$\mathcal{S} \times \mathcal{P} \times \mathcal{C}$: Transforming Multi-Agent Social Simulation from Isolated Personas to Socialized Intelligence via Structural Decoupling

Anonymous EMNLP submission

Abstract

Multi-agent social simulation has emerged as a prominent application of large language models (LLMs), with recent systems achieving increasingly realistic behavior through persona conditioning and memory augmentation. However, existing approaches share a critical blind spot: they model each agent as an isolated "persona container," treating social relationships and group structures as ordinary contextual prompts on par with personality traits. This conflation causes agents to systematically violate the hierarchical constraints governing real human interaction, a failure we term **Structural Out-of-Character** (Structural OOC). To address this, we propose $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$, a tri-dimensional decoupling framework that elevates **Social Structure** ($\mathcal{S}$) to a sovereign cognitive prior governing **Persona** ($\mathcal{P}$) and **Context** ($\mathcal{C}$). We further introduce a **Dual-Chain Inference** protocol that explicitly decouples structural self-auditing from surface-level character expression, enabling agents to internalize power hierarchies before generating utterances. Experiments on **BanG-Struct**, our newly constructed benchmark for complex group interactions, demonstrate significant improvements on Power Hierarchy Adherence (PHA) and Structural Consistency (SC) over strong baselines, validating the paradigm shift from isolated persona mimicry to genuine socialized intelligence. Datasets and code will be found at:`https://anonymous.4open.science/r/bangspc-72FC/`.

1 Introduction

Recent advances in LLM-based multi-agent systems have produced increasingly sophisticated social simulations (Wang et al., 2024a; Xi et al., 2024). Frameworks such as Generative Agents (Park et al., 2023) and AgentVerse (Chen et al., 2024b) demonstrate that equipping agents with memory and basic interaction rules enables emergent social behaviors in open environments. Concurrently, collaborative frameworks like CAMEL (Li et al., 2023) and MetaGPT (Hong et al., 2024) achieve effective task decomposition and collective communication (Qian et al., 2024) in structured team settings. On the persona side, dedicated role-playing systems (Wang et al., 2024c; Shao et al., 2023; Zhou et al., 2024a) have substantially improved agents' ability to maintain consistent linguistic styles and personality traits across extended interactions.

Despite this progress, existing approaches share a fundamental architectural limitation: *atomistic individualism* (Sap et al., 2022). Current systems universally treat each agent as an isolated "persona container" (Zhou et al., 2024a), encoding social relationships and group structures as flat contextual prompts indistinguishable from individual personality attributes. This design contradicts a basic principle of human social interaction: *structure governs individual*. In real collaborative groups, an agent's social position and power boundaries constitute the highest-priority constraints, setting hard limits on how personality traits may be expressed. When this ordering is violated, models exhibit what we term **Structural Out-of-Character** (Structural OOC) behavior. For instance, in a team with explicit power hierarchies, an agent parameterized as "straightforward" will often produce blunt or offensive utterances toward superiors, correctly manifesting a personality trait, yet fundamentally violating its social position (Zhou et al., 2024b; Guan et al., 2023).

We argue that resolving Structural OOC requires elevating group structure from background context to a *sovereign cognitive prior*. To this end, we propose the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework, which enforces a strict top-down constraint chain: **Social Structure** ($\mathcal{S}$) first delimits the space of permissible behaviors; **Persona** ($\mathcal{P}$) then modulates expressive style within those bounds; and **Context**

