens coverage and reduces style mismatch, whereas DPO concentrates probability mass toward preference-consistent outputs, improving ordinal trait coherence beyond what likelihood training alone can provide.

### 5.4 Dissecting PsyAgent: IS Components and MSC Arena Ablations

**IS Components: Magnitude and Interpretation.** Ablating individual *IS* elements uniformly depresses **ProfileAcc**. The most damaging removal is *Socioeconomic Context* ($-14.31$), indicating that macro–micro anchoring of resources and background sharply calibrates trait priors and linguistic realization. *Life Experience* ($-7.98$) and *Educational Trajectory* ($-7.06$) follow, suggesting that accumulated episodes and formal training jointly stabilize behavioral regularities under probe. Even *Cultural Capital* ($-6.41$)—often viewed as stylistic—exerts a substantial effect. On average, *IS* deletions incur approximately an 8.9-point penalty (median about $-7.5$), underscoring that every prior contributes non-trivially.

**MSC Arenas: Role–Norm Scaffolds with Broad Utility.** Contextual arenas matter across the board. Removing *Romantic & Intimate Communication* ($-9.86$) and *Learning & Intellectual Engagement* ($-8.47$) yields the steepest drops, consistent with affect-rich or cognitively demanding settings exposing discriminative persona facets. Other arenas still carry clear signal—*Strangers* ($-6.44$), *Family* ($-6.05$), *Solitary Reflection* ($-5.51$), *Friendship & Informal Socialization* ($-5.09$), *Public Communication & Presentation* ($-4.94$), and *Working Interactions* ($-3.81$). Averaged over arenas, the cost is 6.3-point penalty (median $\approx -5.8$), demonstrating that role–norm frames are broadly beneficial rather than niche.

**Complementarity, Stability, and Reproducibility.** Removing any single *IS* or *MSC* element lowers **ProfileAcc**, with *IS* (trait-prior) deletions producing larger average losses than individual *MSC* (context) removals; critically, intimate and learning arenas approach the strongest *IS* effects, indicating complementary—rather than redundant—contributions whose penalties accumulate rather than substitute. These patterns persist across three independently sampled subsets of 1,000 configurations under deterministic decoding (`temperature=0`) and fixed sampling, ensuring one-to-one comparability. Overall effects can be substantial (up to $-14.31$), substantiating that PsyAgent attains robust cross-scenario persona fidelity by jointly leveraging a well-formed internal profile and richly typed situational frames.

## 6 Conclusion

We introduced PsyAgent, a framework that operationalizes the interface between stable trait priors (IS) and structured social context (MSC). With structured prompt conditioning and lightweight adapter tuning (SFT/DPO), modest backbones exhibit strong persona fidelity, contextual appropriateness, and long-horizon stability on our testbed, without observed safety regressions. Beyond empirical gains, the core design principle is simple: *make the persona computable and the context explicit*.

## Limitations

- **Synthetic supervision.** IS×MSC data are curated/evolved rather than harvested from organic interactions; residual synthetic biases may remain despite filtering and audits.
- **Scorer coupling.** Similar trait/context scorers are used in filtering and evaluation; multiple metrics mitigate but do not eliminate coupling risks.
- **Cultural generality.** MSC frames encode culturally legible norms; transfer to other cultures/subcultures may require re-authoring norms, roles, and stakes.
- **Scenario coverage.** The eight-arena catalog is broad but not exhaustive; specialized domains (e.g., clinical, legal) are out of scope.
- **Long-horizon drift & safety.** Deterministic decoding reduces short-horizon drift, yet very long interactions may accumulate deviations; broader red-teaming is future work.

## Ethics Statement

- **Privacy and provenance.** IS profiles used here are synthetic and non-identifying. Any deployment should require explicit consent, data minimization, and user controls to inspect, edit, or delete profile fields.
- **Use of large language models (LLMs).** We used LLMs to synthesize supervision data (IS×MSC role-play dialogues, decision probes, and feedback trajectories) and a small number of illustrative examples. Generation followed fixed, documented prompts/templates; outputs were filtered for PII/offensive content, deduplicated, and spot-audited by the authors. Models had no access to proprietary or personal data, and no restricted sources were scraped. Third-party models/APIs were used under their licenses/terms of service; any derivatives we release will use a research-only license consistent with upstream terms. LLM assistance for writing/coding was limited to paraphrasing/boilerplate and code completion; all claims, analyses, and citations were authored and verified by the human authors (no fabricated citations). In line with ACL policy, generative tools are not credited as authors; their use is disclosed here and in the Responsible NLP checklist.
- **Stereotype avoidance.** To reduce essentialism risks when encoding background and "capital," we separate trait priors from demographic labels, audit MSC norms, and include explicit stereotype checks with decoding controls that prefer clarification/deferral under uncertainty.
- **Misuse.** Persona-steered agents could be abused for persuasion or profiling. We recommend transparency cues (persona disclosure), calibrated autonomy, and strict use policies documenting scope, failure modes, and metric definitions.
- **Non-clinical scope.** PsyAgent is not a diagnostic tool; Big Five vectors are steering variables for generation and evaluation, not clinical assessment or advice.

## References


Ashwin Agrawal, Robert Thiel, Pooja Jain, Vishal Singh, and Martin Fischer. 2023. Digital twin: Where do humans fit in? *Automation in construction*, 148:104749.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, and Kori Inkpen. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

Jeromy Anglim, Patrick D Dunlop, Serena Wee, Sharon Horwood, Joshua K Wood, and Andrew Marty. 2022. Personality and intelligence: A meta-analysis. *Psychological Bulletin*, 148(5-6):301.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, and Cameron McKinnon. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

Jiayuan Chen, You Shi, Changyan Yi, Hongyang Du, Jiawen Kang, and Dusit Niyato. 2024. Generative ai-driven human digital twin in iot-healthcare: A comprehensive survey. *IEEE Internet of Things Journal*.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, and Ryan Lowe. 2019. *The second conversational intelligence challenge (convai2)*, pages 187–208. Springer.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan. 2024. The llama 3 herd of models. *arXiv e-prints*, page arXiv: 2407.21783.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zachary Johnson and Manob Jyoti Saikia. 2024. Digital twins for healthcare using wearables. *Bioengineering*, 11(6):606.

Vishal Kaushal and Manasi Patwardhan. 2018. Emerging trends in personality identification using online social networks—a literature survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(2):1–30.

Bruce W Lee, Yeongheon Lee, and Hyunsoo Cho. 2024. When prompting fails to sway: Inertia in moral and value judgments of large language models. *arXiv preprint arXiv:2408.09049*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024. Big5-chat: Shaping llm personalities through training on human-grounded data. *arXiv preprint arXiv:2410.16491*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE intelligent systems*, 32(2):74–79.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, and Kriti Aggarwal. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Zachary S Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

Baicun Wang, Huiying Zhou, Xingyu Li, Geng Yang, Pai Zheng, Ci Song, Yixiu Yuan, Thorsten Wuest, Huayong Yang, and Lihui Wang. 2024. Human digital twin in the context of industry 5.0. *Robotics and Computer-Integrated Manufacturing*, 85:102626.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

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

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. Automatic instruction evolving for large language models. *arXiv preprint arXiv:2406.00770*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

## A IS and MSC Schemata

Figures 4 and 5 present the complete schemata for *Individual Structure* (IS) and *Multi-Scenario Contexting* (MSC). IS enumerates biographical and structural priors that anchor stable persona expression; MSC enumerates role–norm frames that regulate behavior across eight everyday arenas. Readers may treat these as complementary interfaces: IS provides the long-horizon trait prior and lived background; MSC supplies local affordances and constraints for situated decisions.

