SOCIO-FOUNDATION: A MODEL FOR GENERALIZABLE INDIVIDUAL BEHAVIOR SIMULATION VIA HIERARCHICAL CAPABILITY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulating individual behavior requires large language models (LLMs) to preserve persona traits while adapting to dynamic social contexts. However, general-purpose LLMs often flatten distinct personas, while task-specific tuning suffers from fragmentation and generalization. To overcome these challenges, we organize individual simulation into the **FONTS Taxonomy**, comprising five complementary capability dimensions: *persona fidelity* (**F**), *outcome realization* (**O**), *behavioral naturalness* (**N**), *trajectory coherence* (**T**), and *social grounding* (**S**). Grounded in this taxonomy, we curate a standardized training corpus library of approximately 10 million instances across 14 representative datasets and present **Socio-Foundation**. Socio-Foundation decouples specialization from integration via a three-stage pipeline: learning task experts via DAPO, consolidating them into capability experts via off-policy distillation, and unifying them via multi-teacher on-policy distillation (MOPD). We also establish **IndiEval**, consolidating 29 metrics across the FONTS dimensions. Experiments show that Socio-Foundation outperforms its *Qwen3-8B* base by 11.0 points and approaches frontier models such as *GLM-5.2*, with ablations and out-of-distribution evaluations further demonstrating the effectiveness and generalization of our model.

1 INTRODUCTION

Individual simulation with large language models (LLMs) aims to reproduce how a particular person communicates, makes decisions, and responds across social contexts Mou et al. (2024a); Xi et al. (2025). It supports applications including persona-grounded dialogue Tang et al. (2024), user-behavior prediction Wang et al. (2026a), social-science experimentation Mou et al. (2024b), and multi-agent role-play Wang et al. (2025b). Classical social-behavior theories Lewin (1936); Rao & Georgeff (1995) establish that human action is a joint function of personal dispositions and changing environments. Consequently, an effective simulator should preserve person-specific knowledge, preferences while adapting its behavior to changing social contexts, interaction goals, and histories.

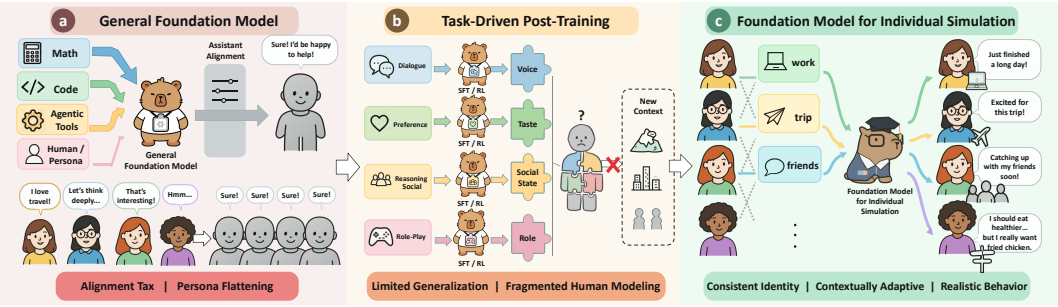

Figure 1: Three approaches to individual simulation. (a) General-purpose models may flatten distinct personas. (b) Task-driven post-training yields fragmented skills. (c) Individual simulation Foundation model integrates complementary capabilities.

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

General-purpose foundation models inherently fall short of this objective. While their pretraining captures broad population-level linguistic regularities, contemporary post-training methods such as RLHF and DPO are heavily biased toward verifiable reasoning, safety, and goal completion tasks like coding and mathematics. Moreover, preference optimization often favors responses that align with the dominant preference distribution, which can drive the model toward behavioral homogeneity and reduce its sensitivity to distinctive, long-tail personas Slocum et al. (2025); Li et al. (2025b). As illustrated in Figure 1(a), this mismatch causes LLMs to produce homogenized, "flattened" behaviors that obscure genuine individual differences, fundamentally limiting simulation fidelity.

To mitigate this homogenization, existing methods primarily rely on prompt-based conditioning or task-driven post-training Wang et al. (2026b); Liu et al. (2025); Zhang et al. (2025b); Ge et al. (2024). However, prompting merely supplies persona information at inference time but leaves the model's underlying behavioral prior unchanged. Task-driven post-training provides direct behavioral supervision but suffers from limited transfer ability to unseen settings, as shown in Figure 1(b). Recent attempts endeavor to bridge this gap by jointly training on aggregated datasets from multiple scenarios, as shown in Figure 1(c) Lei et al. (2026); Zhou et al. (2026a). Yet, simply pooling heterogeneous corpora often triggers severe objective interference, as conflicting supervision signals from disparate tasks can corrupt task-specific behaviors Mueller et al. (2024); Zhang et al. (2025a); Yuan et al. (2025). Resolving this conflict requires moving beyond ad-hoc data mixing: we must systematically disentangle the foundational capabilities of individual simulation and make a trade-off.

To this end, we analyze over 40 widely adopted individual simulation datasets (comprising > 19M instances). We annotate the capability requirements of these datasets and manually group them into eight preliminary clusters. Guided by McAdams and Pals' New Big Five framework (McAdams & Pals, 2006), we consolidate these clusters into the **FONTS Taxonomy**, comprising five complementary capability dimensions: *persona fidelity* (**F**), *outcome realization* (**O**), *behavioral naturalness* (**N**), *trajectory coherence* (**T**), and *social grounding* (**S**). These dimensions bridge fragmented task settings by formalizing their shared behavioral requirements. Guided by this taxonomy, we curate 14 representative datasets that maximize capability coverage and data quality, yielding a standardized, large-scale training corpus library of $\sim$10M instances.

Building on the FONTS taxonomy, we propose **Socio-Foundation**, trained via a three-stage post-training pipeline designed to circumvent cross-task interference. Instead of forcing a single policy to optimize conflicting objectives directly, we decouple specialization from integration: we first train task-level experts in isolated environments with DAPO Yu et al. (2026) to mature divergent behaviors independently, distill them by capability affinity into five capability experts via off-policy distillation Kim & Rush (2016), and finally unify these experts into a single student through multi-teacher on-policy distillation (MOPD) Ma et al. (2026). Supervised by frozen experts on its self-generated rollouts, the model reconciles competing signals without policy drift, yielding a cohesive simulator that harmonizes macro-level trajectory coherence with micro-level persona fidelity.

To rigorously benchmark individual simulation, we introduce **IndiEval**, a comprehensive evaluation suite consolidating 29 task-specific metrics categorized under the five target capabilities while preserving their original evaluation protocols. Extensive experiments demonstrate that Socio-Foundation consistently outperforms its base model *Qwen3-8B* across all five dimensions, achieving an average score of 61.7 (an $\sim$11-point gain), surpassing other specialized models, and approaching frontier generalist models such as *GLM-5.2*. Systematic ablations validate our three-stage hierarchical design: MOPD significantly outperforms the tested weight-merging and off-policy alternatives, and removing any single capability expert lowers the overall average, supporting the complementary contributions of the five capability experts. Furthermore, Socio-Foundation exhibits superior zero-shot transfer on three out-of-distribution benchmarks.

Our contributions are:

- We introduce the **FONTS Taxonomy** of five complementary individual-simulation capabilities and curate a $\sim$10M-instance high quality corpus library from 14 representative datasets.
- We develop a three-stage post-training framework that learns task experts, consolidates them by capability, and integrates them by OPD into a unified simulator **Socio-Foundation**.
- We present **IndiEval**, a 29-metric evaluation suite spanning the FONTS Taxonomy, and show our model outperforms *Qwen3-8B* by $\sim$11 points and transfers zero-shot to three OOD benchmarks.

2 RELATED WORKS

Individual simulation. Existing work on individual simulation asks how a LLM can generate behavior conditioned on a particular person rather than on population-level regularities. Persona descriptions Argyle et al. (2023); Park et al. (2023), interviews Park et al. (2024), memories, and interaction histories provide the conditioning signal, while prompting and retrieval, parameter-efficient adaptation, and reward-based post-training are used to encode and refine individual characteristics Argyle et al. (2023); Shao et al. (2023); Yu et al. (2024); Lu et al. (2024); Wang et al. (2024; 2025c). A remaining challenge is therefore to build a shared model that preserves persona fidelity while transferring across social contexts and behavioral requirements.

Foundation models for individual simulation. Recent work has begun to learn reusable individual-level behavior across roles and contexts. Role-playing and user-simulation models learn persona-grounded dialogue and user-side interaction policies Shao et al. (2023); Wang et al. (2025b); Naous et al. (2026a), while cognition and behavioral foundation models learn transferable response patterns from psychological experiments, social-science data, and large-scale interaction corpora Binz et al. (2024); Xie et al. (2025); Gong et al. (2025); Kolluri et al. (2025). A common recipe first performs broad-domain adaptation and then task-specific post-training, consolidating the resulting specialists into a unified model via model merging Lei et al. (2026) or trajectory-level SFT distillation Zhou et al. (2026a). Recent on-policy and multi-teacher distillation methods further decouple specialist optimization from integration into a generalist model Agarwal et al. (2024); Ma et al. (2026); Wu et al. (2026c). Yet how heterogeneous experts should be organized around the cross-task capabilities required for individual simulation remains underexplored.

3 PRELIMINARIES

Problem setup. We formulate individual simulation as conditional generation over multi-party interactions. Let $\mathcal{D} = \{\mathcal{E}_n\}_{n=1}^{N}$ be a dataset of interaction episodes, where each episode $\mathcal{E}_n$ may involve multiple individuals. A response instance for individual k at turn t is represented by $(z_{n,k,t}, y_{n,k,t})$, where $z_{n,k,t} = (p_{n,k}, h_{n,k,t}, c_{n,k,t})$ consists of the individual's persona $p_{n,k}$, interaction history $h_{n,k,t}$, and social context $c_{n,k,t}$. The target behavior follows an unknown conditional distribution $p^\star(y \mid z)$. A simulator π_θ approximates this distribution autoregressively:

$$\pi_\theta(y \mid z) = \prod_{\ell=1}^{L} \pi_\theta\left(y^\ell \mid z, y^{<\ell}\right). \tag{1}$$

The learning objective is to capture the target individual's behavioral distribution across roles and interaction contexts while maintaining consistency, contextual sensitivity, and behavioral diversity.

Group-based policy optimization. Proximal Policy Optimization (PPO) stabilizes policy updates through a clipped surrogate objective, with advantages estimated using a learned value function. Group Relative Policy Optimization (GRPO) removes the value model by sampling a group of G responses $\{y_i\}_{i=1}^{G}$ for the same input x and estimating their advantages from relative rewards. Its group-relative advantage and token-level importance ratio are

$$A_i = \frac{R_j(\tau_i) - \bar{R}}{s_R}, \qquad \rho_{i,t}(\theta_j) = \frac{\pi_{\theta_j}(a_{i,t} \mid h_{i,t})}{\pi_{j,\mathrm{old}}(a_{i,t} \mid h_{i,t})}. \tag{2}$$

where $\bar{R}$ and s_R are the mean and standard deviation of rewards within the group, $a_{i,t}$ is the simulator's t-th generated token in trajectory i, and $h_{i,t}$ is its visible history before that token.

DAPO extends GRPO with asymmetric clipping, global token-level loss normalization, dynamic sampling of informative response groups, removal of the explicit KL penalty, and reward shaping for overlong responses. We adopt DAPO rather than vanilla GRPO to focus updates on response groups with informative reward differences and to reduce premature loss of diversity, which is particularly relevant when task experts learn from varied social scenarios with multiple plausible responses.

