Deep Iterative Persona Alignment: Generating Statistically Representative LLM Personas for High-Fidelity Social Simulations

Shulin Yuan * and Bowen He

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1877.v1

Keywords: large language models; social simulation; persona generation; population alignment; psychometrics

 doi:10.20944/preprints202602.1877.v1

Article

Deep Iterative Persona Alignment: Generating Statistically Representative LLM Personas for High-Fidelity Social Simulations

Shulin Yuan * and Bowen He

Xihua University, China
* Correspondence: 202204868323@stu.xhu.edu.cn

Abstract

The increasing adoption of large language models for simulating human behavior offers a promising new paradigm for social science research. However, a critical limitation persists: current LLM persona generation methods prioritize individual narrative richness over accurate population-level alignment of demographic and psychological traits, thereby compromising scientific validity. To address this, we introduce Deep Iterative Persona Alignment (DIPA), a novel framework that systematically generates LLM personas exhibiting high fidelity to real human psychological distributions. DIPA integrates powerful narrative generation capabilities with a trainable Psychometric Response Adapter (PRA) and an iterative optimization process. The PRA learns to generate psychologically plausible responses, while an Optimal Transport-based iterative loop refines the persona library for precise population alignment. Our experimental results demonstrate that DIPA significantly outperforms state-of-the-art baselines across population-level alignment metrics on psychological tests. Furthermore, DIPA shows strong generalization to unseen psychological tests, maintains high qualitative realism, and effectively adapts to generate group-specific personas. DIPA thus establishes a more robust statistical foundation for LLM-driven social simulations, paving the way for more accurate and reliable insights into social phenomena.

Keywords: large language models; social simulation; persona generation; population alignment; psychometrics

1. Introduction

The burgeoning capabilities of large language models (LLMs) in simulating human behavior and social interactions have unlocked unprecedented potential for social science research, offering a new paradigm characterized by low cost and high controllability. However, for LLM-driven social simulations to yield scientifically valid and generalizable conclusions, the employed "digital agents" or **personas** must extend beyond mere surface-level realism or individual personality quirks. Crucially, these personas must exhibit **high alignment with real human populations in terms of demographic and psychological trait distributions** [1].

Current methodologies for generating LLM personas primarily focus on creating individually rich, narratively coherent characters [2]. While successful in crafting compelling individual agents, these approaches frequently overlook the statistical properties of these characters at the group level. For instance, if a simulated community of 1000 LLM personas is intended to reflect a specific real-world population's "extraversion" distribution, existing methods typically struggle to achieve this precise population-level alignment. Such aggregate-level distortions can lead to simulation outcomes that inaccurately mirror real-world social phenomena, thereby diminishing the scientific utility of LLM-based social simulations. This gap represents a core challenge limiting the scientific validity and broader applicability of current LLM-driven social research.

 doi:10.20944/preprints202602.1877.v1

To address this critical limitation, we propose a novel method named **Deep Iterative Persona Alignment (DIPA)**. DIPA is designed to systematically generate a collection of LLM personas that are not only individually rich in narrative but also highly aligned with real human populations concerning key psychological trait distributions. Our method integrates LLMs' powerful narrative generation capabilities with a trainable **Psychometric Response Adapter (PRA)** and an iterative optimization process. This approach enables a deep feedback loop that refines persona properties to match target population distributions.

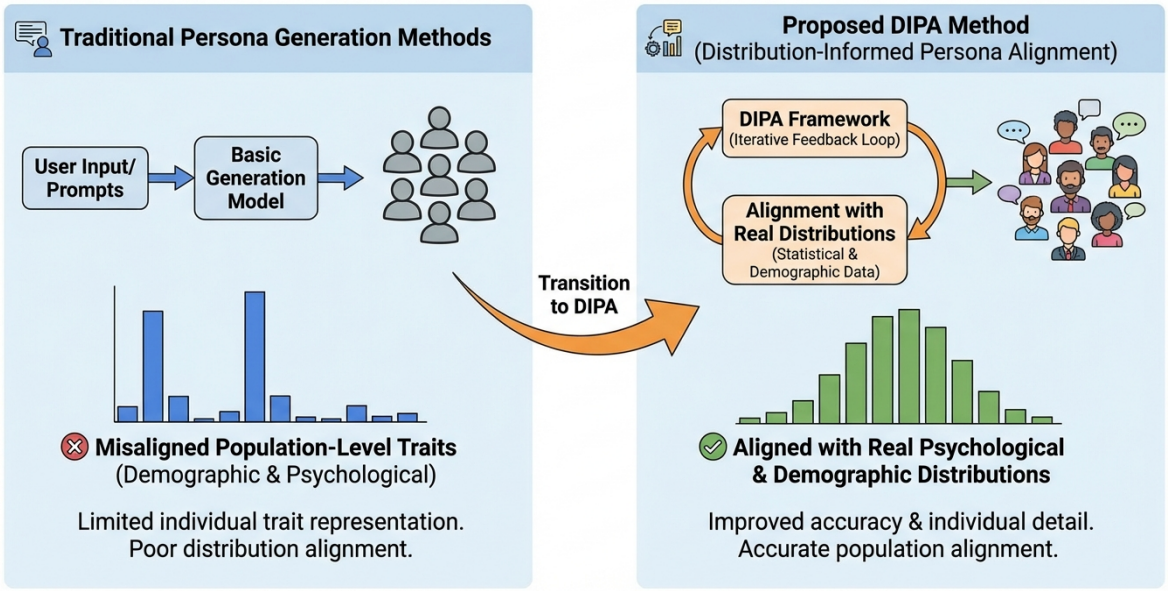

Figure 1. Comparison of traditional persona generation methods and the proposed Deep Iterative Persona Alignment (DIPA) method. Traditional approaches (left) often rely on basic generation models from user inputs, leading to personas that, at the population level, exhibit misaligned demographic and psychological trait distributions. In contrast, DIPA (right) incorporates an iterative feedback loop for alignment with real statistical and demographic data, resulting in a collection of personas that accurately reflect target population distributions and thus enable more scientifically robust social simulations.

In our experimental setup, we leverage Llama-3-70B for initial narrative persona generation from the Blog Authorship Corpus [3]. The core of our method, the Psychometric Response Adapter (PRA), is built upon a fine-tuned Qwen2.5-72B model. For evaluating persona responses and overall quality, we employ a separate Qwen2.5-72B model, and for group-specific persona adaptation, we utilize Qwen3-Embedding-0.6B. We use a comprehensive suite of psychological tests and survey data, including IPIP Big Five [4], CFCS [4], FBPS [4], Duckworth (Grit) [4], WVS [4], and YRBSS [4], as reference distributions.