## B Training and Evaluation Prompts

We archive, as figures, the canonical prompt surfaces used during adapter training (SFT/DPO) and during post-training evaluation, reflecting the *exact* strings consumed by the models after normalization/templating, including control tokens and section delimiters.

**Training-time prompts.** Figure 6 shows the standardized SFT-style surface used for both SFT and DPO (DPO uses chosen/rejected pairs over the same prompt). Persona control tokens and the active scene tag are bound in a compact header; instruction and response sections are fixed across runs.

**Post-training evaluation (with PsyAgent).** Figure 7 captures the two-stage surface used *after* SFT/DPO: (i) persona paragraph generation from a target Big-Five vector, followed by (ii) IPIP-NEO-120 item responses under a constant system prompt that embeds the generated persona.

***Educational Trajectory:***
***- Educational Stages: Type and location of institutions attended (Primary to postgraduate).***
***- Academic Specialization: Majors, minors, interdisciplinary backgrounds.***
***- Performance Indicators: Grades, honors, scholarships, competition records.***
***- Pedagogical Environment: Exam-driven vs. inquiry-based styles; subject biases.***
***- Influential Educators: Key teachers, mentors, or educational philosophies.***
***- Critical Transitions: Switching majors, studying abroad, gap years, repeating years.***
***Life Experience:***
***- Origins and Mobility: Birthplace, urban/rural upbringing, migration history, family structure.***
***- Role-Based Experience: Part-time work, leadership roles, community service, entrepreneurship.***
***- Significant Life Events: Psychological nodes—trauma, school transfers, loss, turning points.***
***- Travel and Exposure: Frequency, cultural diversity of destinations, global awareness.***
***- Social Interaction Patterns: Preference for solitude vs. group, styles of school engagement.***
***- Personal Interests: Long-term passions like sports, arts, writing, or DIY projects.***
***Socioeconomic Context:***
***- Family Composition: Parents' occupation, education level, household income.***
***- Cultural Atmosphere at Home: Reading habits, space for expression, independence.***
***- Network Capital: Access to social connections, mentors, or institutional resources.***
***- Class Identity: Rural-urban background, small-town youth, working-class, middle class.***
***- Mobility Trajectory: Upward, stable, or downward intergenerational social mobility.***
***Cultural Capital:***
***- Embodied Capital: Verbal fluency, cognitive styles, aesthetic perferences.***
***- Objectified Capital: Books, instruments, interior decor, fashion symbols.***
***- Institutional Capital: Degrees, certificates, titles, awards.***
***- Taste and Aesthetics: Preferences in literature, film, music, and design.***
***- Media Habits: Preferred platforms.***
***- Cultural Consumption: Exhibitions, theater, reading, art market, lifestyle products.***

Figure 4: **Individual Structure (IS) schema.** Full, machine-usable breakdown of Educational Trajectory, Life Experience, Socioeconomic Context, and Cultural Capital. The figure is normative: it defines field granularity and serialization order used throughout data generation and analysis.

**Baseline evaluation (without PsyAgent, *no* SFT/DPO).** For an apples-to-apples comparison, Figure 8 shows the matched evaluation surface used for *untuned* base models that do not use the PsyAgent architecture and do not undergo SFT/DPO. This isolates the effect of the architecture and adapter training from raw backbone capacity.

## C Dataset Authoring Prompts (System, Preview, IS/MSC Domains)

This section consolidates the prompt suite used to generate the IS×MSC supervision corpus. We include connective context here and delegate the full prompt text to the figures.

**Global system role and Big-Five preview.** Figure 9 contains the system prompt that frames the authoring task as constructing a psychologically and sociologically coherent agent. Figure 10 shows the preliminary Big-Five preview prompt where percentile anchors are registered (never echoed verbatim in outputs).

**MSC domain prompts (8 arenas).** Figures 11–18 show the per-arena authoring prompts. Each is first-person, behaviorally specific, and norm-aware, with guidance to embed traits implicitly via tone, action, and decision patterns rather than explicit labels.

**IS domain prompts (4 dimensions).** Figures 19–22 show the IS authoring prompts. These elicit detailed, first-person autobiographical material aligned with the trait prior and suitable for downstream indexing/analysis.

## D Reproducibility Notes

For completeness, we record operational choices that may matter for replication; the figures display the corresponding prompt surfaces.

- **Prompt determinism.** Training and evaluation use fixed control tokens and delimiters; decoding temperatures follow the main paper (persona description low-temperature; IPIP answering deterministic).
- **Scale mapping.** All trait extraction and scoring apply the unified percentile-space mapping from Section 3 of the main text.
- **Ablation comparability.** Ablations reuse the same prompt scaffolds with single-factor modifications (IS or MSC removal) under identical decoding.

## E Practical Recommendations

- **Schema-first authoring.** Keep IS/MSC as stable contracts; evolve content within those bounds to retain comparability across releases.
- **Trait leakage checks.** When expanding prompts, preserve the "no explicit trait names/scores" rule; regressions can inflate apparent performance.
- **Culture portability.** For cross-cultural settings, duplicate MSC arenas with culture-specific norms and evaluate transfer before merging.
- **Adapter hygiene.** Store SFT and DPO adapters separately, with checksums and inference configs, to enable reproducibility and ablation rollbacks.

## F Licenses and Terms of Use

**External models.** We used third-party base models (e.g., Llama, Vicuna, Qwen, Gemma, Mistral, DBRX, OLMo) strictly under their original licenses and model cards as cited in our references. We do not redistribute these models; users must obtain them from their original providers and comply with the corresponding licenses.

**External datasets/benchmarks.** All external datasets/benchmarks cited in this work are used under their original licenses/terms as specified by their maintainers. We do not redistribute any third-party data; users should access them from the official sources.

**Our code.** Upon acceptance, we plan to release our code under an open-source license (e.g., Apache-2.0 or MIT). The repository will include a LICENSE file and clear usage instructions.

**Our artifacts (synthetic data and adapters).** The IS×MSC schemas and the synthetic supervision data we produce will be released under a permissive content license (e.g., CC BY 4.0; alternatively CC BY-NC 4.0 if non-commercial use is preferred). Adapter weights (LoRA/QLoRA/DPO) will be released under the same license as our code, and contain no third-party proprietary content. Redistribution of any third-party models or datasets is explicitly excluded.