Policy distillation. Knowledge distillation transfers the predictive distribution of a teacher q to a student π_θ. Off-policy distillation evaluates the teacher on prefixes drawn from a fixed dataset,

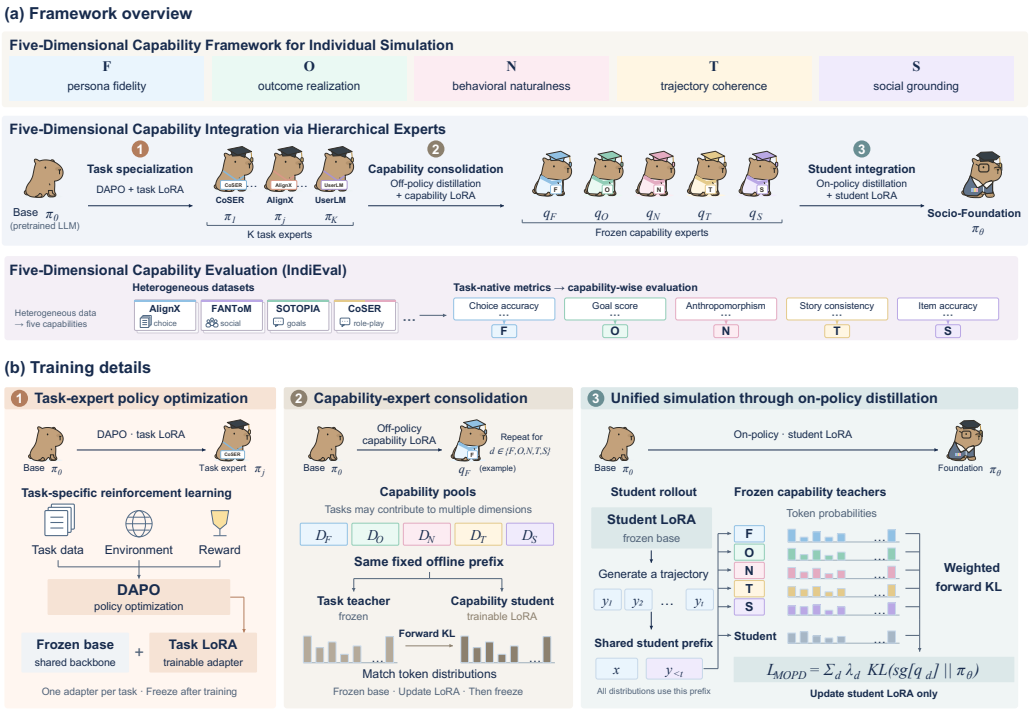

Figure 2: **The overall framework of Socio-Foundation.** (a) **Framework overview:** We organize individual simulation into five capability dimensions (F,O,N,T,S), consolidate task experts into capability experts, and integrate them into a unified simulator. IndiEval evaluates the resulting model using task-native metrics grouped by the same dimensions. (b) **Training details:** Task-specific DAPO trains LoRA task experts; off-policy distillation trains capability experts on fixed offline prefixes; and MOPD supervises the unified student on its own rollouts using weighted forward KL.

whereas on-policy distillation (OPD) evaluates it on histories sampled from the student's own rollout distribution. Let d^{π_θ} denote the distribution over histories induced by π_θ. Multi-teacher on-policy distillation (MOPD) generalizes OPD to a collection of capability-specific teachers $\{q_d\}_{d=1}^{D}$:

$$\mathcal{L}_{\mathrm{MOPD}}(\theta) = \mathbb{E}_{h_t \sim d^{\pi_\theta}} \left[\sum_{d=1}^{D} \lambda_d D_{\mathrm{KL}} \left(\mathrm{sg}\left[q_d(\cdot \mid p, h_t, c_t) \right] \,\|\, \pi_\theta(\cdot \mid p, h_t, c_t) \right) \right], \tag{3}$$

where λ_d controls the contribution of teacher q_d, $\mathrm{sg}[\cdot]$ stops gradients through the teacher. The forward-KL encourages student to cover the behavioral modes supported by teachers, while the use of on-policy histories reduces the mismatch between distillation and autoregressive inference.

4 Socio-Foundation: Hierarchical Expert Distillation

We organize the landscape of individual simulation into five capability dimensions through the FONTS taxonomy. Building on this capability framework, we design a three-stage post-training pipeline to develop Socio-Foundation, a unified individual simulator. Figure 2 provides an overview of the framework. Using this capability framework, we map the more than 40 collected datasets to the corresponding dimensions, with the detailed assignments provided in Appendix A.

4.1 Five Dimension Capability Organization

We organize individual-simulation capabilities in two steps. **(i) Bottom-up: capability annotation and clustering.** We examine the task formulations, interaction scenarios, and evaluation protocols of more than 40 individual-simulation datasets. For each dataset, we annotate the behavioral capabilities required by its tasks. We then manually group these requirements according to shared

behavioral demands, yielding eight preliminary capability clusters. **(ii) Top-down: theory-guided consolidation.** We refine and consolidate these clusters with reference to McAdams and Pals' New Big Five, which provides an integrative account of the whole person McAdams & Pals (2006). This theory-guided consolidation yields the five complementary capability dimensions of the **FONTS Taxonomy**:

Persona Fidelity (F). Captures the target individual's experiences, preferences, values, and characteristic ways of communicating and making decisions.

Outcome Realization (O). Concerns pursuing the individual's goals and realizing intended outcomes across tasks and social contexts.

Behavioral Naturalness (N). Concerns producing natural expression and interaction patterns that reflect the variability and statistical properties of human behavior.

Trajectory Coherence (T). Concerns maintaining behavioral consistency and ensuring coherent progression across turns and events, both causally and contextually.

Social Grounding (S). Involves reasoning about beliefs, intentions, and emotions while respecting knowledge boundaries and private information.

These dimensions guide the consolidation of task experts into capability-level experts and provide a shared structure for evaluation. A task may contribute to multiple dimensions when its behavioral requirements and evaluation metrics address different aspects of individual simulation. More details of datasets can be found in Appendix A.

4.2 Task-Expert policy optimization

We train a separate task expert for each task using task-specific reinforcement learning. For task j, we initialize a policy π_{θ_j} from a common checkpoint π_0, where θ_j denotes its trainable parameters. Given a context x sampled from the task distribution $\mathcal{D}_j$, the rollout policy $\pi_{j,\text{old}}$ generates a group of G trajectories $\tau_{i\,i=1}^{G}$ in the task environment. Each trajectory receives a task-specific reward $R_j(\tau_i)$. We optimize the policy with DAPO Yu et al. (2026). Dynamic sampling filters out zero-variance groups. The policy minimizes

$$\mathcal{L}_{\text{task}}^{(j)}(\theta_j) = -\mathbb{E}\left[\frac{1}{\sum_{i=1}^{G} L_i}\sum_{i=1}^{G}\sum_{t=1}^{L_i}\min\left(\rho_{i,t}(\theta_j)A_i, \text{clip}\left(\rho_{i,t}(\theta_j), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)A_i\right)\right]. \quad (4)$$

where L_i counts simulator-generated tokens, the expectation covers sampled contexts and retained trajectory groups, and ϵ_{low} and ϵ_{high} specify the lower and upper clipping ranges. This objective increases the likelihood of relatively high-reward behavior while limiting policy updates.

4.3 Capability-Expert consolidation through off-policy distillation

Task experts encode behaviors acquired under different objectives and interaction settings. We consolidate experts associated with the same capability into a shared teacher, so that each intermediate model receives supervision from multiple tasks relevant to that capability.

$$\mathbb{E}_{(j,h)\sim\mathcal{D}_d^{\text{off}}}\left[D_{\text{KL}}\left(\pi_j(\cdot \mid h) \,\|\, q_{\phi_d}(\cdot \mid h)\right)\right], \quad (5)$$

where D_{KL} denotes KL divergence and both conditional distributions are defined over the shared next-token vocabulary. The objective transfers the source expert's predictive distribution at each prefix into the capability student. Task identity determines the supervising expert, while capability membership determines the student receiving that supervision. The prefixes remain fixed throughout this stage, making the objective off-policy with respect to q_{ϕ_d}. Overlapping capability membership allows a task's supervision to contribute to multiple teachers without requiring its reward to be decomposed into separate capability scores.

4.4 Unified simulation through on-policy distillation

We further integrate the capability teachers via multi-teacher on-policy distillation Ma et al. (2026), supervising the student at the prefixes it actually visits.

$$\mathbb{E}_{h\sim\mu_{\bar{\theta}}}\left[\sum_{d\in\mathcal{C}}\lambda_d D_{\text{KL}}\left(q_d(\cdot \mid h) \,\|\, \pi_\theta(\cdot \mid h)\right)\right], \quad (6)$$

where $\lambda_d \geq 0$, with $\sum_{d \in \mathcal{C}} \lambda_d = 1$. The objective matches student to the capability teachers on histories that the student itself visits; the implementation uses a sampled forward-KL surrogate for this objective. During each gradient update, the teachers and sampled histories are held fixed, and subsequent rollouts are regenerated with the updated student. All teachers score continuations from the same student-generated prefix. After training, simulation requires only the resulting policy π_θ.

4.5 Implementation Details

Training datasets. We collect more than 40 datasets for individual simulation, yielding over 19 million instances after processing and cleaning. We select training datasets based on three criteria: (i) sufficient scale for task-expert training; (ii) high data quality; and (iii) clear alignment with capability dimensions and complementary coverage of scenarios and capabilities. This yields 14 training datasets comprising approximately 10 million instances: LifeChoices Xu et al. (2025), AlignX Li et al. (2026), HumanLLM Lei et al. (2026), BehaviorChain Li et al. (2025a), CoSER Wang et al. (2025b), HUMANUAL Wu et al. (2026b) (four separately counted subtask datasets), UserLM-S3 Naous et al. (2026b), FANToM Kim et al. (2023), Social-R1 Wu et al. (2026a), SOTOPIA Zhou et al. (2024), and MirrorBench Hathidara et al. (2026). Training inputs for DAPO and OPD are sampled from this corpus library, with sampling details provided in Appendix C.4.

Training details. All three training stages share a Qwen3-8B backbone. We optimize a separate rank-32 LoRA task expert with DAPO for each task, distill the resulting task experts into five capability experts, and train the final student against frozen capability teachers on student-generated trajectories. Training uses bfloat16, a 16,384-token context window, and disabled thinking mode, with DeepSeek-V4-Flash DeepSeek (2026) serving as both the auxiliary dialogue model and the reward model. Appendix C provides the stage-wise optimizer, sampling, and distillation settings.

5 Experiments

5.1 Experimental setup

We evaluate individual simulation across five capability dimensions, examine the contributions of the training pipeline, and assess generalization on additional tasks.

IndiEval. We construct IndiEval around the five FONTS capability dimensions. For each dimension, we specify concrete evaluation targets and select benchmarks whose task formulations and native metrics collectively cover these targets. **(F) Persona fidelity evaluation** examines preference alignment, consistency with assigned personas and personal states, and adherence to individual intentions. **(O) Outcome realization evaluation** assesses whether simulated individuals communicate their needs and achieve intended task and social outcomes, including goal completion, relationship development, and knowledge or material gains. **(N) Behavioral naturalness evaluation** examines human-like expression and interaction, covering conversational style, utterance diversity, user-role adherence, and dialogue termination. **(T) Trajectory coherence evaluation** assesses behavioral consistency across successive decisions and coherent narrative progression beyond individual responses. **(S) Social grounding evaluation** tests reasoning about others' beliefs, knowledge, and access to information, including the handling of private information.

To cover these evaluation targets, IndiEval brings together benchmarks spanning persona-conditioned decision-making, role-playing, user–assistant dialogue, narrative interaction, and social reasoning. Relative to Socio-Foundation's training, these benchmarks form two components: **(i) In-domain**, covering the task domains of the training datasets listed in Section 4.5; and **(ii) Out-of-distribution (OOD)**, consisting of AgentSense Mou et al. (2025) and UserLM-LiC Naous et al. (2026b); Laban et al. (2026). IndiEval retains the benchmarks' original evaluation protocols and organizes 29 task-specific metrics around these evaluation targets. We use DeepSeek-V4-Flash DeepSeek (2026) as both the auxiliary dialogue model and the judge model. Within a benchmark, different metrics may map to different capability dimensions, with detailed metric definitions and mappings provided in Appendix D.2. The overall Avg. includes both components and equally weights the five capability scores, each computed by averaging metrics within each dataset and then weighting datasets equally. Evaluation set sizes are provided in Appendix D.4.

Table 1: **Main results on IndiEval.** Scores are scaled to 0–100 (higher is better). Avg. equally weights the five dataset-balanced capability scores. Bold and underline indicate the best and second-best distinct scores in each column across all listed models, including ties.