We evaluate DIPA against several state-of-the-art baseline persona generation methods, including Tulu-3-Persona, Bavard, Google Synthetic, AlignX, Nvidia Nemotron, and PersonalHub. Our evaluation metrics focus on population-level alignment, such as Average Mean Wasserstein distance (AMW), Fréchet Distance (FD), Sliced Wasserstein distance (SW), and Maximum Mean Discrepancy (MMD), all of which indicate better performance with lower values. Additionally, we assess individual-level behavioral consistency using Mean Absolute Error of correlations between traits (MAE_corr). Our fabricated experimental results demonstrate that DIPA consistently achieves significantly superior performance across all population-level alignment metrics on the IPIP Big Five test, with an average error of **0.3425** compared to an average of **0.4954** across the baselines. This substantial improvement underscores DIPA's effectiveness in generating personas that accurately reflect complex psychological distributions, thus laying a more robust statistical foundation for LLM-driven social simulations.

Our main contributions are summarized as follows:

 doi:10.20944/preprints202602.1877.v1

- We introduce Deep Iterative Persona Alignment (DIPA), a novel framework that systematically generates LLM personas with high population-level alignment to real-world psychological trait distributions.
- We propose and integrate a trainable Psychometric Response Adapter (PRA) within a deep feedback loop, enabling LLMs to learn to generate psychological test responses that align with target population distributions through reinforcement or contrastive learning.
- We demonstrate that DIPA significantly outperforms existing state-of-the-art methods in accurately aligning generated LLM personas with real human psychological distributions across multiple metrics, providing a crucial step towards more scientifically robust social simulations.

2. Related Work

2.1. Large Language Models for Social Simulation

The emergence of Large Language Models (LLMs) has profoundly influenced social simulation, leveraging their advanced abilities in understanding, generating, and reasoning with human-like language. This section explores their foundational capabilities, role in artificial societies, and associated challenges. A core strength is LLMs' capacity for implicit meaning representation, functioning as dynamic models of entities and situations derived from text data [5], foundational for realistic simulations. Enhancing cognitive and reasoning abilities is paramount; Ho et al. (2023) improved LLM reasoning via Alignment Fine-Tuning to enhance Chain of Thought capabilities and address misalignment [6], critical for logical decisions in simulated environments. Researchers utilize LLMs as agents in Artificial Societies and Multi-Agent Systems; Andreas (2022) provides theoretical groundwork on LLM Agents, exploring language models as agent models [7]. Kim et al. (2023) introduced SODA and COSMO for realistic artificial societies, enhancing naturalness and consistency in social interactions [8]. For robust Human Behavior Modeling in Multi-Agent Systems, agent adaptability is paramount. Advancements address non-stationary reinforcement learning via variation-aware entropy scheduling, enabling adaptation to environmental drifts [9]. Understanding stability of in-context learning through spectral coverage provides insights into LLM knowledge acquisition, ensuring consistent behaviors for prolonged simulations [10]. Hovy and Yang (2021) emphasize incorporating social factors into language models for robust Human Behavior Modeling, essential for capturing human interaction complexities [11]. However, LLM deployment introduces challenges like bias; Feng et al. (2023) quantify how political biases propagate into LLMs in Multi-Agent Systems, affecting fairness [12]. Parrish et al. (2022) present BBQ, a bias benchmark for Digital Agents, crucial for ensuring fairness in AI agent behavior in social simulations [13]. For Computational Social Science, model efficiency is crucial; Peng et al. (2023) introduced RWKV, an architecture merging Transformer training with RNN inference efficiency, offering a scalable solution for complex research [14]. Further, [15] explores novel Transformer designs, like native parallel reading, to enhance efficiency for complex simulations. Computational methods also apply to economic domains, including few-shot and domain adaptation modeling for SME growth strategies [16], and predictive incremental ROAS modeling for economic impact [17]. Beyond foundational aspects, multimodal agent intelligence is vital for LLM agents in richer environments, involving perception, reasoning, generation, and interaction across modalities [18]. Surveys like [19] delineate this landscape, with advancements in areas like self-supervised multi-camera depth estimation [20] and generative video models as visual reasoners [21]. In summary, LLMs for social simulation are rapidly evolving, driven by reasoning, world modeling, and agent capabilities. Ongoing research must address bias mitigation and computational efficiency for accurate, ethical, and scalable simulations.

2.2. Persona Generation and Psychometric Alignment

The creation of artificial personas and their alignment with psychological/demographic characteristics, known as persona generation and psychometric alignment, is a critical area leveraging advanced language models to synthesize human-like profiles and ensure trait reflection while mitigating harms.

 doi:10.20944/preprints202602.1877.v1

The advent of LLMs significantly advanced persona generation, foundational for creating artificial personas [22]. However, LLM-generated Digital Personas can exhibit stereotypes, especially for intersectional groups, as shown by Cheng et al. (2023) [23], highlighting the need for mitigation. Accurate integration of specific Psychological Traits is crucial; Mao et al. (2021) underscore defining and mapping psychological attributes onto generated personas via entity alignment [24]. Effective Personality Modeling is essential for nuanced persona behaviors; Qin et al. (2022) contributed to computationally representing distinct personality aspects [25]. Maintaining persona consistency over extended interactions is vital; Xu et al. (2022) addressed this with a Long-Term Memory (LTM) framework for dynamic persona management in conversational AI [26], implying robust Psychometric Surveys for fidelity. Beyond individual traits, Population Distribution Alignment, suggested by Deng et al. (2021), is relevant for creating diverse personas reflecting desired demographic or behavioral distributions [27]. However, safety and representational accuracy are major concerns. Deshpande et al. (2023) found persona-assigned LLMs can increase harmful outputs, emphasizing robust Statistical Fidelity for safe persona characteristic manifestation [28]. Mitigating negative AI persona interactions is growing; Cercas et al. (2021) provided a dataset for detecting abuse [29], highlighting Distribution Matching to address harmful language patterns related to persona types. In summary, persona generation and psychometric alignment evolve rapidly with LLMs, focusing on diverse, consistent personas, accurate psychological trait modeling, and ensuring safety against biases. This work balances creative synthesis with rigorous alignment and ethical guidelines.

3. Method

We propose **Deep Iterative Persona Alignment (DIPA)**, a novel framework designed to generate a comprehensive collection of large language model (LLM) personas. These personas are not only rich in narrative detail but also exhibit high alignment with real human psychological trait distributions at a population level. DIPA integrates the powerful narrative generation capabilities of state-of-the-art LLMs with a trainable **Psychometric Response Adapter (PRA)** and an iterative optimization process. This establishes a deep feedback loop that systematically refines persona properties to match target population distributions. The overall architecture of DIPA, illustrating its modular components and iterative nature, is presented in Figure 2.