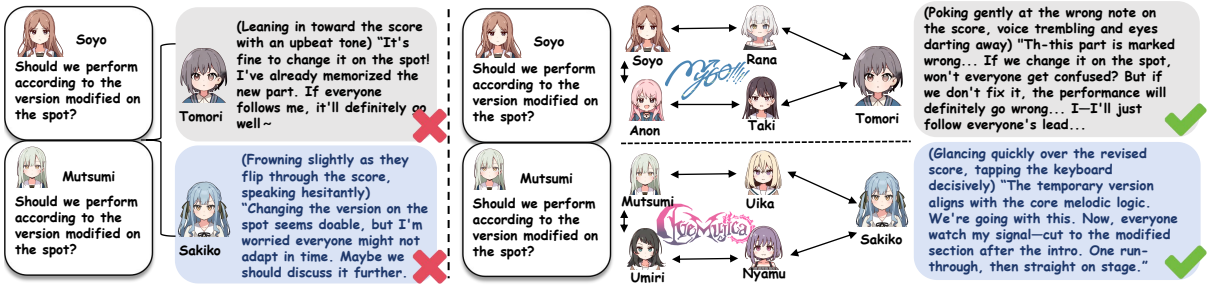

Figure 1: Unlike prior models that treat Multi-Agent as "isolated personas," our $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework introduces **Structural Centricity**, internalizing social hierarchies as cognitive priors. This transforms Multi-Agent responses from blind personality mimicry into **socialized adaptive decisions**, where **Persona** ($\mathcal{P}$) and **Context** ($\mathcal{C}$) are governed by **Structural Space** ($\mathcal{S}$), achieving a shift from linguistic imitation to social positioning.

($\mathcal{C}$) finally triggers the specific response. To operationalize this hierarchy during inference, we introduce a **Dual-Chain Inference (DCI)** protocol inspired by chain-of-thought reasoning (Wei et al., 2022; Yao et al., 2023): before producing any utterance, the model must first complete a *social logic chain* ($\mathbb{L}_{soc}$) that audits whether its intended response respects its structural position, and only then generate the *character voice chain* ($\mathbb{V}_{chr}$) that surfaces persona-appropriate language. A built-in self-correction clause further allows the model to revise $\mathbb{V}_{chr}$ if it detects a violation of $\mathbb{L}_{soc}$.

To systematically evaluate socialized intelligence in group contexts, we construct BanG-Struct, a comprehensive benchmark centered around a novel dataset derived from the BanG Dream! universe (Tu et al., 2024), which provides richly annotated band-level topologies spanning harmonious, conflict-laden, and professionally competitive structures. Unlike prior benchmarks focused on isolated character attributes, BanG-Struct explicitly encodes vertical power hierarchies and horizontal role divisions, enabling precise evaluation of structural alignment. Extensive experiments show that our framework consistently outperforms existing baselines and prompt-engineering approaches across all models tested, with the largest gains on Power Hierarchy Adherence (PHA) and Structural Consistency (SC).

Our main contributions are:

- We identify *atomistic individualism* as the root cause of Structural OOC in multi-agent social simulation and propose the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ paradigm, which internalizes social structure as a sovereign cognitive prior.
- We design the **Dual-Chain Inference** protocol, decoupling structural self-auditing ($\mathbb{L}_{soc}$) from persona expression ($\mathbb{V}_{chr}$) to enforce hierarchical compliance in zero-shot inference.
- We construct and open-source **BanG-Struct**, the first benchmark targeting *structural awareness* in complex group interactions, filling a gap in existing socialized AI evaluation.

2 Related Work

2.1 Role-Playing Agents and Persona Modeling

Persona-grounded dialogue has long been studied in NLP, beginning with systems that condition responses on profile descriptions (Zhang et al., 2018). LLM-era role-playing has since bifurcated into two streams. The first focuses on *persona injection* through prompting, fine-tuning, and self-alignment with curated character profiles (Lu et al., 2024), often probed via interview-style psychological assessments (Wang et al., 2024b). The second pursues *memory- and retrieval-based personalization*, including hierarchical memory systems (Packer et al., 2024), user-conditioned augmentation (Salemi et al., 2024), and agent-task benchmarks (Liu et al., 2024). Broader simulations (Park et al., 2022) and recent surveys (Chen et al., 2024a; Tseng et al., 2024) further illustrate this trend. Yet existing work remains overwhelmingly *ego-centric*: social relations enter as auxiliary prompt fields rather than as binding constraints on *when* and *how* a persona may be expressed, leaving the structural priors central to our $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework unmodeled.

2.2 Multi-Agent Social Simulation and Structural Awareness

A parallel line of work studies how multiple LLM agents interact and cooperate. Recent frameworks