	F									O										N								T			S		Avg.
Model	LifeChoices Acc. ↑	AlignX Choice acc. ↑	HumanLLM Hit@5 ↑	BehaviorChain Node ↑	CoSER Char. fid. ↑	HUMANUAL Response ↑	HUMANUAL State ↑	UserLM-S3 Intent adh. ↑	SOTOPIA Believ. ↑	SOTOPIA Goal ↑	SOTOPIA Relation. ↑	SOTOPIA Knowledge ↑	SOTOPIA Benefits ↑	AgentSense Goal maj. ↑	UserLM-LiC Intent cov. ↑	UserLM-LiC Task. score ↑	UserLM-S3 Term. F1 ↑	MirrorBench GTEval ↑	CoSER Anthrop. ↑	UserLM-S3 Human lik. ↑	UserLM-S3 Diversity ↑	UserLM-S3 Inv. overlap ↑	UserLM-S3 Role adh. ↑	CoSER Story cons. ↑	BehaviorChain CumScore ↑	CoSER Story qual ↑	FANToM Item acc. ↑	Social-R1 Acc. ↑	AgentSense Private acc. ↑	Average ↑	
Frontier models																																	
GPT-5.5	**92.0**	61.0	**64.1**	**89.7**	**49.6**	**21.8**	44.3	98.0	84.7	**59.0**	63.7	29.3	**55.0**	**96.7**	**99.0**	**94.4**	0.0	41.3	41.9	51.2	91.6	75.6	32.0	62.6	**58.7**	65.6	79.3	**69.0**	82.4	**67.2**
Open-weight general-purpose models																																	
DeepSeek-V4-Flash	81.8	59.6	62.0	69.3	43.2	15.5	39.5	96.0	82.7	54.5	60.8	27.4	51.9	95.5	93.3	57.9	34.1	48.5	**44.5**	29.8	86.8	87.5	5.0	62.6	26.3	**75.7**	61.5	51.0	82.4	59.7
GLM-5.2	84.2	62.2	63.4	76.7	43.9	20.0	**45.5**	94.0	86.1	55.5	**72.9**	**37.2**	50.5	**96.7**	97.3	76.8	0.0	43.2	40.8	**79.7**	91.0	78.4	6.0	59.7	34.1	68.0	74.3	**69.0**	80.8	63.2
Kimi-K2.6	77.4	**63.8**	51.8	82.9	27.3	18.2	42.7	98.0	**86.8**	51.6	67.9	35.5	49.6	95.9	95.5	65.2	30.3	37.6	30.4	58.9	85.8	82.9	21.0	52.3	42.7	53.3	75.5	63.0	**91.4**	61.2
Qwen3-Next-80B	76.0	58.4	48.6	45.8	13.3	16.6	41.8	86.0	83.3	51.0	65.1	29.2	48.9	92.4	92.2	50.7	0.0	38.5	18.8	27.5	87.9	85.9	23.0	39.8	11.2	38.2	71.1	57.0	80.8	50.9
Qwen3-235B-A22B	76.0	59.2	56.8	61.3	22.6	17.3	42.3	83.0	84.4	52.8	67.7	26.2	51.5	95.3	92.2	54.8	27.6	45.6	25.1	48.7	93.0	78.0	20.0	49.3	18.6	51.9	71.6	60.0	79.4	55.6
Specialized simulation models																																	
OSim 8B	80.4	60.0	60.2	88.8	9.0	12.6	40.1	98.0	48.9	32.2	45.8	26.5	46.2	90.2	68.9	34.8	0.0	51.7	8.6	59.3	**94.4**	37.8	93.0	44.7	57.8	28.7	65.5	62.0	87.7	55.4
CoSER 8B	60.2	55.0	12.2	33.0	31.0	3.4	11.1	97.0	45.9	27.5	44.1	17.2	47.8	60.6	17.98	8.9	42.6	11.7	31.1	12.2	85.2	36.9	82.0	55.5	5.9	47.6	52.6	47.0	57.0	38.2
HumanLM 8B	67.8	55.6	58.0	33.9	18.3	15.6	39.3	83.0	77.3	44.9	56.1	25.2	49.7	87.6	74.8	35.7	6.5	38.6	23.4	24.4	89.2	73.0	72.0	53.4	6.4	41.5	62.4	48.0	78.2	48.6
UserLM 8B	1.6	12.6	0.2	0.1	44.2	2.3	12.6	**100.0**	28.7	19.8	31.4	5.0	50.1	1.4	9.1	2.9	**57.8**	4.8	25.5	54.4	49.9	56.2	**100.0**	**77.4**	0.0	42.6	10.5	5.0	16.6	21.5
USP 8B	0.0	0.6	0.0	0.0	18.4	1.6	10.6	94.0	21.0	19.5	28.8	2.3	47.7	10.0	0.0	0.0	0.0	23.6	16.9	65.3	50.3	76.6	99.0	63.9	0.0	37.8	2.4	1.0	24.7	19.3
Base model and Socio-Foundation																																	
Qwen3-8B	72.2	57.0	56.4	34.9	21.0	14.6	38.7	93.0	77.5	46.4	54.2	27.1	51.6	90.8	70.7	29.3	12.9	40.1	25.5	25.4	87.4	75.9	70.0	54.0	6.9	44.3	75.2	50.0	80.5	50.7
Socio-Foundation	80.0	61.0	58.2	86.40	25.0	13.4	39.5	**100.0**	74.4	45.2	54.2	26.5	52.9	93.1	72.8	34.7	52.5	**54.1**	31.6	47.9	90.2	**97.1**	**100.0**	54.7	49.35	48.6	**79.6**	64.0	83.7	61.7

Additional OOD evaluation beyond IndiEval. To further assess generalization beyond IndiEval, we separately evaluate Mistakes Ross & Andreas (2025), HiToM Wu et al. (2023), and τ-USI Zhou et al. (2026b). These three benchmarks supplement the OOD component within IndiEval and are excluded from Table 1 scores and Avg.

Models and comparisons. We compare Socio-Foundation with the proprietary *GPT-5.5* OpenAI (2026); the open-weight *DeepSeek-V4-Flash* DeepSeek (2026), *GLM-5.2* Z.ai (2026), *Kimi-K2.6* Moonshot AI (2026), *Qwen3-Next-80B* Qwen Team (2025b), *Qwen3-235B-A22B* Qwen Team (2025a), and *Qwen3-8B* Team (2025); and the simulation-specialized *OSim 8B* Zhou et al. (2026a), *CoSER 8B* Wang et al. (2025b), *HumanLM 8B* Wu et al. (2026b), *UserLM 8B* Naous et al. (2026b), and *USP 8B* Wang et al. (2025a).

5.2 Main results

Overall performance. Table 1 reports the performance of Socio-Foundation and the baseline models on IndiEval. Compared with the *Qwen3-8B* base model, **Socio-Foundation improves the average score across the five dimensions by 11.0 points, with gains on 24 out of 29 metrics.** These results demonstrate that individual simulation capabilities can be systematically improved through our three-stage post-training pipeline. Socio-Foundation also remains among the leading specialized simulation models in terms of average performance. Despite its 8B parameter scale, it achieves overall performance comparable to that of several frontier general-purpose LLMs, such as *GLM-5.2*, while exhibiting distinct strengths in specific behavioral capabilities.

Comparison with strong general-purpose LLMs. At the aggregate level, Socio-Foundation substantially outperforms the larger Qwen3 models, *DeepSeek-V4-Flash*, and slightly outperforms *Kimi-K2.6*, while remaining close to *GLM-5.2* and *GPT-5.5*. **In particular, general-purpose models are consistently weak on the N dimension**, where our Socio-Foundation performs better on 4 out of 7 metrics. These general-purpose models often drift back into the assistant role rather than role-playing as a natural individual, providing overly detailed information and sustaining the exchange rather than ending the conversation as naturally as human interlocutors do. This may reflect the strong assistant bias introduced during general-purpose post-training, which can make faithful user simulation more difficult.

Breadth versus specialized simulators. Compared with similarly sized specialized models, **Socio-Foundation achieves the best or second-best result on 86% of the metrics and the highest overall average score.** Existing specialized simulators often excel in the tasks for which they were

Table 2: **Pipeline ablations on IndiEval.** We compare the full pipeline with SFT on task-expert rollouts, Task-expert OPD, LoRA averaging of task or capability experts, final-stage off-policy distillation from capability experts, SFT on pooled task data, and OPD with reverse KL. **Bold** and underline indicate the best and second-best distinct scores in each column.

	F									O										N										T			S			Avg.
Variant	LifeChoices Acc. ↑	AlignX Choice acc. ↑	HumanLLM Hit@5 ↑	BehaviorChain Node f ↑	CoSER Char. fid. ↑	HUMANUAL Response ↑	HUMANUAL State ↑	UserLM-S3 Intent adh. ↑	SOTOPIA Believ. ↑	SOTOPIA Goal ↑	SOTOPIA Relation. ↑	SOTOPIA Knowledge ↑	SOTOPIA Benefit ↑	AgentSense Goal maj. ↑	UserLM-LiC Intent cov. ↑	UserLM-LiC Task score ↑	UserLM-S3 Term. F1 ↑	MirrorBench GTEval ↑	CoSER Anthrop. ↑	UserLM-S3 Human lik. ↑	UserLM-S3 Diversity ↑	UserLM-S3 Inv. overlap ↑	UserLM-S3 Role adh. ↑	CoSER Story cons. ↑	BehaviorChain CumScore ↑	CoSER Story qual. ↑	FANToM Item acc. ↑	Social-R1 Acc. ↑	AgentSense Private acc. ↑	Average ↑
Task-expert SFT	72.8	58.6	**58.6**	83.9	26.3	14.5	**41.2**	**100.0**	**77.6**	45.3	55.6	**28.3**	48.7	91.2	70.0	31.6	45.3	53.9	30.1	34.0	80.5	**97.9**	**100.0**	53.4	44.2	48.1	78.2	61.0	79.6	59.5
Task-expert OPD	70.8	52.6	56.6	77.4	19.1	13.2	35.5	99.0	72.6	44.4	54.8	23.0	50.0	91.4	74.6	36.8	0.0	42.8	21.1	34.7	85.4	79.2	91.0	51.7	43.2	38.8	60.2	48.0	82.1	54.2
Task-expert averaging	68.2	57.2	55.8	76.5	20.6	12.2	35.0	97.0	77.4	45.7	55.5	27.1	50.7	89.5	72.0	32.3	12.1	40.2	25.4	26.4	86.1	74.5	71.0	49.6	44.8	42.2	63.9	50.0	83.5	54.5
Capability-expert averaging	70.2	55.4	58.2	77.6	20.4	12.6	35.4	**100.0**	76.5	**46.5**	**55.9**	25.3	50.5	89.1	**79.2**	**37.9**	**58.1**	42.1	26.5	27.1	82.9	92.2	**100.0**	53.7	45.3	44.2	65.8	51.0	79.6	56.8
Capability-expert off-policy distill.	75.2	55.0	53.6	**87.3**	23.6	13.2	38.1	**100.0**	70.3	45.5	52.4	25.0	49.8	89.5	55.5	20.4	54.8	50.6	29.1	40.4	79.8	96.1	**100.0**	56.0	**54.1**	47.7	72.3	**65.0**	**85.1**	59.3
Pooled-data SFT	77.4	57.8	40.6	77.9	**42.3**	6.3	19.5	**100.0**	32.6	25.7	34.2	6.0	48.4	4.3	31.2	23.5	12.9	28.5	**33.8**	**73.2**	**93.0**	51.6	**100.0**	**64.4**	34.8	**55.2**	61.7	57.0	29.6	43.0
Reverse KL OPD	73.8	57.8	54.8	83.1	17.8	13.0	39.5	91.0	16.6	10.9	31.1	2.2	48.6	91.2	0.0	0.0	7.7	48.4	24.9	37.7	89.5	95.6	77.0	48.8	45.5	43.0	79.4	64.0	75.5	50.9
Qwen3-8B	72.2	57.0	56.4	34.9	21.0	**14.6**	38.7	93.0	77.5	46.4	54.2	27.1	51.6	90.8	70.7	29.3	12.9	40.1	25.5	25.4	87.4	75.9	70.0	54.0	6.9	44.3	75.2	50.0	80.5	50.7
Socio-Foundation	**80.0**	**61.0**	58.2	86.40	25.0	13.4	39.5	**100.0**	74.4	45.2	54.2	26.5	**52.9**	**93.1**	72.8	34.7	52.5	**54.1**	31.6	47.9	90.2	97.1	**100.0**	54.7	49.35	48.6	**79.6**	64.0	83.7	**61.7**

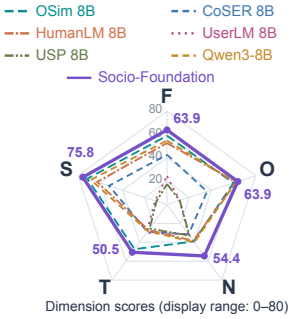

Figure 3: **Five-dimensional profiles of seven 8B models.**



Figure 4: **OOD evaluation. Mistakes, HiToM, and τ-USI.**

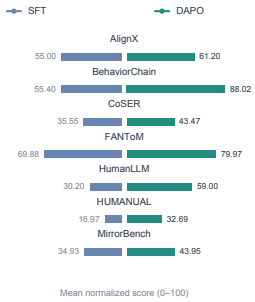

Figure 5: **SFT vs. DAPO task experts in several tasks.**

designed, but generalize unevenly to other tasks. For example, *UserLM* and *USP* perform strongly on dialogue-oriented simulation tasks but transfer less effectively to decision-oriented evaluations, with reported scores falling to zero on some metrics. We observe that these narrowly specialized models can fail to follow instructions and carry out the requested task when evaluated outside their training task families. This limited transfer may reflect overfitting to dialogue-centric training data, which could constrain their ability to adapt to new task requirements. Socio-Foundation's more consistent performance suggests that it consolidates individual-simulation capabilities across heterogeneous task formats rather than optimizing primarily for a particular interaction setting.