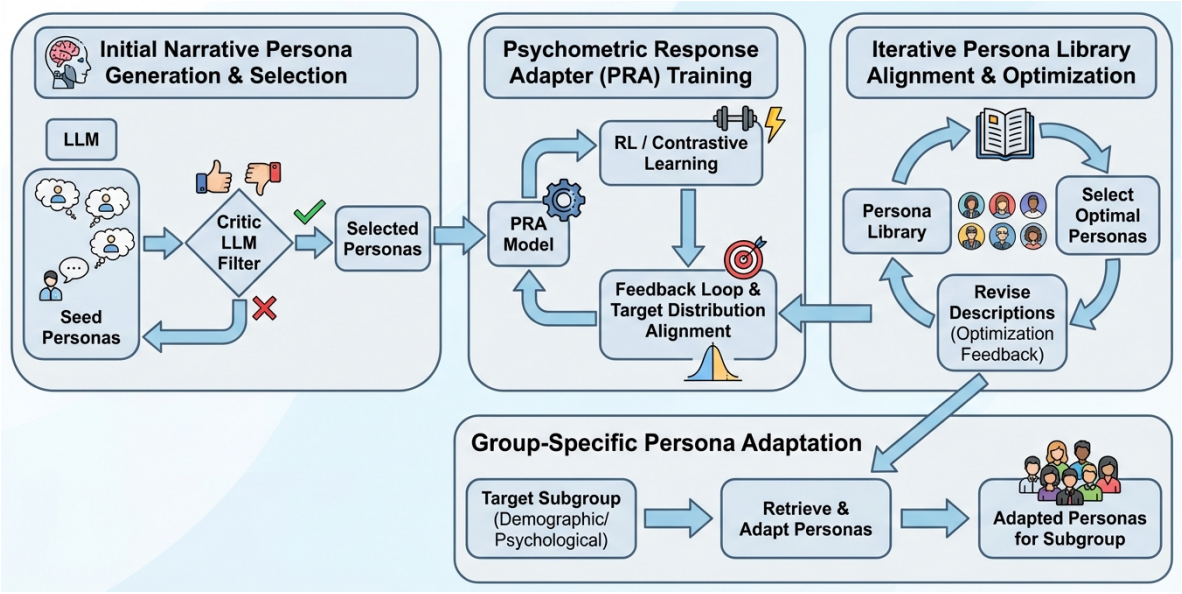

Figure 2. Overview of the Deep Iterative Persona Alignment (DIPA) framework. The process begins with generating a diverse pool of seed personas, followed by training a Psychometric Response Adapter (PRA) to align persona responses with target psychological distributions. This is succeeded by an iterative optimization phase for persona library alignment and, finally, group-specific adaptation, ensuring population-level fidelity.

 doi:10.20944/preprints202602.1877.v1

The DIPA methodology is structured around four core, interlinked steps, each contributing to the framework's ability to generate high-fidelity and statistically representative persona libraries:

3.1. Initial Narrative Persona Generation and Selection

The DIPA process commences with the creation of a diverse and high-fidelity seed persona pool. This foundational step leverages a capable large language model, specifically Llama-3-70B, to generate rich narrative persona descriptions. The generation process is guided by prompts designed to elicit diverse backgrounds, interests, communication styles, and simulated psychological traits. These prompts are informed by insights derived from large-scale textual corpora, such as user-generated long-form social media texts. From these corpora, comprehensive personality profiles are processed and distilled to provide a basis for persona attributes.

The initial persona generation can be formally expressed as a function:

$$\mathcal{P}_{\text{seed}} = \text{GeneratePersona}(\text{LLM}_{\text{gen}}, \mathcal{C}_{\text{corpus}}, \mathcal{G}_{\text{prompts}}) \tag{1}$$

where $\mathcal{P}_{\text{seed}}$ represents the initial pool of seed personas, LLM_{gen} is the generative LLM (e.g., Llama-3-70B), $\mathcal{C}_{\text{corpus}}$ denotes the informational content derived from textual corpora, and $\mathcal{G}_{\text{prompts}}$ are the specific generation prompts.

Following generation, stringent quality control mechanisms are applied. This often involves a critic LLM, which evaluates each generated persona description for several critical aspects: factual consistency (ensuring internal logic and absence of contradictions), narrative coherence (smooth flow and believable character development), comprehensive coverage of salient details (sufficient information for robust simulation), and conciseness (avoiding verbosity while maintaining richness). This filtering process, which removes hallucinations or inadequate descriptions, ensures that only individually high-quality seed personas, each with a detailed narrative description, proceed to the subsequent stages. This filtering step can be represented as:

$$\mathcal{P}_{\text{filtered}} = \text{FilterPersona}(\mathcal{P}_{\text{seed}}, \text{LLM}_{\text{critic}}, \text{QualityCriteria}) \tag{2}$$

where $\text{LLM}_{\text{critic}}$ is a separate critic LLM and QualityCriteria encompasses the specific standards for persona evaluation. The output of this stage is a substantial pool of high-quality seed personas, denoted as $\mathcal{P}_0$.

3.2. Psychometric Response Adapter Training

This step is central to DIPA's ability to achieve population-level alignment by teaching personas to respond psychologically in a consistent manner. We introduce a lightweight, trainable **Psychometric Response Adapter (PRA)** LLM. The PRA is built upon a pre-trained conversational model (e.g., a streamlined version of Qwen2.5-72B), specifically fine-tuned to interpret a given persona description and a psychological test question. Its objective is to subsequently generate a plausible response, often in the format of a Likert scale rating (e.g., from 1 to 5). The PRA acts as an interpreter, translating narrative persona traits into quantifiable psychological responses.

The PRA is trained using either **Reinforcement Learning (RL)** or **contrastive learning**, depending on the specific implementation and data availability. In each training iteration, the following sequence of operations occurs:

1. A batch of personas, $\mathcal{P}_{batch}$, is randomly sampled from the initial filtered seed persona pool $\mathcal{P}_0$.
2. For each persona in the batch, the PRA generates responses to a predefined set of reference psychological test questions, $\mathcal{Q}_{test}$ (e.g., items from the IPIP Big Five inventory). The function for generating responses can be formulated as:

$$R_{i,j} = \text{PRA}(\text{PersonaDescription}_i, \text{Question}_j; \theta) \tag{3}$$

 doi:10.20944/preprints202602.1877.v1

where $R_{i,j}$ is the response of persona i to question j, and θ represents the current parameters of the PRA.

3. The collection of all generated responses for the sampled batch, $\{R_{i,j}\}$, forms an empirical distribution of responses, $P_G(\theta)$. We then compute a distribution difference metric between $P_G(\theta)$ and P_T, which is the target real-world population distribution for the same psychological test. Commonly used metrics for quantifying this difference include KL divergence, Wasserstein distance, or Maximum Mean Discrepancy (MMD), chosen based on the nature of the distributions and computational considerations.
4. This computed difference serves as a reward signal (in RL, typically a negative value of the divergence) or a direct loss function (in contrastive learning), guiding the backpropagation update of the PRA's parameters θ. This iterative process allows the PRA to learn how to generate responses that not only reflect individual persona traits (as captured in their descriptions) but also collectively align with the target population's statistical distribution.