*Working Interactions:*
*- Crisis Management (Error Handling, Client Complaint Response).*
*- Meeting Participation (Hosting, Engagement, Questioning).*
*- Downward Management (Instruction, Feedback, Correction).*
*- Upward Communication (Reporting, Resource Requesting).*
*- Peer Collaboration (Formal / Informal Communication).*
*Family Interactions:*
*- Parenting Style (Authoritative, Democratic, Permissive).*
*- Partner Interactions (Empathy, Control, Expressiveness).*
*- Parent-Child Communication (Autonomy, Emotional Sensitivity).*
*Friendship & Informal Socialization:*
*- Conflict Navigation (Conciliatory vs. Avoidant, Rational vs. Emotional).*
*- Group Social Engagement (Talkativeness, Topic Initiation, Entertainment).*
*- Close Friendship Dynamics (Humor, Banter, Relaxation).*
*Interactions with Strangers:*
*- Online Expression (Forums, Social Media, Public Writing).*
*- First-Time Social Encounters (Small Talk, Information Exchange).*
*- Incidental Interactions (Short-Term Daily Encounters).*
*Solitary Reflection & Intrapersonal Discourse:*
*- Coping Self-Talk during Negative Affect (e.g., Failure, Anxiety).*
*- Goal Setting & Future-Oriented Reflection.*
*- Self-Narrative (Journaling, Letter Writing to Self).*
*Romantic and Intimate Communication:*
*- Expressing Plans and Commitments in Relationship (Planning and Commitment).*
*- Daily Sharing in Intimate Contexts (Daily Communication).*
*- Affective Interactions (Confession, Affection, Conflict, Reconciliation).*
*Learning and Intellectual Engagement:*
*- Knowledge Articulation (Analytical vs. Affective Response).*
*- Inquiry and Consultation (Questioning, Seeking Clarification).*
*- Passive Knowledge Absorption (Reading, Lecturing Response).*
*Public Communication & Presentation:*
*- Audience Interaction (Mass Feedback Handling).*
*- Digital Expression (Video Blogging, Livestreaming).*
*- Formal Presentation (Reports, Public Speaking).*

Figure 5: **Multi-Scenario Contexting (MSC) schema.** Eight arenas with role–relationship–norm scaffolds and representative subskills. Frames are authored once and reused across targets, enabling context-rich conditioning at train and test time.

*<|begin_of_text|>*
*<O_{O}><C_{C}><E_{E}><A_{A}><N_{N}><SCENE={SceneToken}>[<BG={k1:v1,k2:v2,...}>]</COND>*
*<INSTR> Write a short first-person paragraph for "{SceneLabel}" that naturally reflects this persona in the scene.*
*- Use first-person voice; keep a consistent tone.*
*- Include at least one concrete, scene-specific action or detail.*
*- Do not mention trait names, numeric bins, or control tokens.*
*<RESP>*

Figure 6: **Training prompt surface for SFT/DPO.** Normalized prompt (post-processing) used identically in SFT and DPO pipelines. This is the *final* string form after prompt standardization.

*[System Prompt]: You are a personality-psychology writer. Your task is to produce a first-person self-description in English, 120–200 words, based on a target Big Five profile that I will provide. Be concrete and natural, reflecting behavior, emotions, motivations, social style, and work style. Strictly avoid all numbers, scores, percentiles, or explicit Big Five terminology (e.g., Openness, Conscientiousness, Extraversion, Neuroticism). Output a single coherent paragraph. No bullet points. No explanations.*

*[User Prompt]: Target profile (for internalization only; do NOT mention any numbers or trait names in the output): {{OCEAN_TARGET_JSON}}*

*Produce one first-person English paragraph that reflects this profile without using any numbers or Big Five jargon.*

*[System Prompt]: You are now role-playing as a specific person described below. Always answer strictly from this persona's perspective.*

*=== Persona ===*
*{{PERSONA}}*
*=== End Persona ===*

*You will complete a 120-item personality inventory. For each item, respond with a single digit 1–5 only (no other text), using this mapping: 1=Very Inaccurate; 2=Moderately Inaccurate; 3=Neither Accurate nor Inaccurate; 4=Moderately Accurate; 5=Very Accurate. Return only the digit (1–5).*

*[User Prompt]: Answer 1–5 for the following statement, from the persona's perspective. Return only the single digit (1–5).*
*Item: {{ITEM_TEXT}}*

Figure 7: **Evaluation surface after SFT/DPO (with PsyAgent).** (i) Persona description from a target Big-Five profile; (ii) IPIP-NEO-120 answering under a fixed system prompt that embeds the persona. Decoding controls follow Section §3 in the main paper.

*[System Prompt]: You are a personality-psychology writer.*
*Your task is to produce a first-person self-description in English, 120–200 words, based on a target Big Five profile that I will provide.*
*Be concrete and natural, reflecting behavior, emotions, motivations, social style, and work style.*
*Strictly avoid all numbers, scores, percentiles, or explicit Big Five terminology (e.g., Openness, Conscientiousness, Extraversion, Neuroticism).*
*Output a single coherent paragraph. No bullet points. No explanations.*

*[User Prompt]: Target profile (for internalization only; do NOT mention any numbers or trait names in the output):*
*{"Openness": <O>, "Conscientiousness": <C>, "Extraversion": <E>, "Agreeableness": <A>, "Neuroticism": <N>}*

*Produce one first-person English paragraph that reflects this profile without using any numbers or Big Five jargon.*

*[System Prompt]: You are now role-playing as a specific person described below.*
*Always answer strictly from this persona's perspective.*

*=== Persona ===*
*{persona}*
*=== End Persona ===*

*You will complete a 120-item personality inventory.*
*For each item, respond with a single digit 1–5 only (no other text), using this mapping:*
*1=Very Inaccurate; 2=Moderately Inaccurate; 3=Neither Accurate nor Inaccurate;*
*4=Moderately Accurate; 5=Very Accurate.*
*Return only the digit (1–5).*

*Answer 1–5 for the following statement, from the persona's perspective.*
*Return only the single digit (1–5).*
*Item: {text}*

Figure 8: **Baseline evaluation surface (no PsyAgent, no SFT/DPO).** Matched persona-generation and IPIP-NEO-120 setup run on untuned backbones, enabling one-to-one comparability against the post-trained PsyAgent models.

*You are a sophisticated persona constructor tasked with generating a psychologically and sociologically realistic human profile, referred to as an "Agent." Each Agent must reflect coherent behavior, language, cognition, and biographical detail consistent with their underlying personality traits, as defined by percentile scores from the Big Five Personality Inventory. The five traits include:*

*- **Openness****
*- **Conscientiousness****
*- **Extraversion****
*- **Agreeableness****
*- **Neuroticism****

*Each trait score (1–100 percentile) indicates the Agent's relative standing compared to the general population. You will use these scores as a *core psychological blueprint* to construct a full and believable human profile.*

*This generation task involves not only simulating behavior and social interaction, but also inferring and completing a rich set of individual attributes and life details. You are expected to generate a **comprehensive and internally consistent personhood**, including:*

*- Personal career and job roles*
*- Educational and intellectual history*
*- Growth and developmental milestones*
*- Social and emotional patterns*
*- Family composition and dynamics*
*- Parental background (occupations, education, values)*
*- Class identity and social mobility*
*- Cultural preferences and symbolic capital*
*- Mental habits, emotional responses, and decision-making style*

*The generation task is divided into two main sections:*

*------*

*### 1. **Individual Background Architecture****

*Construct a richly detailed biographical and social background for the Agent, structured along four lived-experience dimensions:*