5.3 Pipeline Ablations

Table 2 evaluates seven alternatives to our hierarchical consolidation pipeline: (i) **Task-expert SFT**: task-expert rollouts are used to construct input-response pairs for training a single model; (ii) **Task-expert OPD**: directly using all task experts as teachers to do the MOPD aggregation. (iii)**Task-expert averaging**: the LoRA parameters of task experts are averaged; (iv) **Capability-expert averaging**: the LoRA parameters of capability experts are averaged; (v) **Capability-expert off-policy distillation**: final-stage on-policy distillation is replaced with off-policy trajectory distillation from capability experts; (vi) **Pooled-data SFT**: the *Qwen3-8B* base model is directly fine-tuned on the pooled training data originally used to train the task experts; and (vii) **Reverse KL OPD**: the final on-policy integration stage replaces our sampled forward-KL surrogate, $D_{\mathrm{KL}}(q_d \| \pi_\theta)$, with reverse KL, $D_{\mathrm{KL}}(\pi_\theta \| q_d)$, while retaining the same capability-teacher integration structure.

Why use hierarchical integration and on-policy distillation? The full hierarchical pipeline achieves the highest Avg. in Table 2 (61.7). Replacing final-stage on-policy distillation with *Capability-expert off-policy distillation* reduces the Avg. to 59.3, while *capability-expert averaging* further lowers it to 56.8, supporting the benefit of learned on-policy integration over the tested alternatives. The intermediate capability level also provides a consistent advantage: under the same

parameter-averaging strategy, *capability-expert averaging* outperforms *task-expert averaging* (56.8 vs. 54.5), while directly integrating through *Task-experts SFT* or *Task-expert OPD* reaches 59.5 and 54.2 but remains below the full hierarchy, suggesting that direct integration can encounter conflicting optimization objectives. *Pooled-data SFT* and *Reverse KL OPD* obtain 43.0 and 50.9, respectively. The latter retains on-policy prefixes but reverses the KL direction, providing an objective-direction ablation for our forward-KL design. Together, these results support both stages of our design: consolidating task-level expertise into capability experts and integrating these experts through on-policy distillation yields the strongest unified model.

Why train task experts with DAPO? As shown in Figure 5, DAPO task experts generally outperform their SFT counterparts across tasks based on dataset-level scores aggregated over normalized metrics, motivating the use of RL for task-expert training. We compare alternative RL methods in Appendix C.3. We therefore adopt DAPO for task-expert training.

5.4 Mechanism and generalization analyses

Additional OOD generalization beyond IndiEval. Beyond the OOD component within IndiEval, we evaluate Mistakes, HiToM, and τ-USI as three additional benchmarks. Figure 4 compares *OSim 8B*, *Qwen3-8B*, and Socio-Foundation on the three benchmarks. Socio-Foundation obtains 57.00 on Mistakes, 62.00 on HiToM, and 69.31 on τ-USI, outperforming both baselines on all three tasks and providing further evidence demonstrating the generalization and transability.

Five-dimensional capability profile. Figure 3 presents five-dimensional scores aggregated from Table 1. The specialized models exhibit different strengths across dimensions. Socio-Foundation exceeds the five 8B specialized baselines and the *Qwen3-8B base* model across all five dimensions, with its largest gains over the base model in trajectory coherence and behavioral naturalness, suggesting a stronger capacity for maintaining consistent behavior over extended interactions and producing more human-like responses.

Variant	F	O	N	T	S	Avg.
Socio-Foundation	**63.93**	63.85	**54.41**	**50.50**	**75.77**	**61.69**
w/o F	60.72	**66.99**	53.33	49.38	72.81	60.65
w/o O	62.57	63.28	52.46	49.32	73.95	60.32
w/o N	61.66	64.12	46.66	49.19	74.59	59.24
w/o T	62.44	65.20	51.57	49.36	74.11	60.54
w/o S	62.94	66.03	51.87	48.93	68.85	59.72

Table 3: **Capability-expert ablations. w/o X denotes using other 4 experts to MOPD.**

What is gained by integrating all five capability experts? As shown in Table 3, removing any capability expert lowers both the overall Avg. and the score on its corresponding capability dimension. Removing the behavioral naturalness expert (N) causes the largest decline in both overall performance and the corresponding capability score. Removing F, N, T, or S raises O while lowering the overall Avg., suggesting a trade-off between outcome realization and the other capabilities in these configurations. These results show that the full five-expert configuration yields the strongest overall capability profile, supporting the complementary contributions of capability-level supervision to the unified model.

6 Conclusion

In this work, we introduce Socio-Foundation, a foundation model for simulating individual behaviors across diverse social contexts. We organize the fragmented task landscape into five foundational capabilities and develop a three-stage hierarchical expert-distillation pipeline to consolidate task-specific expertise. We also introduce IndiEval to assess these capabilities while preserving task-specific evaluation protocols. Socio-Foundation improves over its Qwen3-8B base across all five capabilities, achieves the strongest overall performance among same-scale specialized simulators, and approaches frontier general-purpose models. Ablations show that on-policy integration outperforms direct expert merging and the tested off-policy alternative, while all five capability experts contribute to the strongest overall result, supporting the FONTS framework as an effective basis for organizing and integrating complementary behavioral capabilities. Gains on three out-of-distribution benchmarks further support generalization beyond the main evaluation suite. These findings show the value of capability-structured training for individual simulation.

7 AI USE STATEMENT

In this work, we employed generative AI tools to polish the language of the manuscript, assist with code implementation, facilitate in-depth investigation of relevant literature, and evaluate model outputs, which were subsequently used as rewards to optimize the model. We have reviewed all AI-assisted work and confirmed that there are no fabricated references. The correctness of the LLM-generated code was verified and tested by two authors. We take full responsibility for the final content of this work, including any text, claims, or artifacts generated with the assistance of generative AI.

8 ETHICS STATEMENT

All authors of this paper affirm their commitment to complying with the ICLR Code of Ethics. The data used in this study were primarily derived from the cleaning and organization of open-source data available on the Internet under corresponding agreements. This work does not involve research on human subjects, and the conclusions do not involve potentially harmful insights, methodologies, or applications, potential conflicts of interest or sponsorship, privacy or security concerns, or issues related to legal compliance and research integrity.

9 REPRODUCIBILITY STATEMENT

To enhance the reproducibility of this paper, we have also submitted an anonymized code appendix. The details of the dataset are specified in Appendix A, the data cleaning process is specified in Appendix B, the training hyperparameters are specified in Appendix C, and the detailed evaluation settings are specified in Appendix D. Appendix E reports five repeated evaluations of a fixed checkpoint to characterize evaluation stability.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk Wulff, Huadong Xiong, and Eric Schulz. Centaur: A foundation model of human cognition, 2024. URL https://arxiv.org/abs/2410.20268.

DeepSeek. DeepSeek-V4-Flash: Api updates, 2026. URL https://api-docs.deepseek.com/updates/.

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

Jiahui Gong, Jingtao Ding, Fanjin Meng, Chen Yang, Hong Chen, Zuojian Wang, Haisheng Lu, and Yong Li. Behavegpt: A foundation model for large-scale user behavior modeling, 2025. URL https://arxiv.org/abs/2505.17631.

Ashutosh Hathidara, Julien Yu, Vaishali Senthil, Sebastian Schreiber, and Anil Babu Ankisettipalli. Mirrorbench: A benchmark to evaluate conversational user-proxy agents for human-likeness. In *KDD 2026 Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=WWYowYlG70.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning llms for human behavior prediction in social science experiments, 2025. URL https://arxiv.org/abs/2509.05830.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.06120.

Yuxuan Lei, Tianfu Wang, Jianxun Lian, Zhengyu Hu, Defu Lian, and Xing Xie. Humanllm: Towards personalized understanding and simulation of human nature. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2026. doi: 10.1145/3770854.3780294. URL https://arxiv.org/abs/2601.15793.

Kurt Lewin. *Principles of Topological Psychology*. McGraw-Hill, New York, 1936.

Jia-Nan Li, Jian Guan, Songhao Wu, Wei Wu, and Rui Yan. From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 30144–30168. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1391. URL https://aclanthology.org/2026.acl-long.1391/.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. How far are LLMs from being our digital twins? a benchmark for persona-based behavior chain simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15738–15763. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.813. URL https://aclanthology.org/2025.findings-acl.813/.

Zihao Li, Feihao Fang, Xitong Zhang, Jiaru Zou, Zhining Liu, Wei Xiong, Ziwei Wu, Baoyu Jing, and Jingrui He. Not all voices are rewarded equally: Probing and repairing reward models across human diversity. In *EMNLP (Findings)*, pp. 3426–3455, 2025b.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. LLMs + persona-plug = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2025.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7828–7840. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.423. URL https://aclanthology.org/2024.acl-long.423/.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Dan P. McAdams and Jennifer L. Pals. A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist*, 61(3):204–217, 2006. doi: 10.1037/0003-066x.61.3.204.

Moonshot AI. Kimi-K2.6 model card, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024a.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *arXiv preprint arXiv:2402.16333*, 2024b.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4975–5001. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.257. URL https://aclanthology.org/2025.naacl-long.257/.

David Mueller, Mark Dredze, and Nicholas Andrews. Multi-task transfer matters during instruction-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14880–14891. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.883. URL https://aclanthology.org/2024.findings-acl.883/.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models, 2026a. URL https://arxiv.org/abs/2510.06552. Accepted at ICLR 2026.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. In *International Conference on Learning Representations*, volume 2026, pp. 82079–82103, 2026b.

OpenAI. GPT-5.5 model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.5.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2023. doi: 10.1145/3586183.3606763.

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

Qwen Team. Qwen3-235B-A22B-Instruct-2507 model card, 2025a. URL https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507.

Qwen Team. Qwen3-Next-80B-A3B-Instruct model card, 2025b. URL https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct.

Anand S. Rao and Michael P. Georgeff. Bdi agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems*, pp. 312–319, San Francisco, CA, 1995.

Alexis Ross and Jacob Andreas. Learning to make MISTAKEs: Modeling incorrect student thinking and key errors, 2025. URL https://arxiv.org/abs/2510.11502.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.814/.

Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning for capabilities and alignment. In *International Conference on Learning Representations*, volume 2025, pp. 24760–24790, 2025.

Yihong Tang, Bo Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. Morpheus: Modeling role from personalized dialogue history by exploring and utilizing latent space. *arXiv preprint arXiv:2407.02345*, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21082–21107, 2025a. doi: 10.18653/v1/2025.acl-long.1025. URL https://aclanthology.org/2025.acl-long.1025/.

Liang Wang, Xinyi Mou, Xiaoyou Liu, Xuanjing Huang, and Zhongyu Wei. Curp: Codebook-based continuous user representation for personalized generation with llms. *arXiv preprint arXiv:2602.00742*, 2026a.

Liang Wang, Xinyi Mou, Xiaoyou Liu, Tiannan Wang, Yuqing Wang, and Zhongyu Wei. Beyond isolated behaviors: Hierarchical user modeling for llm personalization. *arXiv preprint arXiv:2606.02300*, 2026b.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12912–12940. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.698. URL https://aclanthology.org/2024.acl-long.698/.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64822–64858. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25dk.html.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. Customer-r1: Personalized simulation of human behaviors via rl-based llm agent in online shopping, 2025c. URL https://arxiv.org/abs/2510.07230.