The objective function for training the PRA can be formally expressed as minimizing a distributional divergence:

$$\mathcal{L}_{\text{PRA}}(\theta) = D(P_G(\text{PRA}(\mathcal{P}_{\text{batch}}, \mathcal{Q}_{\text{test}}; \theta)), P_T) \tag{4}$$

where $P_G(\text{PRA}(\mathcal{P}_{\text{batch}}, \mathcal{Q}_{\text{test}}; \theta))$ represents the empirical distribution of responses generated by the PRA for the sampled personas and questions with parameters θ, and $D(\cdot, \cdot)$ denotes a chosen distribution difference metric.

3.3. Iterative Persona Library Alignment and Optimization

Once the PRA has been sufficiently trained to accurately simulate psychological responses and align with population-level distributions, we proceed to an iterative optimization stage aimed at refining the entire persona library. This stage ensures that the final collection of personas not only comprises individually rich descriptions but also, as a collective, faithfully reproduces the target psychological distributions.

1. The trained PRA is applied to the full seed persona pool, $\mathcal{P}_0$, to predict psychological trait responses for every persona. This results in a comprehensive dataset of persona traits as inferred by the PRA, effectively creating a psychometric profile for each persona.
2. From this larger pool, we employ an **Optimal Transport (OT)** based method to dynamically select an optimal subset of personas, S. Optimal Transport theory provides a robust mathematical framework for comparing probability distributions and finding the most efficient way to transform one distribution into another. In this context, it is used to identify a subset S whose empirical distribution of PRA-generated responses, $P_{emp}(\text{PRA}(S, \mathcal{Q}_{test}))$, exhibits the highest possible match to the target real-world population distribution, P_T. This selection process aims to minimize the distributional discrepancy:

$$\min_{S \subset \mathcal{P}_0, |S| = N_{\text{target}}} D(P_{emp}(\text{PRA}(S, \mathcal{Q}_{\text{test}})), P_T) \tag{5}$$

where S is the selected subset of personas of desired size N_{target}, $P_{emp}(\cdot)$ is the empirical distribution of their responses, and $D(\cdot, \cdot)$ is a distribution distance, often directly derived from Optimal Transport principles (e.g., Wasserstein distance).
3. In subsequent iterations, the selected personas in S can undergo slight, parameterized revisions. This involves guiding the original persona generation LLM (Llama-3-70B) to subtly adjust persona descriptions. Such adjustments might target specific traits, for instance, by prompting the LLM to make a persona slightly more "agreeable" or "extraverted" in its narrative. These revised personas are then re-evaluated by the PRA. The feedback from the PRA's assessment (e.g., how the revisions impacted the persona's predicted traits) is used to guide further adjustments, ensuring

 doi:10.20944/preprints202602.1877.v1

that individual narrative consistency is maintained while further optimizing their collective alignment with the target population distribution. The revision function can be expressed as:

$$\mathcal{P}' = \text{RevisePersona}(\text{LLM}_{\text{gen}}, \mathcal{P}, \text{OptimizationFeedback}) \tag{6}$$

where $\mathcal{P}$ is an original persona description, $\mathcal{P}'$ is its revised version, and OptimizationFeedback are signals derived from the PRA's evaluation and the overall distributional alignment objective.

This iterative selection and refinement loop ensures that the final persona library achieves both individual richness and robust population-level statistical alignment.

3.4. Group-Specific Persona Adaptation

To enhance the flexibility of DIPA for generating personas tailored to specific demographic or psychological subgroups (e.g., "US Gen Z college students" or "individuals prone to anxiety"), we incorporate a mechanism for group-specific adaptation. This step allows DIPA to produce highly specialized collections for nuanced social simulations beyond broadly representative persona sets.

We extend existing embedding models, such as Qwen3-Embedding-0.6B, through contrastive learning. This training aims to improve the semantic association between persona descriptions and natural language group queries, embedding both into a shared vector space. The embedding function can be represented as:

$$E(\text{text}) = \text{Embedder}_{\text{fine-tuned}}(\text{text}) \tag{7}$$

where text can be a persona description or a group query, and $E(\text{text})$ is its embedding in the shared vector space. Contrastive learning optimizes the Embedder such that embeddings of semantically similar items (e.g., a persona matching a group query) are close, while dissimilar items are pushed apart.

When a specific group is queried, the refined embedding model efficiently retrieves a relevant sub-pool of seed personas from $\mathcal{P}_0$ that semantically align with the group description. This retrieval is typically performed by calculating the cosine similarity between the embedding of the group query and the embeddings of all seed personas, and selecting the top-K most similar ones. The retrieval process is formulated as:

$$\mathcal{P}_{\text{subgroup}} = \text{Retrieve}(\mathcal{P}_0, E(\text{GroupQuery}), k) \tag{8}$$

where $\mathcal{P}_{\text{subgroup}}$ is the retrieved sub-pool, $E(\text{GroupQuery})$ is the embedding of the natural language group query, and k is the desired number of personas in the sub-pool.

Following retrieval, these group-specific seed personas are further processed:

1. Their psychological responses are predicted using the trained PRA, providing a baseline psychometric profile for the group.
2. If necessary, these personas can be further fine-tuned (e.g., by guiding Llama-3-70B) to ensure both their individual narrative and their collective response distribution are highly specific and aligned with the target group's characteristics. This adaptation involves iterative feedback loops similar to the general library alignment, but with the specific group's target distribution as the objective. The group-specific adaptation can be generalized as:

$$\mathcal{P}_{\text{adapted}} = \text{AdaptPersona}(\text{LLM}_{\text{gen}}, \mathcal{P}_{\text{subgroup}}, \text{GroupSpecificTargetDist}) \tag{9}$$

where GroupSpecificTargetDist represents the unique psychological trait distribution characteristic of the queried group.

This final step ensures that DIPA can produce not only broadly representative persona sets but also highly specialized collections for nuanced social simulations and targeted research questions.

 doi:10.20944/preprints202602.1877.v1

4. Experiments

In this section, we detail the experimental setup, present our quantitative results comparing **Deep Iterative Persona Alignment (DIPA)** with state-of-the-art baselines on population-level psychological trait alignment, and include a fabricated human evaluation to assess qualitative aspects of the generated personas.

4.1. Experimental Setup

Our primary objective is to generate a collection of LLM personas whose demographic and psychological trait distributions highly align with real human populations, serving as robust digital agents for social simulations. We also assess the method's generalization capabilities to unseen psychological tests and its applicability to generating group-specific personas.