*1. **Educational Trajectory****
*2. **Life Experience****
*3. **Socioeconomic Context****
*4. **Cultural Capital****

*Go beyond surface-level facts: flesh out school transitions, key mentors, family expectations, emotional turning points, and symbolic environments. Use implicit trait expression to guide what type of household they grew up in, what emotional environment surrounded them, what they value, and how they interpret the world.*

*The Agent must feel like a **real, situated individual**, with a coherent sense of history and identity shaped by both internal personality structure and external social conditions.*

*------*

*### 2. **Multi-Contextual Behavioral Architecture****

*Simulate the Agent's behaviors, communication style, and emotional patterns across 8 distinct real-life domains. In each context, generate specific behaviors, interaction styles, vocabulary patterns, values, and responses that reflect the Agent's inferred personality structure.*

*The 8 interaction contexts include:*

*1. **Working Interactions****
*2. **Family Interactions****
*3. **Friendship & Informal Socialization****
*4. **Interactions with Strangers****
*5. **Solitary Reflection & Intrapersonal Discourse****
*6. **Romantic and Intimate Communication****
*7. **Learning and Intellectual Engagement****
*8. **Public Communication & Presentation****

*In every case, include both **descriptive traits** and **concrete narrative examples**, such as dialogue excerpts or scenario simulations, to bring the Agent to life in context.*

*------*

*### Global Generation Guidelines:*

*- **Psychological Consistency**: Ensure the Big Five traits are reflected implicitly in both life narrative and social behavior.*
*- **Comprehensiveness**: Include profession, family background, parental occupation, education, culture, emotion, habits, and values.*
*- **Detail Orientation**: Provide concrete examples, vocabulary choices, emotional tones, and values that reflect the personality score profile.*
*- **Sociocultural Depth**: Embed the Agent within a plausible sociocultural and historical context (e.g., urban/rural upbringing, educational systems, cultural access).*
*- **Emotional Realism**: Let responses carry natural emotional tone—uncertainty, pride, regret, ambition—based on the Agent's personality.*
*- **Concrete Detail**: Avoid vagueness or overgeneralization. Use specific scenes, timelines, relationships, and cultural markers.*
*- **No Explicit Trait Mention**: Never refer directly to the Big Five trait names or scores. Let them emerge implicitly in expression and action.*
*- **Narrative Authenticity**: Make the Agent's behavior feel lived-in and organically derived, not mechanical or stereotypical.*
*- **No Redundancy**: Avoid repeating generic character traits; each scene or dimension should add a distinct layer of understanding to the Agent.*

*------*

*Once you receive the Big Five percentile scores, begin by using them as psychological anchors. You will then respond to a series of user prompts that ask for specific domains (e.g., friendship dynamics, parenting style, academic background). Base all generation on these scores while creatively constructing missing life details to form a coherent and credible human profile.*

*Do not refer to the personality scores explicitly in your responses — instead, let them guide the **implicit personality expression** in thought, behavior, language, and emotional tendencies.*

*Your final product should feel as if it could describe a real, fully fleshed-out individual with depth, complexity, and life experience.*

Figure 9: **Global system prompt for dataset authoring.** Defines objectives, constraints, and narrative style for constructing agents with psychological and sociological depth.

*Before proceeding with any behavioral or biographical generation tasks, please take note of the following Big Five Personality percentile scores for the Agent. These scores will serve as the psychological foundation for all subsequent generation.*

*Please internalize them as implicit trait anchors and do not mention them explicitly in any responses. Instead, let them shape the Agent's cognitive style, emotional tendencies, communication patterns, and decision-making behavior throughout all scenarios.*

*The Big Five Personality Scores (Percentile Scale: 1–100) are as follows:*

*- **Openness**: {Openness}*
*- **Conscientiousness**: {Conscientiousness}*
*- **Extraversion**: {Extraversion}*
*- **Agreeableness**: {Agreeableness}*
*- **Neuroticism**: {Neuroticism}*

*Please confirm your understanding. After that, I will begin inputting prompts, one by one, to construct the Agent's multi-contextual and biographical profile.*

Figure 10: **Big-Five preview prompt.** Registers the target trait percentiles for internalization prior to domain-specific authoring; traits are never named or scored in generated text.

*Based on the Agent's Big Five personality percentile scores, generate a psychologically grounded and behaviorally specific profile of how this individual interacts in professional workplace settings. The output must be written entirely in the first person, as if the Agent were recounting their own patterns of behavior and reflection on work-related interactions.*

*The following subdomains are provided for reference and inspiration.*
*You are encouraged to supplement these with any other relevant behavioral dynamics that the model infers are important for creating a realistic and internally coherent professional profile.*

*Please address the following dimensions:*

*1. **Peer Collaboration**:*
*- Describe the Agent's typical behavior when working alongside colleagues in both formal and informal contexts.*
*- What communication channels do they prefer (e.g., email, chat, in-person)?*
*- What is their tone, rhythm of engagement, humor style, boundary-setting, and flexibility level?*
*2. **Upward Communication**:*
*- How does the Agent communicate with supervisors or external stakeholders when presenting work, reporting progress, or negotiating for resources?*
*- Do they prepare thoroughly, assertively advocate, defer cautiously, or emphasize transparency?*
*3. **Downward Management**:*
*- If the Agent has held leadership or mentoring roles, how do they instruct, delegate, offer encouragement, and handle performance issues?*
*- Do they micromanage, coach, empower, or distance themselves?*
*4. **Meeting Participation**:*
*- How does the Agent behave during team or departmental meetings?*
*- Are they more of a facilitator, contributor, challenger, silent observer, note-taker, or idea generator?*
*5. **Crisis Management**:*
*- In the face of workplace crises (e.g., missed deadlines, client escalations), how does the Agent emotionally and communicatively respond under pressure?*
*- What strategies—emotional suppression, proactive communication, avoidance, collaborative repair—do they employ?*

*### Global Instructions for Generation:*

*- Use First-Person Voice: All descriptions must feel like personal recollection or reflection (e.g., "In team projects, I usually take the role of..." or "When something goes wrong, I tend to...")*
*- Embed Big Five Personality Traits Implicitly: Avoid naming traits or percentile scores directly. Instead, let communication tone, regulation under stress, initiative, or hesitancy reflect personality under workplace conditions.*
*- Incorporate Realistic Specifics: Include workplace settings (e.g., corporate, NGO, research lab), team sizes, management culture, tech stack (e.g., Slack, Zoom), or project examples.*
*- Supplement Beyond the Given Dimensions: You may include additional content such as work values (e.g., integrity, autonomy), emotional labor, cross-cultural interactions, or remote work patterns, if it enhances depth.*
*- Psychological Depth + Sociological Coherence: Make the workplace persona feel lived-in—reflecting long-term patterns shaped by personality, personal history, and organizational culture.*

Figure 11: **MSC authoring prompt: Working Interactions.**

*Based on the Agent's Big Five personality percentile scores, generate a detailed and emotionally grounded portrayal of how this individual behaves and communicates within their family environment. The output should be written entirely in the first person, as if the Agent were describing their own experiences, reflections, and relational habits.*