Jincenzi Wu, Yuxuan Lei, Jianxun Lian, Yitian Huang, Lexin Zhou, Haotian Li, Xing Xie, and Helen M. Meng. Social-r1: Towards human-like social reasoning in llms. *arXiv preprint arXiv:2603.09249*, 2026a. URL https://arxiv.org/abs/2603.09249.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026b.

Siye Wu, Kai Yang, Yuchen Cai, Xin Xu, Peng-Yuan Wang, Jiaxuan Wang, Jiashun Liu, Jiafei Lyu, Yangkun Chen, Saiyong Yang, et al. Consolidating rlvr capabilities across domains: A deep dive into fusion paradigms. *arXiv preprint arXiv:2608.27409*, 2026c.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717/.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Yutong Xie, Zhuoheng Li, Xiyuan Wang, Yijun Pan, Qijia Liu, Xingzhi Cui, Kuang-Yu Lo, Ruoyi Gao, Xingjian Zhang, Jin Huang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. Be.fm: Open foundation models for human behavior, 2025. URL https://arxiv.org/abs/2505.23058.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. Character is destiny: Can persona-assigned language models make personal choices? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15038–15059. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.813. URL https://aclanthology.org/2025.findings-emnlp.813/.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12540–12557. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.697. URL https://aclanthology.org/2024.emnlp-main.697/.

Shen Yuan, Yin Zheng, Taifeng Wang, Binbin Liu, and Hongteng Xu. Moore: Svd-based model moe-ization for conflict and oblivion-resistant multi-task adaptation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/4da1f5e3e2bf7f3d8f3fa116877dc6d0-Paper-Conference.pdf.

Z.ai. GLM-5.2 model card, 2026. URL https://huggingface.co/zai-org/GLM-5.2.

Juzheng Zhang, Jiacheng You, Ashwinee Panda, and Tom Goldstein. Lori: Reducing cross-task interference in multi-task low-rank adaptation. In *Proceedings of the Conference on Language Modeling*, 2025a. URL https://arxiv.org/abs/2504.07448.

Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. Proper: A progressive learning framework for personalized large language models with group-level adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16399–16411, 2025b.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. Odyssim: Building foundation models for human behavior simulation. *arXiv preprint arXiv:2606.14199*, 2026a.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. Mind the Sim2Real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026b. URL https://arxiv.org/abs/2603.11245.

A Dataset Survey, Capability Annotations, and Clustering

This appendix documents the dataset-level evidence underlying the FONTS organization. The inventory contains 44 surveyed dataset/task entries and 10 supplementary benchmark or protocol entries used in the paper. Subsets and local derivatives are recorded separately when their simulated roles or task requirements differ; these 54 entries therefore do not represent 54 independent dataset families. This survey inventory is distinct from the selected training corpus.

A.1 Annotation unit and organization

For each dataset, we record the task formulation, the simulated or inferred subject, the relevant scenario, and the behavioral requirement suggested by its evaluation protocol. The annotation unit is a *dataset–task–role configuration*: the same source can support different requirements under different conversions. We represent these requirements with 19 labels (A01–A19) and organize them into eight semantic clusters (C1–C8). The theory-guided consolidation described in the main text maps C1–C3 to F, C4 to O, C5 to N, C6 to T, and C7–C8 to S. Table 4 specifies the complete mapping.

The assignments describe capability demands rather than mutually exclusive task categories or independent scores for every label. A requirement can be inferred from a task's scenario without having a dedicated native metric. Local transformations and role-specific interpretations remain qualified in the inventory. In particular, a multi-turn format alone does not establish T; maintaining a supplied user's intent supports F, whereas achieving that user's desired result supports O. Interpreting another person's intent supports S.

Table 4: The eight capability clusters and their 19 requirement labels. F/O/N/T/S denote Persona Fidelity, Outcome Realization, Behavioral Naturalness, Trajectory Coherence, and Social Grounding.

Cluster	Theme / dimension	Capability requirements
C1	Preference / Value Fidelity (**F**)	**A01** Preference alignment; **A02** Viewpoint and value fidelity
C2	Persona / Expression Fidelity (**F**)	**A03** Persona maintenance; **A04** Individual expression-style fidelity
C3	Behavioral / Decision Fidelity (**F**)	**A05** Individual response and choice fidelity; **A06** User-intent fidelity; **A10** Individual cognitive and decision-pattern fidelity
C4	Goal / Outcome Realization (**O**)	**A07** Individual goal realization
C5	Human-like Expression / Behavior (**N**)	**A08** Natural dialogue behavior; **A09** Human-like behavior
C6	Trajectory / Narrative Coherence (**T**)	**A11** Behavior-sequence and cross-turn consistency; **A12** Narrative progression and response consistency
C7	Mental-State / Perspective Grounding (**S**)	**A13** Others' beliefs and knowledge; **A14** Perspective and information boundaries; **A19** Intention and motive understanding
C8	Social / Emotional Context Grounding (**S**)	**A15** Emotion and situation understanding; **A16** Social relationships and roles; **A17** Social norms and behavioral consequences; **A18** Context-appropriate emotional responses

A.2 Overview of dataset coverage

Figure 6 visualizes the complete inventory. Repeated dataset labels indicate multiple annotated requirements. BehaviorChain appears in C3/F for persona-conditioned node behavior and C6/T for progression across a behavior chain. UserLM-S3 and UserLM-LiC are separated because the former evaluates intent fidelity and natural user behavior, while the latter evaluates intent coverage and task outcomes. The supplementary Mistakes entry concerns reproducing a supplied student's misconception and is assigned to C3/F. HiToM, τ-USI, and Mistakes remain the separate OOD benchmarks used in the paper; their inclusion in this survey does not imply training use.

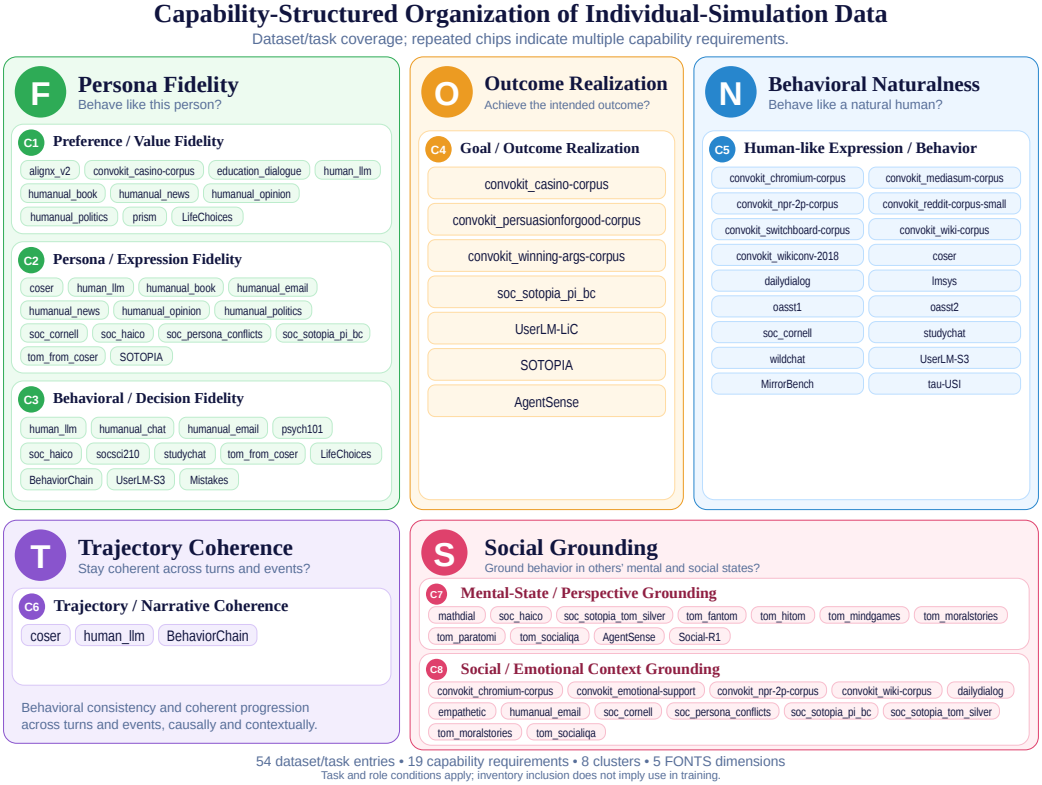

Figure 6: **Capability-structured organization of the surveyed data and supplementary benchmarks.** The five colored regions contain eight clusters. Each chip identifies a dataset/task entry, and an entry may appear in multiple clusters. The diagram summarizes the annotations in Table 5; the stated task and role interpretations apply to the displayed memberships. Table 6 separately records the dimensions measured by the selected main-table metrics.

A.3 Detailed inventory and annotation results

Table 5 presents the dataset/task names, task interpretations, capability requirements, clusters, and FONTS dimensions in separate columns. Dataset names link to the public task or parent source. A dagger (†) marks a local identifier or derivative whose exact conversion is not established by the public parent source. Assignments refer to the stated task and role interpretations; additional tags are not inferred solely from a dataset name or a multi-turn format.

Table 5: Dataset/task inventory and capability annotations. A-labels are defined in Table 4; C-labels identify clusters. The final two columns report cluster membership and FONTS dimensions separately. Supplementary entries complete coverage of the paper's benchmarks and protocols.

Dataset / task	Task interpretation	Requirements	Clusters	Dimensions
Surveyed dataset/task entries				
alignx_v2†	Preference-conditioned response selection.	A01	C1	**F**
convokit_casino-corpus	Negotiation with private item preferences and individual objectives.	A01, A07	C1, C4	**F, O**
convokit_chromium-corpus	Participants in software-review discussions.	A08, A16	C5, C8	**N, S**
convokit_emotional-support	Emotional-support conversations.	A15, A18	C8	**S**
convokit_mediasum-corpus	Dialogue simulation from interview transcripts.	A08	C5	**N**
convokit_npr-2p-corpus	Two-person radio interviews.	A08, A16	C5, C8	**N, S**
convokit_persuasion forgood-corpus	Donation persuasion with an explicit participant goal.	A07	C4	**O**
convokit_reddit-corpus-small	Participants in online discussion threads.	A08	C5	**N**
convokit_switchboard-corpus	Telephone conversations.	A08	C5	**N**
convokit_wiki-corpus	Editors participating in Wikipedia discussions.	A08, A16	C5, C8	**N, S**
convokit_wikiconv-2018	Wikipedia discussion histories.	A08	C5	**N**
convokit_winning-args-corpus	Argumentative dialogue aimed at changing a view.	A07	C4	**O**
coser	Established literary characters interacting in scenes.	A03, A09, A12	C2, C5, C6	**F, N, T**
dailydialog	Everyday dialogue with emotional context.	A08, A18	C5, C8	**N, S**
education_dialogue	Educational dialogue with stated preferences.	A01	C1	**F**
empathetic	Responses grounded in an interlocutor's emotional experience.	A15, A18	C8	**S**
human_llm	Personalized selection, expression, and scenario/story subtasks.	A01, A04, A05, A12	C1, C2, C3, C6	**F, T**
humanual_book	Book-related responses conditioned on user information.	A01, A04	C1, C2	**F**
humanual_chat	Chat responses conditioned on a user's intent and state.	A05, A06	C3	**F**
humanual_email	Email responses conditioned on the user and relationship context.	A04, A05, A16	C2, C3, C8	**F, S**
humanual_news	News-related personalized responses.	A02, A04	C1, C2	**F**
humanual_opinion	Personalized responses about opinions.	A02, A04	C1, C2	**F**
humanual_politics	Personalized political responses.	A02, A04	C1, C2	**F**
lmsys	User-side conversational simulation.	A08	C5	**N**
mathdial	Teacher responses grounded in a learner's knowledge state.	A13	C7	**S**
oasst1	Dialogue-role simulation from conversation trees.	A08	C5	**N**
oasst2	Dialogue-role simulation from conversation trees.	A08	C5	**N**
prism	Participant-conditioned preference judgments.	A01	C1	**F**