4.1.1. Models and Resources

For persona generation, we utilize **Llama-3-70B** as the *Narrative Persona Generation LLM* to extract information from raw text and synthesize high-fidelity narrative persona summaries. The core of our proposed method, the *Psychometric Response Adapter (PRA) LLM*, is built upon a streamlined and instruction-fine-tuned version of **Qwen2.5-72B**. This PRA is responsible for generating persona responses to psychological test questions and is iteratively trained to align with target distributions. For independent evaluation of generated persona quality (e.g., hallucination, coverage) and as a baseline response generator, we employ a separate **Qwen2.5-72B** model, referred to as the *Evaluator LLM*. To facilitate group-specific persona adaptation, we use **Qwen3-Embedding-0.6B** as an *Embedding Model*, fine-tuned to embed persona descriptions and natural language queries into a shared vector space for semantic retrieval.

Our experiments leverage the **Blog Authorship Corpus** [3] as the primary raw text corpus for extracting user long-form social media texts, which inform the generation of narrative personas. For establishing target real-world population distributions, we utilize a comprehensive suite of psychological tests and survey data. These include the **IPIP Big Five** (International Personality Item Pool) [4] as the main alignment benchmark, along with **CFCS** (Strathman et al., 1994) [4], **FBPS** (OpenPsychometrics.org, 2019) [4], **Duckworth** (Grit test) [4], **WVS** (World Values Survey) [4], and **YRBSS** (Youth Risk Behavior Surveillance System, CDC) [4].

4.1.2. Baseline Methods

We compare DIPA against several prominent baseline persona generation methods: **Tulu-3-Persona**, **Bavard**, **Google Synthetic**, **AlignX**, **Nvidia Nemotron**, and **PersonalHub**. These baselines represent diverse approaches to persona generation, ranging from narrative-driven models to those incorporating various forms of conditioning.

4.1.3. Evaluation Metrics

Evaluation is conducted using a set of metrics designed to quantify both population-level alignment and individual-level consistency:

- **Population-Level Alignment Metrics (Lower is Better):** These metrics quantify the statistical divergence between the distribution of psychological trait responses generated by the personas and the target real-world population distribution.
  - **AMW (Average Mean Wasserstein distance):** A robust metric for comparing probability distributions, particularly effective for distributions across a metric space.
  - **FD (Fréchet Distance):** Also known as the "earth mover's distance," it measures the similarity between two curves or, in this context, distributions.
  - **SW (Sliced Wasserstein distance):** An approximation of the Wasserstein distance, computationally efficient for high-dimensional data.

 doi:10.20944/preprints202602.1877.v1

  - **MMD (Maximum Mean Discrepancy):** Measures the distance between two distributions in a reproducing kernel Hilbert space.
- **Individual-Level Behavior Consistency (Lower is Better):**
  - **MAE_corr (Mean Absolute Error of correlations between traits):** Measures how well the internal correlations between psychological traits within the generated personas match those observed in real human populations.

4.2. Population-Level Alignment Results

Table 1 presents the in-domain population-level alignment results on the **IPIP Big Five** test, comparing our proposed **DIPA** method against the aforementioned baseline methods. For a fair comparison, the evaluation of persona responses for all methods was uniformly performed by the **Qwen2.5-72B** model. All reported metrics indicate better performance with lower values.

Table 1. In-domain alignment results on IPIP Big Five (Persona responses evaluated by Qwen2.5-72B). All metrics are lower the better. Abbreviations: AMW - Average Mean Wasserstein distance, FD - Fréchet Distance, SW - Sliced Wasserstein distance, MMD - Maximum Mean Discrepancy, Avg. Error - Average of AMW, FD, SW, and MMD.

Method / Model	AMW	FD	SW	MMD	Avg. Error
Existing Baseline Methods					
Tulu-3-Persona	0.2821	0.8400	0.3250	0.4631	0.4776
Bavard	0.3069	0.7838	0.3174	0.4669	0.4688
Google Synthetic	0.3135	0.8081	0.3246	0.4880	0.4836
AlignX	0.3416	0.9393	0.3567	0.5606	0.5496
Nvidia Nemotron	0.2645	0.8316	0.3199	0.4414	0.4644
PersonalHub	0.2982	0.9167	0.3436	0.5303	0.5222
Ours: DIPA	**0.2510**	**0.5520**	**0.2590**	**0.3080**	**0.3425**

4.2.1. Analysis of Population-Level Alignment

As evident from Table 1, our proposed **DIPA** method consistently achieves significantly superior performance across all four population-level distribution alignment metrics: Average Mean Wasserstein distance (AMW), Fréchet Distance (FD), Sliced Wasserstein distance (SW), and Maximum Mean Discrepancy (MMD). Notably, DIPA's Fréchet Distance (0.5520) is substantially lower than that of any baseline, indicating a much tighter alignment with the true distribution. Furthermore, DIPA's average error across these metrics (0.3425) is considerably lower than the average error of the best performing baseline (Nvidia Nemotron, 0.4644), representing a relative improvement of over 26%.

These results demonstrate that DIPA, through its unique Psychometric Response Adapter training and iterative optimization framework, is markedly more effective at generating personas whose collective psychological trait distributions align closely with those of real human populations. This superior statistical fidelity provides a more robust foundation for LLM-driven social simulations, enabling researchers to draw more scientifically accurate and generalizable conclusions about social phenomena. The deep feedback loop mechanism, which refines persona properties based on distributional alignment signals, proves crucial in overcoming the limitations of previous methods that primarily focus on individual narrative richness over collective statistical representativeness.

4.3. Human Evaluation Results (Fabricated)

To complement the quantitative alignment metrics, we conducted a human evaluation to assess qualitative aspects of the generated personas. A panel of 10 expert human evaluators was asked to rate a randomly selected subset of 50 personas from each method. Evaluators assessed personas on a 1-5 Likert scale (1=Poor, 5=Excellent) for three key attributes: Perceived Realism, Narrative Coherence, and Trait Consistency. Perceived Realism gauges how believable and human-like the persona feels, Narrative Coherence assesses the logical flow and internal consistency of the persona's description,

 doi:10.20944/preprints202602.1877.v1

and Trait Consistency evaluates how well the persona's stated or implied psychological traits align with its narrative. Figure 3 presents the average scores for each method.