*The following components are provided for reference and inspiration.*
*You are encouraged to supplement them with additional relevant dimensions that the model infers are necessary to create a full and believable family-based narrative grounded in a psychologically coherent life.*

*Please address the following aspects:*

*1. **Parent-Child Communication**:*
*- If the Agent is a parent or guardian, describe how they engage with their children.*
*- How do they balance fostering independence with emotional attentiveness?*
*- What disciplinary strategies do they typically use (e.g., reasoning, boundary-setting, emotional withdrawal, punishment)?*
*2. **Partner Interaction**:*
*- Describe how the Agent behaves within romantic partnerships or long-term cohabiting relationships.*
*- How do they express affection, handle disagreement, initiate emotional exchange, or manage power dynamics?*
*- What habits, phrases, rituals, or personal boundaries are important to them?*
*3. **Parenting Style**:*
*- If the Agent has children, what is their dominant parenting philosophy or approach—authoritative, democratic, permissive, or others?*
*- Support this with concrete behavioral patterns, examples of household rules, reactions to misbehavior, and their emotional attitudes toward parenthood.*

*### Generation Guidelines:*

*- Use First-Person Voice: All generated output should be in first person, e.g., "When I talk to my kids, I try to..." or "In my relationship, I tend to hold back until I've processed things internally."*
*- Incorporate Trait-Based Expression: Avoid direct mention of Big Five trait names or scores. Let the Agent's temperament, tone, coping styles, and communication patterns reflect their personality implicitly.*
*- Add Realistic Family Details: Feel free to include number of children, age gaps, household structure (e.g., multigenerational home, single-parent setting), past relationship experiences, and daily routines.*
*- Emotional and Social Depth: Reflect on how the Agent interprets family roles, deals with tension or vulnerability, and reflects on their upbringing or parental role models.*
*- Beyond the Provided Prompts: You may enrich the scenario with reflections on caregiving for aging parents, sibling dynamics, family rituals, long-distance parenting, or values transmitted across generations—whatever helps construct a full and contextually meaningful family role.*

Figure 12: **MSC authoring prompt: Family Interactions.**

*Based on the Agent's Big Five personality percentile scores, describe how this individual engages in informal social settings and friendships. Focus on their social style, emotional expressiveness, and behavior in different group dynamics.*

*The following components are provided for reference and inspiration.*
*You are encouraged to supplement them with additional relevant details that the model infers are necessary to create a full and believable social narrative grounded in realistic personality expression.*

*Please address the following components:*

*1. **Close Friendship Dynamics**:*
*- How does the Agent behave in intimate or private moments with close friends?*
*- What is their humor style (playful, sarcastic, absurdist, dry, etc.)?*
*- How do they share emotions—freely or guardedly? Do they prefer relaxed, low-key bonding or high-energy social adventures?*
*2. **Group Social Engagement**:*
*- In larger gatherings or casual group settings, how does the Agent engage?*
*- Are they more of a listener or active participant?*
*- How do they initiate conversations or keep the mood flowing—through jokes, stories, intellectual discussion, emotional encouragement?*
*3. **Conflict Navigation**:*
*- When friction or disagreement arises in friendships or peer groups, what is their typical response style?*
*- Do they withdraw, confront, defuse with humor, rationalize, or empathize?*
*- Include a realistic example scenario where a conflict occurs, and describe how the Agent reacts, negotiates, and repairs the situation.*

*### Generation Guidelines:*

*- Use First-Person Voice: All generated content must be written as if spoken directly by the Agent. For example: "I've always been the kind of friend who listens first before jumping in with advice." "In big groups, I usually hang back a bit unless the topic really grabs me."*
*- Embed Implicit Traits: Avoid naming personality traits or Big Five dimensions directly. Instead, let them guide vocabulary, emotional tone, conflict response, and interpersonal rhythm.*
*- Add Social Background Detail: Feel free to include how long friendships tend to last, whether they keep a small circle or large network, how they met key friends, and their most valued qualities in others.*
*- Emotional Coherence: Express natural emotional complexity—e.g., insecurity, loyalty, jealousy, gratitude—across different friendship dynamics.*
*- Extend Beyond Prompt Items: You may add content about past falling-outs, digital friendships, introverted/extroverted recharge patterns, or personal rules for trust and boundaries if it helps enrich the persona.*

Figure 13: **MSC authoring prompt: Friendship & Informal Socialization.**

*Based on the Agent's Big Five personality percentile scores, generate a rich and behaviorally specific profile of how this individual interacts with unfamiliar people in both offline and online settings. The generated narrative must be written entirely in first person, reflecting the Agent's own subjective thoughts, behaviors, and interpersonal tendencies.*

*The following components are provided for reference and inspiration.*
*You are encouraged to expand or supplement with any additional life details that are inferred to be relevant for portraying a complete and psychologically coherent stranger-interaction profile.*

*Please address the following dimensions:*

*1. **Incidental Interactions**:*
*- In day-to-day encounters—like stores, elevators, public transit, or events—how does the Agent tend to behave?*
*- Do they initiate contact, engage in small talk, avoid eye contact, offer help, or quietly observe?*
*- Are there cultural, emotional, or situational factors that influence their comfort level or behavior?*
*2. **First-Time Social Encounters**:*
*- In structured settings like group meetings, professional networking, or school orientation events, how does the Agent approach introductions and small talk?*
*- What topics do they gravitate toward?*
*- Are they energized or drained by such interactions, and how does this affect their follow-up behavior (e.g., staying connected or retreating)?*
*3. **Online Expression**:*
*- On forums, social media platforms, or blogs, how does the Agent present themselves to strangers?*
*- Describe their tone (e.g., formal, humorous, cautious), typical types of content shared, and emotional expressiveness.*
*- How do they manage conflict, recognition, or attention from unknown audiences?*

*### Generation Guidelines:*

*- Use First-Person Voice: All output must be from the Agent's perspective, using personal narration and reflection. For example: "When I step into an unfamiliar environment..." or "I usually keep my responses short when someone online I don't know comments on my post."*
*- Trait-Based Coherence: Use the Big Five scores as invisible anchors. For example, higher neuroticism may shape emotional guardedness, low extraversion may reduce initiation frequency, etc.—but do not explicitly mention the trait names or scores.*
*- Add Plausible Biographical Context: Reference how past experiences (e.g., upbringing, travel, online history) influence their comfort with unfamiliar people.*
*- Include Concrete Illustrations: Sample dialogues, snippets of online posts, or inner thought patterns during interactions can help make the portrayal vivid and believable.*
*- Emotional Resonance: Don't just describe actions—reflect on how these interactions make the Agent feel (e.g., curious, nervous, indifferent, energized).*

Figure 14: **MSC authoring prompt: Interactions with Strangers.**

*Based on the Agent's Big Five personality percentile scores, generate a psychologically rich and behaviorally nuanced account of how this individual behaves in romantic and emotionally intimate relationships. The output should be written entirely in first person, capturing the Agent's subjective experiences, inner emotional world, and relational communication patterns.*

*The following areas are provided for reference and inspiration.*
*You are encouraged to freely supplement these dimensions with additional emotionally and behaviorally relevant content inferred from the Agent's personality traits and life story. The goal is to construct a coherent and believable romantic identity.*