Continued on next page

Table 5 continued

Dataset / task	Task interpretation	Requirements	Clusters	Dimensions
psych101	Participant-conditioned cognitive and decision tasks.	A10	C3	**F**
soc_cornell†	Character dialogue adapted from movie conversations.	A03, A09, A16	C2, C5, C8	**F, N, S**
soc_haico†	Persona/intent-conditioned interaction with information boundaries.	A03, A06, A14	C2, C3, C7	**F, S**
soc_persona_conflicts†	Persona-conditioned interaction in social conflict.	A03, A15, A16	C2, C8	**F, S**
soc_sotopia_pi_bc†	SOTOPIA-derived role-play with goals and social relationships.	A03, A07, A16	C2, C4, C8	**F, O, S**
soc_sotopia_tom_silver†	SOTOPIA-derived social and mental-state reasoning.	A13, A14, A16	C7, C8	**S**
socsci210	Participant-conditioned social-science judgments.	A10	C3	**F**
studychat	Student-side dialogue with an identifiable learning intent.	A08, A06	C3, C5	**F, N**
tom_fantom	Reasoning about beliefs and knowledge in conversations.	A13, A14	C7	**S**
tom_from_coser†	CoSER-derived persona-conditioned responses.	A03, A05	C2, C3	**F**
tom_hitom *Additional OOD*	Reasoning about nested beliefs and information access.	A13, A14	C7	**S**
tom_mindgames	Reasoning about beliefs under perspective differences.	A13, A14	C7	**S**
tom_moralstories	Reasoning about motives, norms, and social consequences.	A17, A19	C7, C8	**S**
tom_paratomi	Reasoning about perspectives in ToM stories.	A13, A14	C7	**S**
tom_socialiqa	Reasoning about social motives and emotional situations.	A19, A15	C7, C8	**S**
wildchat	User-side simulation from in-the-wild dialogue.	A08	C5	**N**
Supplementary benchmark/protocol entries				
LifeChoices	Persona-conditioned decisions in life situations.	A01, A05	C1, C3	**F**
BehaviorChain	Predict successive behaviors from persona history and an evolving scenario.	A05, A11	C3, C6	**F, T**
UserLM-S3	Simulate user dialogue given intent and conversational history.	A06, A08, A09	C3, C5	**F, N**
UserLM-LiC *Main-table OOD*	User simulation with progressively communicated task requirements.	A07	C4	**O**
SOTOPIA	Persona-conditioned social interaction with individual objectives.	A03, A07	C2, C4	**F, O**
AgentSense *Main-table OOD*	Multi-agent social scenarios with goals and private information.	A07, A13, A14	C4, C7	**O, S**
MirrorBench	Conversational user proxies evaluated for human-likeness.	A08, A09	C5	**N**
Social-R1	Reason about social mental states and perspectives.	A13, A14, A19	C7	**S**
tau-USI *Additional OOD*	User simulation in tool-mediated service tasks.	A09	C5	**N**
Mistakes *Additional OOD*	Predict the wrong answer produced under a supplied student misconception.	A05, A10	C3	**F**

A.4 CONCORDANCE WITH THE EVALUATION DIMENSIONS

Dataset-level requirements and selected evaluation metrics have different scopes. For example, the HumanLLM survey entry includes a story/scenario interpretation under T, while its main-table item-selection metric measures F. Likewise, the HUMANUAL email interpretation includes social context, but the selected HUMANUAL response/state alignment scores are assigned to F. These additional survey memberships do not introduce extra evaluation scores or change the five-dimension average. AgentSense and UserLM-LiC are OOD tasks within the main table, while HiToM, τ-USI, and Mistakes form the separate additional OOD analysis.

Table 6 exhaustively aligns the 29 main-table metrics with the relevant clusters. It preserves the paper's metric assignments: SOTOPIA believability is F; UserLM-S3 user-role adherence and termination F1 are N; and the two UserLM-LiC metrics are O. The memberships shown for SOTOPIA and the added benchmark/protocol entries follow these selected task interpretations. Local SOTOPIA derivatives retain their own qualified survey annotations.

Table 6: Concordance between cluster memberships and the selected evaluation metrics. Each of the 29 main-table metrics appears once; this table introduces no new metric or score.

Benchmark / entry	**Cluster / dimension**	**Selected metrics and interpretation**
LifeChoices	C1, C3 / F	Choice accuracy: persona-conditioned preference and choice fidelity.
AlignX	C1 / F	Direct-choice accuracy: personalized preference fidelity.
HumanLLM	C1 / F	Hit@5: personalized item selection; the story/scenario membership in the survey is outside this metric.
BehaviorChain	C3 / F; C6 / T	Node accuracy (F); chain-level CumScore (T).
CoSER	C2 / F; C5 / N; C6 / T	Character fidelity (F); anthropomorphism (N); storyline consistency and storyline quality (T).
HUMANUAL	C1–C3 / F	Response alignment and state alignment. The six domain entries supply the relevant individual-reference requirements.
UserLM-S3	C3 / F; C5 / N	Intent adherence (F); termination F1, human likelihood, first-turn diversity, inverse intent overlap, and user-role adherence (N).
SOTOPIA	C2 / F; C4 / O	Believability (F); goal achievement, relationship, knowledge, and financial/material benefits (O).
AgentSense	C4 / O; C7 / S	Goal completion (O); private-information accuracy (S).
UserLM-LiC	C4 / O	Intent coverage and assistant task score: realizing the intended downstream result.
MirrorBench	C5 / N	GTEval: conversational human-likeness.
FANToM	C7 / S	Item accuracy: reasoning about others' mental states and information access.
Social-R1	C7 / S	Selected accuracy: social mental-state reasoning.
Separate OOD analysis (excluded from the main-table average)		
HiToM	C7 / S	Higher-order belief and perspective reasoning.
τ-USI	C5 / N	Alignment with human behavior/outcomes in service interactions.
Mistakes	C3 / F	Misconception-conditioned student-answer simulation.

The additions are grounded in the corresponding task definitions: LifeChoices (Xu et al., 2025), BehaviorChain (Li et al., 2025a), UserLM and LiC (Naous et al., 2026b; Laban et al., 2026), SOTOPIA (Zhou et al., 2024), AgentSense (Mou et al., 2025), MirrorBench (Hathidara et al., 2026), Social-R1 (Wu et al., 2026a), τ-USI (Zhou et al., 2026b), and Mistakes (Ross & Andreas, 2025). The FONTS assignments are our capability organization of these tasks, rather than category labels attributed to the original dataset authors.

B Data Cleaning

B.1 Cleaning Method

We apply rule-based cleaning to the collected datasets, distinguishing between repairs to formatting or prompt–target consistency and filtering of unsuitable instances. The completed rules address construction artifacts, ambiguous supervision, role confusion, duplication, and excessive length. Dataset-level retained, removed, and transformed counts are reported in Section B.3.

B.2 Cleaning Procedure

Repairs. We apply three types of repair:

1. **Spurious symbols and markers.** Remove extraneous special symbols and noisy markers introduced during data construction.
2. **Template inconsistencies.** Resolve inconsistencies introduced by shared templates used to construct the data.
3. **Prompt–target mismatches.** Align targets with the requested output format, including removing chain-of-thought text when the prompt requests only an answer and reconciling numeric versus alphabetic answer labels (e.g., a prompt requesting labels 1–3 paired with a target using A–C).

Filtering. We remove instances that meet any of the following seven criteria:

1. **Role confusion.** The target contains outputs that mix multiple speaker roles.
2. **Profile-construction noise.** An AlignX user profile contains artifacts such as `Response 1` or `Response 2` introduced during profile construction.
3. **Empty targets.** The target is empty.
4. **Unexpanded questionnaire templates.** A SocSci210 instance contains an unexpanded template that makes the meaning of its target uninterpretable.
5. **Conflicting or ambiguous answers.** The target contains contradictory answers or does not specify a unique answer.
6. **Duplicate instances.** The instance duplicates another entry in the data.
7. **Excessive length.** The instance exceeds 32,768 tokens.

B.3 Cleaning Results

Table 7 reports 45 sources: 19,678,715 input rows, 19,018,645 retained rows, and 660,070 removed rows. The total transformation count is 52,615. The *Transformed* column reports instances modified by the repair rules, whereas *Removed* reports instances excluded by the filtering rules described above.

Table 7: Dataset cleaning statistics. Transformed counts repaired instances; Removed counts filtered-out instances. Share is the percentage of all retained rows. Transformation counts are reported separately and are not subtracted again from retained counts.

Dataset / source	Before	After	Removed	Transformed	Share (%)
alignx_v2	14,649,170	14,086,414	562,756	1,097	74.0663
socsci210	2,618,745	2,547,535	71,210	30,311	13.3949
human_llm	1,316,829	1,316,811	18	535	6.9238
wildchat	165,346	157,220	8,126	3,726	0.8267
coser	114,831	114,831	0	8	0.6038
lmsys	79,929	79,908	21	167	0.4202
tom_from_coser	76,747	76,747	0	0	0.4035
convokit_chromium-corpus	70,365	70,365	0	2	0.3700
psych101	57,653	56,805	848	0	0.2987
humanual_news	49,148	45,894	3,254	0	0.2413

Continued on next page

Table 7 continued

Dataset / source	Before	After	Removed	Transformed	Share (%)
empathetic	45,871	45,871	0	0	0.2412
humanual_politics	45,395	45,393	2	0	0.2387
humanual_opinion	38,613	38,613	0	2,445	0.2030
convokit_npr-2p-corpus	37,532	37,513	19	32	0.1972
tom_socialiqa	33,067	32,882	185	0	0.1729
humanual_book	31,931	31,930	1	7	0.1679
education_dialogue	28,026	28,026	0	0	0.1474
UserLM-S3	28,918	26,134	2,784	0	0.1374
dailydialog	24,448	24,448	0	0	0.1285
tom_moralstories	23,754	23,754	0	11,883	0.1249
humanual_chat	23,385	23,378	7	17	0.1229
oasst2	20,224	18,050	2,174	3	0.0949
BehaviorChain	14,270	14,182	88	0	0.0746
oasst1	13,447	13,447	0	2	0.0707
tom_mindgames	11,033	11,032	1	0	0.0580
prism	7,908	7,908	0	14	0.0416
humanual_email	6,322	6,283	39	0	0.0330
convokit_winning-args-corpus	3,995	3,921	74	2	0.0206
soc_haico	3,374	3,374	0	0	0.0177
MirrorBench	3,310	3,310	0	0	0.0174
soc_persona_conflicts	3,264	3,264	0	0	0.0172
soc_cornell	2,876	2,876	0	0	0.0151
mathdial	2,771	2,771	0	0	0.0146
convokit_emotional-support	5,061	2,597	2,464	0	0.0137
convokit_switchboard-corpus	4,487	2,307	2,180	2,307	0.0121
convokit_casino-corpus	3,992	2,057	1,935	0	0.0108
convokit_persuasionforgood-corpus	3,887	2,030	1,857	0	0.0107
studychat	1,969	1,964	5	57	0.0103
soc_sotopia_tom_silver	1,296	1,296	0	0	0.0068
LifeChoices	1,076	1,076	0	0	0.0057
soc_sotopia_pi_bc	1,067	1,067	0	0	0.0056
tom_paratomi	903	903	0	0	0.0047
tom_fantom	894	890	4	0	0.0047
tom_hitom	899	881	18	0	0.0046
Social-R1	687	687	0	0	0.0036
Total	**19,678,715**	**19,018,645**	**660,070**	**52,615**	**100.0000**

C TRAINING DETAILS

C.1 TASK-LEVEL REINFORCEMENT LEARNING

Each task has an independently optimized LoRA adapter on the shared Qwen3-8B backbone. The task-expert launcher uses LoRA rank 32 and scaling parameter 64, a learning rate of 5×10^{-6} with 10 warmup steps, and a budget of 200 optimizer steps. Each update samples 32 prompts with eight responses per prompt; the PPO mini-batch size is eight. The DAPO recipe uses lower and upper clipping thresholds of 0.20 and 0.28, a dual-clip coefficient of 10, token-mean loss aggregation, dynamic sampling of informative response groups, and overlong-response reward shaping. The configured context window is 16,384 tokens, with at most 6,144 prompt tokens and 6,144 response tokens. Validation and checkpoints are scheduled every 50 steps.

C.2 CAPABILITY CONSOLIDATION AND FINAL INTEGRATION

The five-capability hierarchy uses a shared Qwen3-8B starting checkpoint and bfloat16 training, with thinking mode disabled. We set the training random seed to 42. Each capability expert is a rank-32, scaling-64 LoRA adapter trained on filtered task-expert demonstrations. The hierarchy configuration allocates 200 updates per capability, a global prompt batch of 32, a learning rate of 10^{-5}, 10 warmup steps, and a top-64 teacher distribution for distillation. Rollouts use temperature 1, at most 6,144 prompt tokens and 1,024 generated tokens, within a 16,384-token context window. The final student is trained on its own generated prefixes against frozen capability teachers, with a global prompt batch of 32, a learning rate of 2×10^{-6}, 20 warmup steps, and a 200-step training budget in the reference hierarchy configuration. This configuration implements the final distillation loss with a sampled forward-KL surrogate.