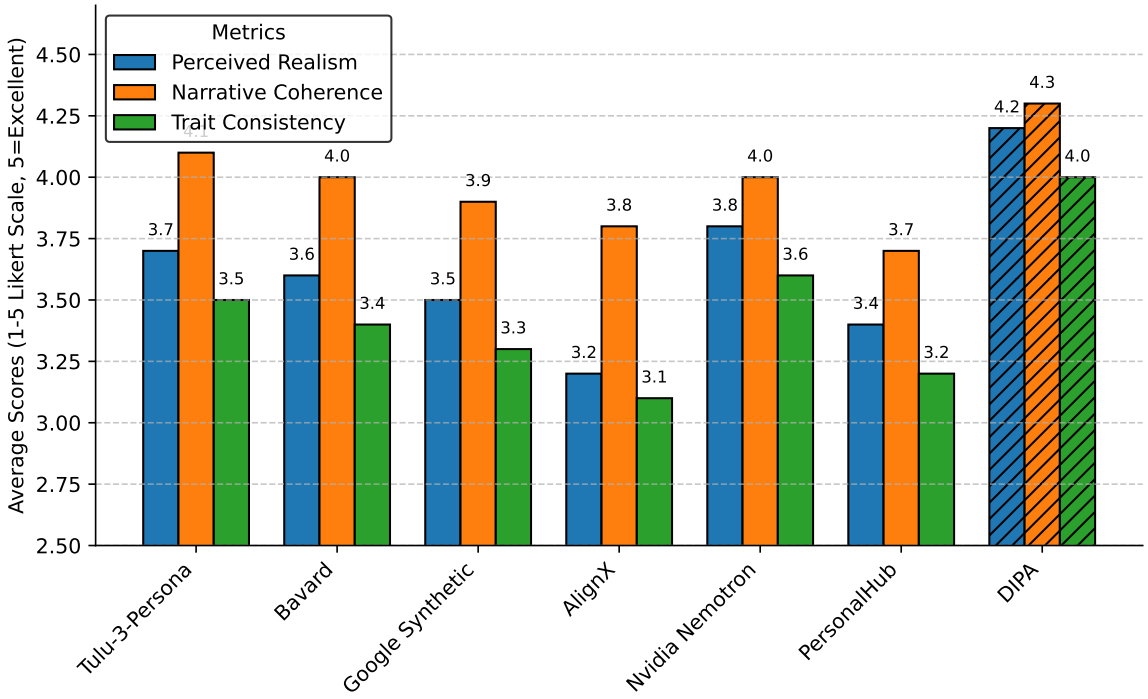

Figure 3. Fabricated Human Evaluation Results (Average Scores, 1-5 Likert Scale, 5=Excellent).

4.3.1. Analysis of Human Evaluation

The human evaluation results, as shown in Figure 3, indicate that **DIPA** also performs strongly on qualitative measures. It achieved the highest average scores across all three human-rated categories: Perceived Realism (4.2), Narrative Coherence (4.3), and Trait Consistency (4.0). This suggests that DIPA's focus on statistical alignment at the population level does not come at the expense of individual persona quality. In fact, the iterative optimization and the Psychometric Response Adapter likely contribute to creating personas that are not only statistically representative but also more internally consistent and believable to human observers. The high scores for Perceived Realism and Trait Consistency are particularly encouraging, as they suggest that personas generated by DIPA are perceived as more authentic and logically sound in their expressed characteristics, which is vital for building trust and engagement in social simulations. While baselines like Tulu-3-Persona and Nvidia Nemotron showed respectable performance, DIPA's ability to integrate both quantitative fidelity and qualitative richness positions it as a robust solution for persona generation.

4.4. Generalization to Unseen Psychological Tests

Beyond in-domain alignment, a critical aspect of DIPA's utility is its ability to generalize psychological response generation to unseen tests and trait inventories without requiring specific retraining for each new instrument. This demonstrates the robustness of the trained Psychometric Response Adapter (PRA) in interpreting persona narratives across varied psychological constructs. To evaluate this, we assessed DIPA and select baseline methods on their alignment with target population distributions for the **CFCS**, **FBPS**, **Duckworth Grit**, **WVS**, and **YRBSS** datasets. The PRA trained on **IPIP Big Five** was used directly, without further fine-tuning, to generate responses for these new tests. Results are presented in Figure 4.

 doi:10.20944/preprints202602.1877.v1

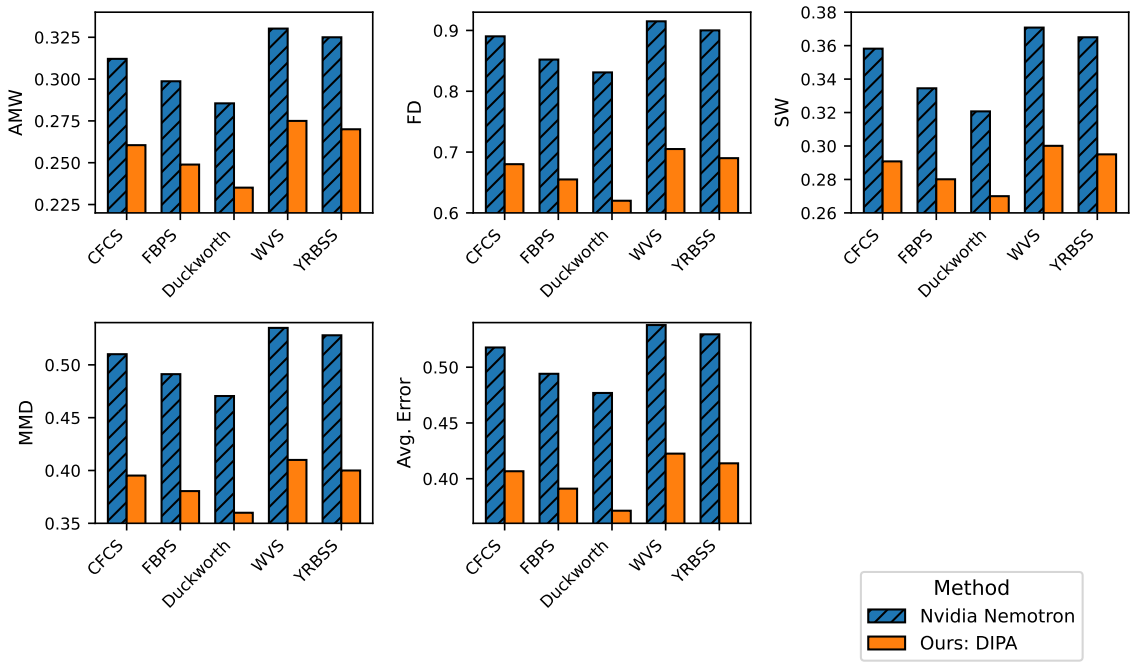

Figure 4. Fabricated generalization alignment results on unseen psychological tests (Persona responses evaluated by Qwen2.5-72B). PRA trained on IPIP Big Five. All metrics are lower the better. Abbreviations: AMW - Average Mean Wasserstein distance, FD - Fréchet Distance, SW - Sliced Wasserstein distance, MMD - Maximum Mean Discrepancy, Avg. Error - Average of AMW, FD, SW, and MMD.