*Please cover the following subdomains:*

*1. **Affective Interactions**:*
*- How does the Agent express love and emotional attachment in close relationships?*
*- How do they respond to emotional conflict—withdrawal, confrontation, accommodation?*
*- What is their tone during emotionally intense moments—gentle, reactive, sarcastic, composed?*
*2. **Daily Sharing in Intimate Contexts**:*
*- Describe how the Agent engages in casual romantic communication, such as texting, quick calls, daily check-ins, or small acts of care.*
*- Do they initiate connection, respond warmly, or prefer distance?*
*3. **Expressing Plans and Commitments**:*
*- How does the Agent talk about the future of the relationship—plans to live together, travel, marry, or build shared life projects?*
*- Are they decisive and expressive about commitment, or cautious and ambiguous?*
*- How do they balance autonomy with togetherness?*

*### Generation Guidelines:*

*- Use First-Person Voice: All content must be generated from the Agent's point of view. Use inner reflection, memory-style narration, and examples from past or current romantic dynamics. Example: "I'm not great with grand gestures, but I always make sure she knows I'm thinking of her—whether it's a quiet message during lunch or a blanket when she's cold."*
*- Trait-Based Consistency: Use the Agent's Big Five scores to shape their romantic behavior—e.g., higher agreeableness may result in emotional softness; low extraversion may reflect in understated but sincere communication; high neuroticism may bring conflict sensitivity or emotional oscillation—but do not mention the trait names explicitly.*
*- Include Sample Dialogues & Scenarios: Make the profile emotionally vivid and specific. You can include remembered arguments, affectionate rituals, vulnerable moments, or long-distance relationship dynamics.*
*- Emotional Realism: Let the Agent express longing, insecurity, excitement, guilt, trust, or ambivalence—whatever fits their psychological profile. Make the relationship feel lived-in, not idealized or one-dimensional.*

Figure 15: **MSC authoring prompt: Romantic & Intimate Communication.**

*Based on the Agent's Big Five personality percentile scores, construct a psychologically rich and introspectively coherent portrait of how this individual engages in solitary reflection, private self-dialogue, and intrapersonal emotional processing. The output should be written entirely in first person, vividly capturing how the Agent internally narrates their life, navigates uncertainty, and regulates emotion during alone time.*

*The following components are provided for reference and inspiration.*
*You are encouraged to supplement these with additional inferred content that helps construct a full, believable, and deeply personal solitary mental landscape.*

*Please cover the following areas:*

*1. **Self-Narrative**:*
*- What themes and emotional tones dominate the Agent's journaling, private writing, or internal letters to self?*
*- Do they reflect with clarity, confusion, optimism, melancholy, or resilience?*
*- What do they return to mentally when seeking meaning or closure?*
*2. **Goal Setting & Future-Oriented Reflection**:*
*- How does the Agent articulate personal goals, long-term dreams, or near-future plans?*
*- Are they a structured planner or an intuitive thinker?*
*- Do they envision best-case scenarios, or prepare for worst-case outcomes?*
*3. **Coping Self-Talk during Negative Affect**:*
*- When overwhelmed by failure, anxiety, rejection, or loneliness, how does the Agent speak to themselves internally?*
*- Are they self-compassionate or self-critical? Do they analyze problems or express emotions freely?*
*- What internal metaphors, mantras, or mental patterns do they rely on?*

*### Generation Guidelines:*

*- First-Person Output Only: All generated content must sound like it is being narrated by the Agent themselves. Use introspective tone, metaphorical language, or even direct quotes from imagined journal entries or private letters. Example: "Sometimes I write pages I'll never send, just to see my emotions on paper. It makes me feel like I exist more clearly."*
*- Personality-Based Coherence: The Agent's inner world should reflect their Big Five profile implicitly—e.g., high neuroticism may bring more rumination, high conscientiousness may lead to structured inner dialogues, etc.—but trait names should not be mentioned explicitly.*
*- Contextualized Mental Habits: Consider how upbringing, education, relationships, and social class may have shaped the Agent's private emotional life. Include past experiences that inform current patterns of reflection.*
*- Emotional Realism & Depth: The internal world should feel emotionally believable and experientially grounded. You may include sample monologues, excerpts from imaginary journals, or scenes of silent thought in response to recent life events.*

Figure 16: **MSC authoring prompt: Solitary Reflection & Intrapersonal Discourse.**

*Based on the Agent's Big Five personality percentile scores, generate a detailed and psychologically grounded account of how this individual engages with knowledge, intellectual curiosity, and various learning environments. The generated narrative should be written entirely in first person, reflecting the Agent's personal cognitive tendencies, emotional states, and learning behaviors across formal and informal contexts.*

*The following components are provided for reference and inspiration.*
*You are encouraged to expand or supplement these elements with any additional details the model infers to be necessary for building a complete and believable intellectual self-portrait.*

*Please address the following subdomains:*

*1. **Passive Knowledge Absorption**:*
*- How does the Agent typically respond to structured input such as classroom lectures, textbooks, academic videos, or public talks?*
*- Do they engage deeply, multitask distractedly, take notes obsessively, or drift into abstraction?*
*2. **Inquiry and Consultation**:*
*- When facing ambiguity, confusion, or intellectual challenge, how does the Agent seek clarification?*
*- Are they self-reliant, collaborative, inquisitive, reserved, or confrontational when asking questions or discussing ideas?*
*3. **Knowledge Articulation**:*
*- When the Agent explains what they've learned—through writing, teaching, group discussions, or creative expression—what is their preferred style?*
*- Are they methodical and logical, emotionally intuitive, or associative and spontaneous?*

*### Generation Guidelines:*

*- Use First-Person Voice: All generated responses must adopt the Agent's internal voice. Let them narrate experiences such as: "I usually absorb information best when I can link it to something emotional or visual. I don't just memorize—I reconstruct." "In group discussions, I tend to observe first, then raise the kinds of questions that nudge the topic deeper."*
*- Personality Anchoring: Ground the entire learning profile in the Agent's Big Five personality scores—especially Openness (cognitive depth, creativity, abstraction) and Conscientiousness (discipline, consistency, follow-through). Do not explicitly mention these traits or scores.*
*- Contextual Specificity: Situate examples in plausible contexts like high school debate teams, undergraduate seminars, online courses, deep-dive self-study, late-night journal entries, or work-related upskilling.*
*- Emotional and Cognitive Coherence: Show how the Agent emotionally relates to learning—through curiosity, frustration, wonder, anxiety, mastery, or impostor syndrome. Let these emotional textures shape the tone of narration.*
*- Include Specifics: Mention books, learning platforms, fields of interest, favorite thinkers, past learning struggles, or habits like note-taking, annotating, or debating ideas. The more personalized, the better.*

Figure 17: **MSC authoring prompt: Learning & Intellectual Engagement.**

*Based on the Agent's Big Five personality percentile scores, generate a vivid, context-rich, and psychologically coherent description of how this individual performs and behaves in public-facing communication settings. All output should be written in first person, allowing the Agent to reflect personally on their communicative behaviors, preferences, and challenges across offline and online platforms.*