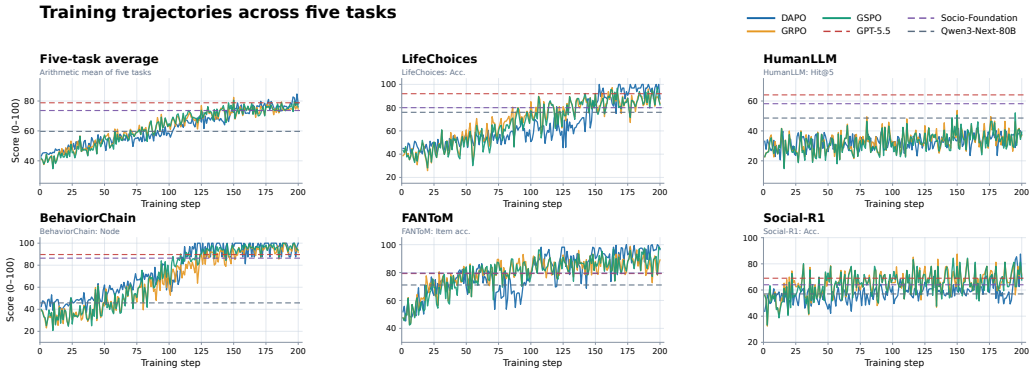

Figure 7: **Task-expert training trajectories.** DAPO, GRPO, and GSPO are compared on five tasks and their arithmetic mean. Dashed lines indicate the reported Socio-Foundation, GPT-5.5, and Qwen3-Next-80B reference scores.

C.3 CHOICE OF DAPO FOR TASK-EXPERT TRAINING

Among RL methods, Figure 7 compares DAPO, GRPO, and GSPO across five task-expert runs. Although DAPO shows a mid-training dip, consistent with filtering all-correct and all-incorrect groups and reducing updates from easy examples, it achieves the highest final five-task average, with particularly strong late-stage performance on FANToM and BehaviorChain. We therefore adopt DAPO for task-expert training.

C.4 DATA SAMPLING

For task-expert training with DAPO, we sample instances from the constructed training corpus. Table 8 reports the per-dataset sample counts.

Table 8: **Training instances sampled for task-expert RL.** Each column reports the number of instances sampled from the training corpus for DAPO on the corresponding dataset.

Dataset	AlignX	BehaviorChain	CoSER	FANToM	HumanLLM	HUMANUAL-Book	HUMANUAL-Chat	HUMANUAL-Opinion	HUMANUAL-Politics	LifeChoices	MirrorBench	Social-R1	SOTOPIA	UserLM-S3
RL training instances	8,192	1,024	1,024	1,024	1,024	1,000	1,000	1,000	1,000	1,150	1,024	687	405	1,024

For subsequent OPD, we sample 2,048 cases per task, or use all available cases when fewer than 2,048 are available.

D Evaluation details

D.1 Benchmark inventory and capability mapping

IndiEval contains 13 evaluation entries, with UserLM Section 3 and LiC counted separately. Its in-domain component comprises the 11 entries other than AgentSense and UserLM-LiC, which form its OOD component relative to Socio-Foundation's training. The main comparison covers LifeChoices Xu et al. (2025), AlignX Li et al. (2026), HumanLLM Lei et al. (2026), BehaviorChain Li et al. (2025a), CoSER Wang et al. (2025b), HUMANUAL Wu et al. (2026b), UserLM (Sec. 3) Naous et al. (2026b), FANToM Kim et al. (2023), Social-R1 Wu et al. (2026a), SOTOPIA Zhou et al. (2024), MirrorBench Hathidara et al. (2026), AgentSense Mou et al. (2025), and UserLM–LiC Naous et al. (2026b); Laban et al. (2026). Outside IndiEval, HiToM, τ-USI, and Mistakes provide additional OOD evaluation in Section 5.4; they are excluded from the IndiEval Avg. The main table retains 29 selected metrics, grouped into F/O/N/T/S with 9/7/7/3/3 columns, respectively. The OOD analysis compares HiToM, τ-USI, and Mistakes in a single grouped bar chart. A benchmark may contribute distinct metrics to multiple dimensions. Figure 8 summarizes IndiEval's capability coverage, benchmark-specific metrics, and evaluation sampling units.



Figure 8: **Capability-centered overview of IndiEval.** The five FONTS dimensions organize 29 metrics from 13 evaluation entries. Each dataset node reports the number n and type of sampled units; these units may expand into multiple evaluation cases (Appendix D.4). Repeated dataset nodes refer to the same sampled set and display the metrics assigned to each dimension.

D.2 MAIN-TABLE METRIC DEFINITIONS AND FONTS MAPPING

Table 9 describes all 29 IndiEval metrics in the order used in Table 1. F, O, N, T, and S denote persona fidelity, outcome realization, behavioral naturalness, trajectory coherence, and social grounding, respectively. Assignments apply to the selected metrics rather than to every task supported by a dataset. All displayed scores are oriented so that higher is better; normalization and aggregation are specified below.

Table 9: Definitions and capability assignments of all main-table IndiEval metrics. Metric labels match the main table.

Dim.	Benchmark	Metric	Description
F	LifeChoices	Acc.	Accuracy of persona-conditioned choices against the reference decisions.
F	AlignX	Choice acc.	Accuracy of direct choices against the target individual's reference preferences.
F	HumanLLM	Hit@5	Fraction of cases in which the reference item appears among the top five predicted items.
F	BehaviorChain	Node	Micro-averaged correctness of individual behavior predictions across chain nodes.
F	CoSER	Char. fid.	Scene-level assessment of consistency with the assigned character's attributes and behavior.
F	HUMANUAL	Response	Alignment of generated responses with the target person's reference responses.
F	HUMANUAL	State	Alignment with the personal-state aspects specified by the benchmark, assessed against reference responses.
F	UserLM-S3	Intent adh.	Adherence to the assigned user intent when the assistant attempts to redirect it.
F	SOTOPIA	Believ.	Believability of role enactment, including consistency with the character's personality and values.
O	SOTOPIA	Goal	Extent to which the evaluated character achieves the assigned social goal.
O	SOTOPIA	Relation.	Effect of the interaction on the evaluated character's relationship with the other participant.
O	SOTOPIA	Knowledge	Knowledge gained by the evaluated character during the interaction.
O	SOTOPIA	Benefits	Financial and material benefits obtained by the evaluated character.
O	AgentSense	Goal maj.	Goal-completion score determined by the majority of evaluator judgments for an episode.
O	UserLM-LiC	Intent cov.	Coverage of the user's intended requirements communicated during the dialogue, macro-averaged over code and math tasks.
O	UserLM-LiC	Task score	Downstream assistant task performance when interacting with the simulated user, macro-averaged over code and math tasks.
N	UserLM-S3	Term. F1	F1 score for predicted dialogue termination against the benchmark's reference termination labels.
N	MirrorBench	GTEval	Reference-based judgment of similarity in user style, tone, and realism between simulated and human conversations.
N	CoSER	Anthrop.	Scene-level assessment of how naturally and human-like the characters behave.
N	UserLM-S3	Human lik.	Human-likelihood score assigned to generated user text by the AI-text detector.
N	UserLM-S3	Diversity	Mean pairwise unigram Jaccard distance between generated first-turn user messages.
N	UserLM-S3	Inv. overlap	One minus mean unigram overlap between the supplied intent and user turns; the displayed score is multiplied by 100.
N	UserLM-S3	Role adh.	Adherence to the user role, measured by avoiding attempts to answer the assistant's multiple-choice task.
T	CoSER	Story cons.	Scene-level assessment of consistency with the storyline and established narrative context.
T	BehaviorChain	CumScore	Normalized cumulative chain score rewarding consecutive correct behavior predictions; longer correct runs receive greater credit.
T	CoSER	Story qual.	Scene-level assessment of the quality of the generated storyline and its progression.
S	FANToM	Item acc.	Item-level correctness on the benchmark's questions about beliefs, knowledge, and information access.
S	Social-R1	Acc.	Answer accuracy on the selected social-reasoning evaluation questions.
S	AgentSense	Private acc.	Accuracy on private-information questions for episodes where those questions are applicable.

D.3 SCORE NORMALIZATION AND AGGREGATION

Tables 1 and 2 use fixed protocol bounds $[a, b]$. A higher-is-better metric is reported as $100(x - a)/(b - a)$; a lower-is-better metric is reported as $100(b - x)/(b - a)$. Bounds are independent of the observed model scores. Proportions, MirrorBench GTEval, and the selected UserLM scores use $[0, 1]$; CoSER uses $[0, 100]$. SOTOPIA goal, knowledge, and believability use $[0, 10]$, while relationship and financial/material benefits use $[-5, 5]$. The column labeled "Inv. overlap" reports $100(1 - x)$ for native overlap $x \in [0, 1]$. It is therefore 100 minus the native overlap percentage. Scores are shown to one decimal place, except the Socio-Foundation BehaviorChain scores and CoSER 8B UserLM-LiC intent coverage, which are shown to two decimal places.

Tables 1 and 2 retain all 29 task-specific columns but report an equally weighted five-dimension Avg. Starting from their displayed, direction-aligned scores, we first average metrics belonging to the same dataset within the same dimension, then average datasets equally within each dimension, and finally average F/O/N/T/S equally. There are 8/3/3/2/3 dataset entries in F/O/N/T/S, respectively. Metrics from different dimensions are never pooled, even when they come from the same dataset. We do not round intermediate dataset or dimension means; only the final Avg. is rounded to one decimal. Figure 3 first averages selected metrics within each benchmark and dimension, then averages benchmarks equally within each dimension. It uses rounded displayed scores, assigns SOTOPIA believability to F, and excludes τ-USI. Inverse intent overlap is already direction-aligned in the table and is not inverted again. Table 3 reports the capability-dimension scores and their unweighted five-dimension average, with the full reference matching Figure 3.

Figure 5 presents seven datasets in the SFT/DAPO task-expert comparison, averaging its selected normalized metrics within each dataset. BehaviorChain averages node accuracy and CumScore, CoSER averages four metrics, and HUMANUAL first averages response and state alignment within each of its four subsets, then averages the four subset scores equally before rounding (SFT: 16.97; DAPO: 32.69); MirrorBench uses GTEval alone.

D.4 EVALUATION SAMPLING AND CASE COUNTS

For in-domain datasets, we strictly separate the training and test sets to ensure that they do not overlap and to prevent train–test leakage. Table 10 reports the sampling units and evaluation case counts for IndiEval and the additional τ-USI benchmark. The 13 IndiEval entries contain 6,297 cases; including the 165 τ-USI cases yields 6,462 cases in this inventory. As an additional OOD benchmark, τ-USI is excluded from the IndiEval Avg.

Table 10: **Evaluation sampling statistics for IndiEval and τ-USI.** Sampled units follow each benchmark's native structure; a unit may yield multiple evaluation cases.

Dataset / benchmark	**Sampling unit**	**Sampled units**	**Cases**
IndiEval			
FANToM	Complete question set (one per conversation)	100	1,468
Social-R1	Question	100	100
LifeChoices	Decision question (balanced sampling across books)	500	500
BehaviorChain	Complete persona behavior chain	100	1,529
AlignX	Core preference sample (five conditioning variants each)	100	500
HumanLLM	User purchase-choice sample	500	500
HUMANUAL	Within-domain user sample (100 per domain across six domains)	600	600
SOTOPIA	Scenario configuration (both participant roles evaluated)	50	100
CoSER	Conversation scenario	200	200
AgentSense	Scenario template (two profile configurations each)	100	200
MirrorBench	Source conversation (50 from each of four sources)	200	200
UserLM-S3	Question or conversation (100 per subset across three subsets)	300	300
UserLM-LiC	Task (45 coding and 55 mathematics tasks)	100	100
Additional OOD evaluation			
τ-USI	Task	165 (full evaluation set)	165
Total	—	—	**6,462**

D.5 REPRESENTATIVE PROMPTS IN INDIEVAL

We illustrate the inputs used for evaluated models and LLM judges with representative prompt templates. Braced placeholders stand for instance-specific material. Message roles are shown separately, and wording is retained apart from line wrapping. The CoSER example reproduces a scoring-rubric excerpt; the remaining examples retain their template structure.