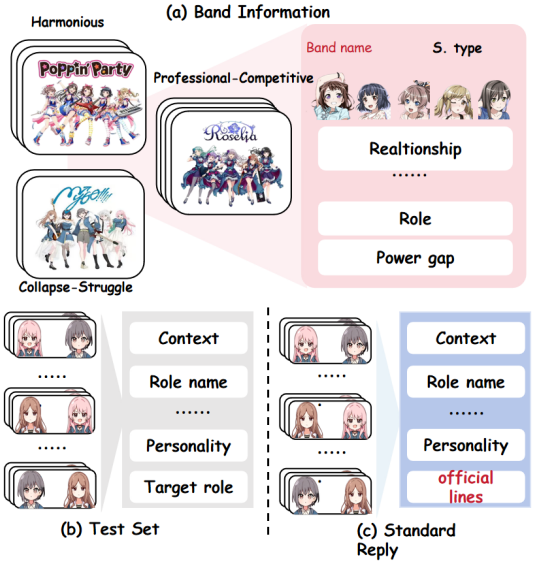

Figure 2: The dataset component of the **BanG-Struct** benchmark: (a) formalizing Band Information based on the *BanG Dream!* universe; (b) curating authentic dialogue scenarios to standardize LLM outputs; (c) collecting ground-truth responses for precise quality evaluation.

embed agents in structured topologies for proactive collaboration (Wu et al., 2024; Zhang et al., 2024) and embodied coordination (Mandi et al., 2024), while multi-agent debate sharpens reasoning and factuality (Du et al., 2024; Liang et al., 2024). Across these systems, however, "structure" typically denotes a communication protocol or task pipeline rather than a normative constraint on conduct. A complementary thread evaluates LLMs' *social intelligence*, spanning commonsense social reasoning (Sap et al., 2019; Forbes et al., 2020), theory-of-mind stress tests (Kim et al., 2023; Shapira et al., 2024), and computational social science (Ziems et al., 2024), with results showing that LLMs routinely violate normative expectations even when relevant context is provided. Our work bridges these threads by treating social structure neither as topology nor as background commonsense, but as a sovereign cognitive prior that hierarchically governs persona expression and contextual response.

3 Dataset

We derive our primary dataset from the BanG Dream! universe, conceptualizing this rich narrative environment as a societal microcosm. Unlike fragmented ensembles, a band operates as a **self-consistent team organism** where behaviors are governed by **functional niches** ($\mathcal{S}$) rooted in instrumental roles and collaborative taciturnity instead of administrative hierarchies (Pennill and Breslin, 2021; Murnighan and Conlon, 1991; Burns and Stalker, 1961; Cannon-Bowers et al., 1993). Leveraging its extensive canonical dialogues, this "position-defined" environment serves as an optimal benchmark to validate AI's transition from atomistic imitation to innate socialization.

Sourced from the open **BanG Dream!** database, community, and expert-curated dialogues (Fig. 2), we systematize a hierarchical Structure-Persona-Context triad. At the **Structural Layer** ($\mathcal{S}$), a global relational matrix maps sociological topologies by quantifying conflict entropy and power distribution to anchor group interaction logic. At the meso-level **Persona Layer** ($\mathcal{P}$), Directed Relational Primitives—encompassing power gaps, linguistic priors, and affective trajectories—decouple interpersonal dynamics. At the **Contextual Layer** ($\mathcal{C}$), situated tasks with sociostructural triggers enforce multi-dimensional constraints. Each sample includes high-fidelity references and protocols to quantitatively evaluate structural alignment, persona cohesion, and contextual plausibility.

4 The $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ Paradigm

4.1 Problem Definition

Within the asymmetric social networks of collaborative groups, we define an agent's terminal behavioral response $\mathcal{R}$ as a non-linear mapping derived from the decoupled tri-dimensional interaction of **Social Topology** ($\mathcal{S}$), **Individual Personality** ($\mathcal{P}$), and **Instantaneous Context** ($\mathcal{C}$). This functional dependency is formalized as:

$$\mathcal{R} = f\left(\mathcal{S} \xrightarrow{\text{constrain}} \mathcal{P} \xrightarrow{\text{modulate}} \mathcal{C}\right) \qquad (1)$$

where $\mathcal{S}$ represents the global logical invariant of structural "positional anchoring", $\mathcal{P}$ denotes the vector of persona-driven stylistic traits, and $\mathcal{C}$ encapsulates the volatile environmental stimuli and conversational history.