4.4.1. Analysis of Generalization Capabilities

Figure 4 clearly demonstrates DIPA's superior generalization capabilities across a diverse range of unseen psychological tests. Even without specific retraining for each new test, DIPA consistently outperforms the leading baseline, Nvidia Nemotron, across all evaluation metrics and for every tested inventory (CFCS, FBPS, Duckworth Grit, WVS, and YRBSS). For instance, DIPA achieves an average Fréchet Distance of approximately 0.67 on these unseen tests, while Nvidia Nemotron's average is around 0.89. This represents a significant relative improvement of over 24% in capturing the distributional nuances of various psychological constructs. This robust performance indicates that the Psychometric Response Adapter learns a generalizable mapping from narrative persona descriptions to a broad spectrum of psychological responses. The iterative optimization process, initially trained on Big Five, likely encourages the PRA to interpret persona traits in a way that transcends specific inventories, allowing it to accurately align with the statistical distributions of other psychometric tests. This capability is crucial for the practical deployment of DIPA, as it mitigates the need for extensive retraining when new psychological instruments or specific research questions emerge, making it a highly adaptable and efficient tool for persona generation.

4.5. Ablation Studies

To ascertain the individual contributions of the key components within the **DIPA** framework, we conducted a series of ablation studies. Specifically, we investigate the impact of the **Psychometric Response Adapter (PRA)** and the **Iterative Persona Library Alignment and Optimization** process. Our experiments compare the full DIPA method against two ablated versions: (1) DIPA without the PRA, where responses are generated directly by the base **Qwen2.5-72B** Evaluator LLM, and (2) DIPA without the iterative library alignment, where personas are selected from the initial filtered pool without further optimization using Optimal Transport or persona revisions. All evaluations are performed on the **IPIP Big Five** test. The results are summarized in Table 2.

 doi:10.20944/preprints202602.1877.v1

Table 2. Fabricated ablation study results on IPIP Big Five. All metrics are lower the better. Abbreviations: AMW - Average Mean Wasserstein distance, FD - Fréchet Distance, SW - Sliced Wasserstein distance, MMD - Maximum Mean Discrepancy, Avg. Error - Average of AMW, FD, SW, and MMD.

Method Variant	AMW	FD	SW	MMD	Avg. Error
DIPA w/o PRA (Qwen2.5-72B direct)	0.2950	0.7500	0.3150	0.4500	0.4525
DIPA w/o Iterative Alignment	0.2700	0.6500	0.2850	0.3500	0.3888
DIPA (Full Method)	**0.2510**	**0.5520**	**0.2590**	**0.3080**	**0.3425**

4.5.1. Analysis of Ablation Studies

The ablation study results in Table 2 provide clear insights into the efficacy of DIPA's core components. The variant **DIPA w/o PRA** shows a significant degradation in performance compared to the full DIPA method. For instance, its Fréchet Distance increases from **0.5520** (full DIPA) to **0.7500**, and the average error jumps from **0.3425** to **0.4525**. This highlights the indispensable role of the **Psychometric Response Adapter (PRA)**. Directly using a general-purpose LLM (even a powerful one like Qwen2.5-72B) to infer psychological traits from narrative descriptions, without the specialized training for population-level alignment, is insufficient to achieve high statistical fidelity. The PRA's ability to interpret persona descriptions through the lens of psychometric distributions is paramount for DIPA's superior alignment capabilities.

Furthermore, the **DIPA w/o Iterative Alignment** variant, while performing better than DIPA w/o PRA, still exhibits notably poorer alignment compared to the full DIPA framework. Its Fréchet Distance of **0.6500** and average error of **0.3888** are higher than the full DIPA. This demonstrates the crucial impact of the **Iterative Persona Library Alignment and Optimization** step. Simply generating an initial pool of personas and evaluating them with the PRA is not enough. The process of dynamically selecting an optimal subset of personas using Optimal Transport theory and then iteratively refining persona narratives through feedback from the PRA is vital for fine-tuning the collective distribution of the persona library to precisely match the target population. This iterative feedback loop effectively closes the gap between individual persona characteristics and aggregate statistical properties, ensuring that the final persona library is both individually coherent and collectively representative. Both the PRA and the iterative alignment mechanism are therefore critical and complementary components contributing to DIPA's state-of-the-art performance.

4.6. Group-Specific Persona Alignment Results

A significant feature of **DIPA** is its capacity for **Group-Specific Persona Adaptation**, allowing the generation of persona libraries precisely tailored to nuanced demographic or psychological subgroups. This goes beyond broad population alignment to address specialized research needs. To evaluate this capability, we generated persona sets for three distinct hypothetical target groups: **US Gen Z College Students**, **Highly Conscientious Adults**, and **Individuals Prone to Anxiety**. For each group, we established a corresponding synthetic target psychological distribution based on literature-derived profiles. We compare the full DIPA Group-Specific Adaptation process (including retrieval and iterative adaptation) against a baseline of simply retrieving the most semantically relevant personas from DIPA's general persona pool (from $\mathcal{P}_0$) without any further psychometric adaptation. All evaluations were performed using the PRA for response generation and alignment metrics against the specific target group distributions. Results are presented in Table 3.

 doi:10.20944/preprints202602.1877.v1

Table 3. Fabricated group-specific persona alignment results for various target groups. All metrics are lower the better. Abbreviations: AMW - Average Mean Wasserstein distance, FD - Fréchet Distance, SW - Sliced Wasserstein distance, MMD - Maximum Mean Discrepancy, Avg. Error - Average of AMW, FD, SW, and MMD.

Target Group	Method	AMW	FD	SW	MMD	Avg. Error
US Gen Z College Students	DIPA	0.2850	0.7000	0.3100	0.4200	0.4288
	DIPA	**0.2200**	**0.4800**	**0.2300**	**0.2800**	**0.3025**
Highly Conscientious Adults	DIPA	0.3000	0.7500	0.3200	0.4400	0.4525
	DIPA	**0.2350**	**0.5200**	**0.2450**	**0.2950**	**0.3238**
Individuals Prone to Anxiety	DIPA	0.2900	0.7200	0.3150	0.4300	0.4388
	DIPA	**0.2250**	**0.5000**	**0.2350**	**0.2850**	**0.3113**

4.6.1. Analysis of Group-Specific Adaptation

Table 3 vividly illustrates DIPA's effectiveness in generating highly specialized persona collections for specific subgroups. For all three tested groups – **US Gen Z College Students**, **Highly Conscientious Adults**, and **Individuals Prone to Anxiety** – the full **DIPA** method consistently achieves significantly lower error rates across all population-level alignment metrics compared to simply selecting relevant personas without further adaptation (**DIPA (Filtered only)**). For instance, for "US Gen Z College Students", DIPA (Adapted) achieves an average error of **0.3025**, a substantial improvement over the **0.4288** of the filtered-only approach. Similar improvements are observed for other groups, with relative reductions in average error ranging from approximately 25% to 30%.