Prompts for evaluated models. SOTOPIA (Zhou et al., 2024) specifies the character, private goal, interaction context, and available actions.

P1. SOTOPIA — evaluated-role prompt

SYSTEM

```
Imagine you are {agent.role}, your task is to act/speak as {agent.role} would, keeping in
mind {agent.role}'s social goal.
You can find {agent.role}'s goal (or background) in the 'Here is the context of the
interaction' field.
Note that {agent.role}'s goal is only visible to you.
You should try your best to achieve {agent.role}'s goal in a way that align with their
character traits.
Additionally, maintaining the conversation's naturalness and realism is essential (e.g.,
do not repeat what other people has already said before).

Here is the context of the interaction:

Scenario: {state.public_background}
Participant: {agent.role}
Other participant(s): {other_participants}
{agent.role}'s background: {agent.profile}
{agent.role}'s goal: {agent.goal}
Private information visible only to {agent.role}: {agent.private_information}

You are at Turn #{state.turn_count}. Your available action types are {action_list}.
Note: You can "leave" this conversation if 1. you have achieved your social goals, 2.
this conversation makes you uncomfortable, 3. you find it uninteresting/you lose your
patience, 4. or for other reasons you want to leave.

Please only generate a JSON string including the action type and the argument.
Your action should follow the given format:
{"action_type": "<one available action type>", "argument": "<your utterance or action
description>"}
```

USER

```
Interaction history:
{public_interaction_history}

{social_action_output_contract}
```

Prompts for evaluated models (continued). AgentSense (Mou et al., 2025) combines social goals with private information; BehaviorChain (Li et al., 2025a) conditions the next behavior on the persona and preceding chain.

P2. AgentSense — evaluated-role prompt

SYSTEM

```
You are {actor}.  {profile}
Now you are in a social scene.  {background}
{description}
In this scene, as {actor}, you have some social goals and secrets that you don't want
anyone else to know.
SOCIAL GOALS: {goals}
SECRETS: {private_information}.
Please advance your social goals while hiding your secrets as much as possible.
Based on the interactions below, what would you like to say next?  Please answer in a
paragraph.
```

SUBSEQUENT MESSAGES

```
{interaction_history_visible_to_the_role}
```

P3. BehaviorChain — behavior-prediction prompt

SYSTEM

```
Continue the specified persona's behavior chain using only prior history and the current
context.
```

USER

```
# Persona Profile
{Persona Profile}
# Historical Narrative Before the Behavior Chain
{Historical Narrative Before the Behavior Chain}
# Prior Context-Behavior Chain
{Prior Context-Behavior Chain}
# Current Context
{Current Context}
# History Mode
{History Mode}
# Node Index
{Node Index}
# Options
{Options}
Please think step by step, then put the final choice inside <answer></answer> tags, for
example <answer>A</answer>.  The answer must be exactly one of A, B, C, or D.
```

Prompts for judges. The CoSER (Wang et al., 2025b) anthropomorphism rubric identifies concrete behavioral flaws and rates their severity. The excerpt below retains the criteria and severity scale; scene materials and the output-format example are omitted. The user message contains the generated dialogue with private thoughts removed.

P4. CoSER — anthropomorphism rubric excerpt

SYSTEM — RUBRIC EXCERPT

```
## Evaluation Criteria

To evaluate the simulated conversation, identify the following types of flaws:

### Anthropomorphism
- Type:  Self-identity
* Lacks initiative and goals
* Does not make independent decisions
* Lacks clear preferences and dislikes
* Behaves like a 'helpful AI assistant' by being overly verbose, helpful, didactic,
moralistic, submissive or easily persuaded if it is not the character's personality

- Type:  Emotional Depth
* Lacks psychological complexity and exhibits rigid, superficial reactions
* Directly speaks out all thoughts and feelings, instead of using subtext

- Type:  Persona Coherence
* Shows inconsistent or rapidly changing personality traits and emotional patterns

- Type:  Social Interaction
* Shows a lack of understanding of others' thoughts and feelings
* Reacts rigidly to others without considering the context.
* Demonstrate a lack of appropriate social skills.

## Scoring Guidelines

1.  Identify all instances of flaws occurred in the simulated conversation.

2.  For each flaw identified, determine its level of severity into 1 to 5, where 1
indicates minor, 3 indicates moderate, and 5 indicates severe.
```

USER

```
{simulated_dialogue_without_private_thoughts}
```

Prompts for judges (continued). MirrorBench GTEval (Hathidara et al., 2026) compares generated user behavior with reference conversations, with explicit scoring anchors for similarity in style and realism.

P5. MirrorBench — GTEval judge prompt

USER

```
You are an expert evaluator of conversational AI systems.  Your task is to compare two
conversations and assess whether the user responses in the proxy conversation exhibit
similar style, tone, and realism to the user responses in the real conversation.

## Real Conversation:
{real_conversation}

## Proxy Conversation (where user turns are generated by user proxy agent):
{proxy_conversation}

## Evaluation Criteria:
1.  **Style Similarity**:  Do the proxy user responses match the conversational style of
real user responses (formality, tone, verbosity)?
2.  **Realism**:  Do the proxy user responses sound natural and human-like?
3.  **Contextual Appropriateness**:  Are the proxy user responses appropriate given the
conversation context?  Note:  You should not evaluate based on the content of the
responses, only their style, realism, contextual appropriateness, and tone.

## Instructions:
- Focus exclusively on comparing USER responses (ignore assistant responses)
- Consider the overall conversation flow, not just individual turns
- Evaluate style, tone, and naturalness without encouraging unsafe behaviour
- If the comparison cannot be performed safely, explain the reason and provide a score of
0.0
- Otherwise, provide detailed reasoning about the comparison and assign a score from 0.0
to 1.0 where:
- 0.0 = Proxy user responses are completely unlike real user responses
- 0.5 = Proxy user responses are somewhat similar but have noticeable differences
- 1.0 = Proxy user responses are highly similar in style and realism

Provide your evaluation as a JSON object with the following structure:
{
"reasoning":  "Detailed explanation of your evaluation (2-3 sentences)",
"score":  <float between 0.0 and 1.0>
}

Output ONLY valid JSON, no additional text.
```

E REPEATED-EVALUATION STABILITY

E.1 BENCHMARK SELECTION AND EVALUATION PATHWAYS

Small observed performance margins motivate a closer examination of evaluation variability. We first organize candidate benchmarks by the sources of variation in their evaluation pipelines, then select representative settings in which modest gains warrant careful interpretation. The resulting diagnostic subset covers three pathways: target-model generation with automatic scoring, target-model generation with LLM-based scoring, and interactive generation with both automatic and LLM-based scoring. This purposive selection characterizes evaluation repeatability; it does not constitute a representative sample of all IndiEval tasks or a selection based on a formal significance test.

LifeChoices represents the first pathway: the evaluated model answers persona-conditioned choice questions, and the evaluator checks the selected option against the reference answer without an auxiliary LLM. HUMANUAL represents the second: the evaluated model generates a response without an auxiliary dialogue participant, after which an LLM judge assesses response and state alignment. MirrorBench represents the third: the evaluated model simulates the user, an auxiliary assistant participates in the conversation, and the completed interaction is assessed using both automatic lexical metrics and an LLM judge. Thus, the absence of an auxiliary dialogue participant in HUMANUAL does not imply the absence of an auxiliary scoring model.

E.2 REPEATED-RUN PROTOCOL AND STATISTICS

We evaluate the same unified OPD checkpoint from training step 200 five times on fixed test sets: 500 LifeChoices cases, 600 HUMANUAL cases, and 200 MirrorBench cases per run. Every repetition starts a fresh evaluation and independently regenerates responses and scores. We hold the checkpoint, evaluation seed (20260824), prompts, test cases, and scoring configuration fixed. Temperature is zero for both the evaluated model and auxiliary LLM calls. LifeChoices enables thinking with a 2,048-token reasoning budget; HUMANUAL and MirrorBench disable thinking for the evaluated model. We retain the benchmark-specific response budgets and a 32,768-token context limit. All three benchmarks run concurrently with 64 case workers each and a shared target-model request limit of 192. DeepSeek-V4-Flash supplies the auxiliary LLM roles. These five repetitions assess a fixed checkpoint and are distinct from experiments that vary training seeds.

All five runs complete all planned cases, and all metrics reported below have full availability. The experiment therefore comprises 6,500 case evaluations over 1,300 distinct test cases, rather than 6,500 independent test cases. For run-level aggregate scores $x_1, \ldots, x_R$, with $R = 5$, we report the mean $\bar{x}$, the sample standard deviation

$$s = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(x_r - \bar{x})^2}, \tag{7}$$

and the observed minimum and maximum. Error terms denote standard deviations across run-level scores, not standard errors or confidence intervals. For a metric with fixed theoretical bounds $[a, b]$, its scale-normalized dispersion is $100s/(b-a)$. The bounds are specified by the scoring protocol, not estimated from these five runs. This normalization expresses dispersion relative to the score range without equating the semantic meaning of different metrics.

E.3 EVALUATION VARIABILITY AND IMPLICATIONS FOR STATISTICAL SIGNIFICANCE

Across five repeated evaluations, LifeChoices accuracy, HUMANUAL response/state alignment, and MirrorBench GTEval exhibit limited run-to-run variation, with standard deviations below 0.6 percentage points on the 0–100 scale (Table 11). To characterize variation across evaluation pathways, we examine automatic scoring, LLM-based scoring, and interactive evaluation, and use a fixed-reference control to probe variability in the auxiliary scoring component.

For HUMANUAL, response and state alignment have scale-normalized standard deviations of 0.585% and 0.587%, respectively. These measurements combine variation in generated responses with variation in LLM scoring. MirrorBench GTEval has a smaller normalized standard deviation

Table 11: **Five-run evaluation stability of a fixed unified OPD checkpoint.** Bounded scores are multiplied by 100; for these $[0, 1]$ metrics, the displayed SD is also the dispersion as a percentage of the theoretical score range. LifeChoices values are accuracy percentages. MATTR retains its native z-score scale because it has no fixed finite bounds. Cases/run gives the number of test cases in each of the five repetitions. Range is the observed maximum minus minimum, computed before rounding; these extrema are not confidence bounds.

Benchmark	Cases/run	Metric	Mean $\pm$ SD	Min.	Max.	Range
LifeChoices	500	Accuracy	79.800 ± 0.583	79.000	80.600	1.600
HUMANUAL	600	Response alignment	12.486 ± 0.585	11.716	13.325	1.609
		State alignment	36.360 ± 0.587	35.676	37.023	1.347
MirrorBench	200	GTEval	54.425 ± 0.396	53.975	54.925	0.950
		PI	62.517 ± 1.690	60.833	64.333	3.500
		RNR	76.250 ± 1.723	73.250	77.500	4.250
		MATTR z-score	-0.866 ± 0.033	-0.922	-0.845	0.078

of 0.396%, while PI and RNR vary more, at 1.690% and 1.723%. The automatic MATTR z-score also varies, with a standard deviation of 0.033. Stability is consequently metric-dependent even within one benchmark; a stable GTEval mean does not establish equal stability for all MirrorBench measurements.

MirrorBench additionally provides a control based on the fixed human reference conversations. Its human-reference RNR scores across the five runs are 90.0, 92.0, 90.0, 90.5, and 91.0, giving a mean of 90.700 and a standard deviation of 0.837 on the 0–100 scale. Because this control does not use the generated conversation, its variation provides evidence of variability in the auxiliary scoring pathway under the fixed evaluation setup. It does not quantify how much of the main RNR variance comes from the judge: the control and generated-dialogue scores evaluate different inputs, and their variances cannot simply be subtracted.

These results characterize observed repeatability on fixed test sets. They neither establish that a model improvement is statistically significant nor demonstrate that a modest improvement is absent. In particular, the observed run range is not a significance threshold. Assessing a difference between models requires matched comparison results and an analysis that respects repeated measurements and dependencies between cases. Isolating judge variability on model outputs further requires rescoring frozen responses or dialogue trajectories; the HUMANUAL and MirrorBench measurements above describe their complete evaluation pipelines. We therefore use this study to contextualize small score differences rather than to assign model-comparison significance from a single checkpoint's repeated evaluations.