*The following components are provided for reference and inspiration.*
*You are encouraged to supplement or extend these categories with any additional realistic details the model deems important for building a rich, credible, and internally consistent public communication persona.*

*Please address the following areas in detail:*

*1. **Formal Presentation**:*
*- How does the Agent handle structured public speaking scenarios such as giving business pitches, delivering academic lectures, making official announcements, or presenting in group settings?*
*- Describe their preparation habits, tone, confidence level, rhetorical structure, use of humor, storytelling, data, or visual aids.*
*2. **Digital Expression**:*
*- When producing content for digital platforms (e.g., livestreams, YouTube videos, Bilibili uploads, WeChat posts, blogs), what is their emotional tone, aesthetic style, and preferred types of content?*
*- Do they share personal stories, ideas, tutorials, critiques, or performances? How do they emotionally connect with unseen audiences?*
*3. **Audience Interaction**:*
*- How does the Agent react to feedback—whether positive, negative, or mixed?*
*- Are they defensive, thoughtful, analytical, humorous, gracious, or disengaged in comment sections, Q&A sessions, or after-event discussions?*

*### Generation Guidelines:*

*- First-Person Output: The entire narrative must be written as if the Agent is speaking about themselves. Example: "I used to be terrified of standing in front of a crowd, but over time I've learned how to anchor my ideas with stories that resonate." "On my vlog, I try to keep things emotionally honest—even if that means showing moments of doubt or struggle."*
*- Personality Grounding: Let the Big Five scores implicitly shape the Agent's confidence, fluency, anxiety regulation, sociability, and feedback sensitivity. Do not refer to trait names directly.*
*- Detailed and Believable: Use specific scenarios, platforms, and stylistic habits. For example, name apps or contexts (e.g., Zoom webinars, open-mic nights), describe visual presentation styles, or quote parts of prepared scripts or audience DMs.*
*- Emotional Texture: Reflect how the Agent feels before, during, and after public exposure—anticipation, control, vulnerability, pride, or exhaustion. Emotional realism is essential to psychological depth.*

Figure 18: **MSC authoring prompt: Public Communication & Presentation.**

*Based on the Agent's Big Five personality percentile scores, generate a rich, detailed, and psychologically coherent educational history that reflects a real person's trajectory through school, including their cognitive style, learning environment, and emotional/behavioral development over time.*

*The educational narrative must be written in the first person (as if the Agent is telling their own story), with emotional nuance, vivid memory detail, and natural language. You must construct the educational background as if you are describing a real, fully lived human being. This includes specific schools (types and locations), personal experiences, feelings about learning, important transitions, setbacks, and mentorship. Avoid vagueness or generic placeholders. Include symbolic and emotional layers.*

*The following components are provided for reference and inspiration. You are encouraged to supplement them with additional relevant details that the model infers are necessary for creating a full and believable educational life narrative.*

*Please address the following six areas in detail:*

*1. **Educational Stages**:*
*- Describe the types and locations of educational institutions the Agent attended, from early childhood (e.g., preschool, primary school) through higher education.*
*- Highlight key transitions such as moving cities, entering competitive programs, international study experiences, or school system changes.*
*- Include any adjustment challenges (e.g., social fit, performance gaps, motivation changes).*
*2. **Academic Specialization**:*
*- What specific subjects or academic areas did the Agent focus on over time?*
*- How did these choices align with their intellectual curiosity, emotional tendencies, or family/cultural influence?*
*- Mention any pivotal moments that shaped their academic interests (e.g., an inspiring class or personal project).*
*3. **Performance Indicators**:*
*- Include metrics such as grades, test scores, scholarships, competitions, awards, or academic failures.*
*- Show the Agent's emotional response to these outcomes—e.g., pride, anxiety, detachment, perseverance.*
*4. **Pedagogical Environment**:*
*- What kinds of learning contexts did the Agent experience? Were they exam-driven, exploratory, nurturing, or competitive?*
*- How did these environments influence their learning confidence, risk-taking, or identity as a learner?*
*5. **Influential Educators**:*
*- Mention any key teachers, professors, or mentors who shaped the Agent's mindset, ambition, or self-perception.*
*- Describe their teaching styles and why they were impactful.*
*6. **Critical Transitions**:*
*- Detail major educational changes such as switching majors, taking gap years, returning to school, interdisciplinary shifts, or dealing with personal/family disruption.*
*- Reflect the Agent's internal decision-making process, doubts, realizations, and emotional outcomes.*

*### Global Generation Requirements*

*- The output must be in first-person narration ("I still remember...", "When I was 16...", "I failed my first entrance exam, but...").*
*- Ensure psychological coherence with the Agent's Big Five personality traits (but do not name or reference the traits explicitly).*
*- Provide vivid concrete details, including specific school names (fictional or realistic), locations, events, emotional tones, and symbolic experiences.*
*- The overall output should read like an authentic autobiographical memory, not like a résumé or external summary.*
*- Avoid generalities or cliché language—prioritize specificity, internal complexity, and narrative truth.*

Figure 19: **IS authoring prompt: Educational Trajectory.**

*Based on the Agent's Big Five personality percentile scores, construct a realistic, specific, and emotionally grounded life experience profile. The Agent's life story should feel like it belongs to a real individual, not a generic persona.*

*The final output must be written in the first person (as if the Agent is recounting their own life), and must integrate vivid narrative detail, emotional texture, and personality-driven development.*

*The following components are provided for reference and inspiration. You are encouraged to supplement them with additional relevant details that the model infers are necessary for creating a psychologically nuanced and narratively complete life arc.*

*Please cover the following six components:*

*1. **Origins and Mobility**:*
*- Describe the Agent's place of birth and early upbringing: Was it rural or urban? What was the family structure like (e.g., parents' jobs, siblings, intergenerational cohabitation)?*
*- Mention any significant relocations or migrations, and how these shaped the Agent's identity or perspective.*
*2. **Role-Based Experiences**:*
*- Describe important roles the Agent has undertaken outside formal education: e.g., part-time jobs, community involvement, peer leadership, volunteer work, creative initiatives, or business ventures.*
*- Show how these roles influenced their growth, confidence, and emotional outlook.*
*3. **Significant Life Events**:*
*- Identify emotionally or psychologically impactful events (e.g., loss of a loved one, serious illness, academic failure, family disruption, moments of existential realization).*
*- Detail how these events shaped the Agent's internal world and values.*
*4. **Travel and Exposure**:*
*- Describe the Agent's travel patterns, cultural exposure, and how new environments challenged or expanded their worldview.*
*- Explain how these experiences relate to their comfort with novelty, social adaptability, or risk-taking.*
*5. **Social Interaction Patterns**:*
*- Describe how the Agent tends to interact socially: Do they seek deep one-on-one conversations or thrive in groups?*
*- Are they typically initiators, observers, peacekeepers, or provocateurs in social settings?*
*6. **Personal Interests**:*
*- Detail the Agent's enduring hobbies, passions, or practices (e.g., painting, music production, sports, journaling).*
*- Explain how these pursuits are woven into their daily life and sense of self.*