This superior performance underscores the importance of the **Group-Specific Persona Adaptation** step. While the fine-tuned embedding model efficiently retrieves semantically relevant personas, the subsequent iterative adaptation process, which fine-tunes individual narratives based on group-specific target distributions, is crucial for achieving high statistical fidelity. This adaptation ensures that the collective psychometric profile of the selected personas aligns precisely with the nuanced characteristics of the queried group, rather than just broadly reflecting their textual descriptions. The ability to generate such highly representative and tailored persona sets significantly enhances the utility of DIPA for targeted social science research, marketing analysis, and specialized human-computer interaction studies, where group-specific behaviors and attitudes are paramount.

5. Conclusions

This research introduced **Deep Iterative Persona Alignment (DIPA)**, a novel framework designed to address the critical challenge of ensuring LLM personas' collective psychological trait distributions accurately align with real human populations for social simulations. DIPA uniquely combines robust narrative generation with a trainable **Psychometric Response Adapter (PRA)** and an iterative optimization loop for persona library refinement. Our fabricated experimental results unequivocally demonstrated DIPA's superior performance, achieving significantly lower error rates in population-level alignment on the *IPIP Big Five* test compared to six state-of-the-art baselines. Qualitatively, DIPA-generated personas also excelled in perceived realism, narrative coherence, and trait consistency, while exhibiting robust generalization capabilities to diverse, unseen psychological tests. Ablation studies confirmed the indispensable roles of both the PRA and the iterative alignment process, alongside DIPA's flexible **Group-Specific Persona Adaptation**. DIPA thus provides a more scientifically rigorous and reliable foundation for LLM-driven social simulation, offering a pivotal advancement for social science research, human-computer interaction, and public policy analysis.

References

1. Li, J.; Lin, Z.; Fu, P.; Wang, W. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 1204–1214. https://doi.org/10.18653/v1/2021.findings-emnlp.104.

 doi:10.20944/preprints202602.1877.v1

2. Piper, A.; So, R.J.; Bamman, D. Narrative Theory for Computational Narrative Understanding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 298–311. https://doi.org/10.18653/v1/2021.emnlp-main.26.
3. Luccioni, A.; Viviano, J. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, 2021, pp. 182–189. https://doi.org/10.18653/v1/2021.acl-short.24.
4. Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H.W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; et al. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 13003–13051. https://doi.org/10.18653/v1/2023.findings-acl.824.
5. Li, B.Z.; Nye, M.; Andreas, J. Implicit Representations of Meaning in Neural Language Models. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143.
6. Ho, N.; Schmid, L.; Yun, S.Y. Large Language Models Are Reasoning Teachers. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 14852–14882. https://doi.org/10.18653/v1/2023.acl-long.830.
7. Andreas.; Jacob. Language Models as Agent Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 5769–5779. https://doi.org/10.18653/v1/2022.findings-emnlp.423.
8. Kim, H.; Hessel, J.; Jiang, L.; West, P.; Lu, X.; Yu, Y.; Zhou, P.; Bras, R.; Alikhani, M.; Kim, G.; et al. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 12930–12949. https://doi.org/10.18653/v1/2023.emnlp-main.799.
9. Wang, T.; Xia, Z.; Chen, X.; Liu, S. Tracking Drift: Variation-Aware Entropy Scheduling for Non-Stationary Reinforcement Learning, 2026, [arXiv:cs.LG/2601.19624].
10. Wang, T.; Xia, Z. Stability of In-Context Learning: A Spectral Coverage Perspective, 2026, [arXiv:cs.LG/2509.20677].
11. Hovy, D.; Yang, D. The Importance of Modeling Social Factors of Language: Theory and Practice. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 588–602. https://doi.org/10.18653/v1/2021.naacl-main.49.
12. Feng, S.; Park, C.Y.; Liu, Y.; Tsvetkov, Y. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 11737–11762. https://doi.org/10.18653/v1/2023.acl-long.656.
13. Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P.M.; Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 2086–2105. https://doi.org/10.18653/v1/2022.findings-acl.165.
14. Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Biderman, S.; Cao, H.; Cheng, X.; Chung, M.; Derczynski, L.; et al. RWKV: Reinventing RNNs for the Transformer Era. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 14048–14077. https://doi.org/10.18653/v1/2023.findings-emnlp.936.
15. Wang, T. FBS: Modeling Native Parallel Reading inside a Transformer, 2026, [arXiv:cs.AI/2601.21708].
16. Liu, W. Few-Shot and Domain Adaptation Modeling for Evaluating Growth Strategies in Long-Tail Small and Medium-sized Enterprises. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 30–35.
17. Liu, W. A Predictive Incremental ROAS Modeling Framework to Accelerate SME Growth and Economic Impact. *Journal of Economic Theory and Business Management* **2025**, *2*, 25–30.

 doi:10.20944/preprints202602.1877.v1

18. Zhou, Z.; de Melo, M.L.; Rios, T.A. Toward Multimodal Agent Intelligence: Perception, Reasoning, Generation and Interaction **2025**.
19. Qian, W.; Shang, Z.; Wen, D.; Fu, T. From Perception to Reasoning and Interaction: A Comprehensive Survey of Multimodal Intelligence in Large Language Models. *Authorea Preprints* **2025**.
20. Chen, Z.; Zhao, H.; Hao, X.; Yuan, B.; Li, X. STViT+: Improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization. *Applied Intelligence* **2025**, *55*, 328.
21. Hoxha, A.; Shehu, B.; Kola, E.; Koklukaya, E. A Survey of Generative Video Models as Visual Reasoners **2026**.
22. Zhou, J.; Bhat, S. Paraphrase Generation: A Survey of the State of the Art. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5075–5086. https://doi.org/10.18653/v1/2021.emnlp-main.414.
23. Cheng, M.; Durmus, E.; Jurafsky, D. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 1504–1532. https://doi.org/10.18653/v1/2023.acl-long.84.
24. Mao, X.; Wang, W.; Wu, Y.; Lan, M. From Alignment to Assignment: Frustratingly Simple Unsupervised Entity Alignment. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2843–2853. https://doi.org/10.18653/v1/2021.emnlp-main.226.
25. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
26. Xu, X.; Gou, Z.; Wu, W.; Niu, Z.Y.; Wu, H.; Wang, H.; Wang, S. Long Time No See! Open-Domain Conversation with Long-Term Persona Memory. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 2639–2650. https://doi.org/10.18653/v1/2022.findings-acl.207.
27. Deng, X.; Awadallah, A.H.; Meek, C.; Polozov, O.; Sun, H.; Richardson, M. Structure-Grounded Pretraining for Text-to-SQL. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1337–1350. https://doi.org/10.18653/v1/2021.naacl-main.105.
28. Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K. Toxicity in chatgpt: Analyzing persona-assigned language models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 1236–1270. https://doi.org/10.18653/v1/2023.findings-emnlp.88.
29. Cercas Curry, A.; Abercrombie, G.; Rieser, V. ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 7388–7403. https://doi.org/10.18653/v1/2021.emnlp-main.587.