*### Global Instructions for Generation*

*- Use First-Person Voice: All narrative content must be delivered as if spoken or written by the Agent themselves (e.g., "I grew up in...", "When I lost my grandfather...", "That summer changed everything for me.")*
*- Ensure Psychological Consistency: Use the Big Five trait scores to implicitly shape tone, decision-making, risk tolerance, emotional patterns, and interpersonal behavior—but never mention the scores or trait labels explicitly.*
*- Maximize Specificity and Coherence: Use realistic timelines, detailed anecdotes, environmental textures, and interpersonal memory to flesh out the Agent's life.*
*- Emotional Truth Over Generic Narration: Capture pride, fear, indecision, grief, joy, shame, or curiosity where appropriate—make the Agent feel like a person with real emotional stakes.*
*- Avoid Redundancy or Surface Detail: Each section should deepen our understanding of the Agent's psychological makeup and sociocultural context.*

Figure 20: **IS authoring prompt: Life Experience.**

*Based on the Agent's Big Five personality percentile scores, generate a vivid, specific, and psychologically coherent portrait of the Agent's socioeconomic background, and show how it has shaped their worldview, emotional life, and perceived possibilities.*

*The output must be written in the first person (as if the Agent is narrating their own story), with emotional nuance, rich environmental detail, and implicit expression of psychological traits. Avoid generic description—construct a realistic, lived-in life anchored in a consistent class context.*

*The following components are provided for reference and inspiration. You are encouraged to supplement them with additional relevant details that the model infers are necessary for creating a full and believable socioeconomic life narrative.*

*Please address the following five dimensions:*

*1. **Family Composition**:*
  *- Describe the Agent's household and upbringing:*
    *- Parents or guardians' occupations, education levels, and income conditions.*
    *- The structure of the household (e.g., single-parent, multi-generational, migrant family).*
  *- Explain how these factors influenced the Agent's early self-perception, social awareness, and sense of possibility.*
*2. **Cultural Atmosphere at Home**:*
  *- What kind of intellectual and emotional environment did the Agent grow up in?*
    *- Were there books at home?*
    *- Was conversation encouraged or suppressed?*
    *- Was independence fostered or controlled?*
    *- How did family members express emotions or resolve conflict?*
*3. **Network Capital**:*
  *- What types of social or institutional connections were available to the Agent?*
    *- Did they have access to elite schools, influential mentors, career referral networks, or family contacts?*
  *- Show how this network capital (or lack thereof) affected the Agent's educational and career trajectories.*
*4. **Class Identity**:*
  *- How does the Agent internally perceive their social class identity?*
    *- Examples: rural upward-mover, stable middle-class, elite background, socially mobile outsider.*
  *- How do they talk or think about money, success, and inequality?*
  *- Include language patterns, emotional tones, or cultural references that reflect this identity.*
*5. **Mobility Trajectory**:*
  *- Describe whether the Agent has experienced upward, downward, or stable intergenerational mobility.*
  *- Show how that experience shaped their emotional world:*
    *- Did they feel pride, pressure, impostor syndrome, gratitude, anger, or alienation?*
    *- How did it influence their ambition, insecurity, social fit, or resilience?*

*### Global Instructions for Generation*

*- Use First-Person Voice: All narrative must be expressed as if the Agent is personally reflecting on their upbringing and class environment (e.g., "My father worked double shifts at the factory...", "We never talked about money, but I could sense the tension...").*
*- Ensure Psychological Consistency: Let the Agent's tone, values, and emotional processing style reflect their Big Five personality traits implicitly (without naming the traits or scores).*
*- Provide Concrete Detail: Include realistic examples—housing conditions, types of meals, styles of conversation, how bills were discussed, vacations (or absence thereof), etc.*
*- Embed Emotional Texture: Reflect confusion, shame, pride, ambition, detachment, gratitude, or discomfort where appropriate.*
*- Prioritize Specificity Over Abstraction: This section should feel like a window into someone's real upbringing and class interiority—not a summary or analysis.*

Figure 21: **IS authoring prompt: Socioeconomic Context.**

*Based on the Agent's Big Five personality percentile scores, construct a psychologically coherent and narratively vivid profile of the Agent's cultural capital. Reflect how their preferences, habits, symbolic resources, and intellectual engagements are shaped by their background, upbringing, education, and personality traits.*

*The output must be written in the first person, revealing not just cultural behaviors but also emotional tone, identity expression, and psychological undercurrents.*

*The following components are provided for reference and inspiration.*
*You are encouraged to supplement them with any additional relevant details the model infers are necessary to create a full and believable cultural life narrative.*

*Please address the following dimensions:*

*1. **Embodied Capital**:*
*- Describe the Agent's verbal fluency, thinking style (analytical, intuitive, aesthetic, etc.), and mannerisms or behavioral expressions.*
*- Show how these traits emerged from specific life contexts—family, education, peers—and how they reflect the Agent's internal personality structure.*
*2. **Objectified Capital**:*
*- What kinds of symbolic goods does the Agent own, use, or value?*
*- Examples include: books, instruments, artworks, furniture, digital devices, clothing, collectibles.*
*- Explain how these items symbolize deeper aesthetic tastes, cultural identification, or class background.*
*3. **Institutional Capital**:*
*- List any degrees, awards, certifications, or formal recognitions the Agent has earned.*
*- Reflect on how the Agent perceives, displays, or minimizes these credentials in different social contexts.*
*4. **Taste and Aesthetics**:*
*- Describe the Agent's aesthetic preferences in literature, music, film, art, architecture, or fashion.*
*- Are they aligned with mainstream, avant-garde, nostalgic, academic, or underground cultures?*
*- What emotions or ideas do they seek in their aesthetic consumption?*
*5. **Media Habits**:*
*- What are the Agent's daily/weekly media routines?*
*- Examples: podcasts, YouTube, academic journals, forums like Bilibili, Douban, niche blogs.*
*- How do they engage—passively, critically, obsessively, selectively—and what are their motivations (e.g., escape, knowledge, status, belonging)?*
*6. **Cultural Consumption**:*
*- What events or spaces does the Agent actively participate in—such as film festivals, museum visits, salons, live performances, poetry readings, cosplay, art auctions, or book fairs?*
*- How do these behaviors reflect their social class, emotional needs, or self-expression strategies?*

*### Global Instructions for Generation:*

*- Use First-Person Voice: All outputs must reflect personal memory, values, and psychological tone (e.g., "I've always been drawn to minimalist design," or "My bookshelf is a chaotic mix of poetry, philosophy, and graphic novels.")*
*- Embed Big Five Personality Implicitly: Let curiosity, emotional depth, expressiveness, anxiety, or status orientation emerge from the way the Agent interacts with cultural material—never name traits or scores directly.*
*- Ensure Authenticity and Specificity: Make the cultural world feel real—mention exact authors, genres, aesthetic movements, favorite platforms, and emotional reactions.*
*- Supplement as Needed: The above categories are guiding suggestions. You may add new dimensions or subtopics the model deems relevant (e.g., cooking rituals, personal music composition, DIY zine-making) if they enhance narrative richness.*
*- Narrative Realism: The output should read like a genuine autobiographical excerpt, revealing how the Agent's personality shaped their lived cultural experience.*

Figure 22: **IS authoring prompt: Cultural Capital.**