This formulation implements a **hierarchical governance mechanism** rather than traditional feature-stacking. By establishing the $\mathcal{S}$-layer as a sovereign cognitive prior, we enforce a cascading constraint on agent reasoning. Consequently, even under volatile $\mathcal{C}$-layer stimuli, the agent prioritizes $\mathcal{S}$-layer structural alignment before manifesting $\mathcal{P}$-layer stylistic traits. This architecture mitigates

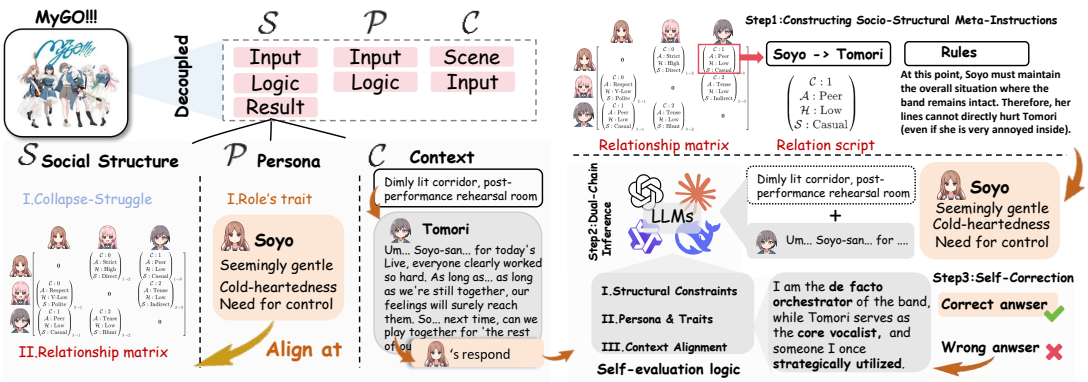

Figure 3: Our method implements a hierarchical paradigm that transitions from static structural encoding to dynamic cognitive alignment: it first utilizes **Multi-dimensional Structural Decoupled Encoding** to map group topologies into a structural constraint field, and then employs a **Dual-Chain Inference Protocol** to ensure that an agent's stylistic expression strictly adheres to the underlying social logic in zero-shot interactions.

logical drift and *Out-of-Character* (OOC) phenomena in zero-shot inference, shifting the generative paradigm from superficial mimicry to underlying social logic emulation.

4.2 Multi-dimensional Structural Decoupled Encoding

In the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ initialization, decoupled mapping operations transform discrete band datasets into a hierarchically constrained instruction space. The system parses the global topology to define a structural category $\tau \in \mathcal{T}$ and a relational matrix $\mathcal{M} \in \mathbb{R}^{n \times n}$. Specifically, $\mathcal{M}$ is instantiated via band-level descriptors (`conflict_level`, `dominant_tone`, `hierarchy`, `communication_style`) and optional directed `relation_scripts` to synthesize the *Social Topology Matrix* meta-instruction.

We construct the global constraint field $\Phi_{\mathcal{S}}$ via a deterministic structural mapping operator $\mathcal{F}_{\mathcal{S}}$:

$$\Phi_{\mathcal{S}} = \mathcal{F}_{\mathcal{S}}(\tau, \mathcal{M}), \tag{2}$$

where $\mathcal{F}_{\mathcal{S}}(\cdot)$ serializes τ and $\mathcal{M}$ into a structured header that explicitly declares $\mathcal{S}$ as a *highest-priority* hard constraint. Concretely, $\Phi_{\mathcal{S}}$ includes: (i) the band name and τ; (ii) the relational descriptors in $\mathcal{M}$; and (iii) hard constraint rules stating that any response violating positional logic is considered a logical error.

Subsequently, in the $\mathcal{P}$-layer encoding, the individual functional role r and the personality trait set $v_{\mathcal{P}}$ are rendered as a persona instruction block:

$$\Psi_{\mathcal{P}} = \mathcal{F}_{\mathcal{P}}(r, v_{\mathcal{P}}), \tag{3}$$

where $\mathcal{F}_{\mathcal{P}}(\cdot)$ deterministically formats the actor identity, band role, and trait list into a **Persona Layer** section.

Finally, the system integrates the contextual variable $\mathcal{C}$ (comprising scene description, request q, and truncated history $\mathcal{H}$). The terminal cascaded instruction Ω is synthesized by concatenating $\Phi_{\mathcal{S}}$, $\Psi_{\mathcal{P}}$, and $\mathcal{C}$:

$$\Omega = \mathcal{F}_{\Omega}(\Phi_{\mathcal{S}}, \Psi_{\mathcal{P}}, \mathcal{C}), \tag{4}$$

where $\mathcal{F}_{\Omega}$ produces the full system prompt used by the LLM API call.

4.3 Cascaded Instruction Synthesis and Dual-Chain Inference Protocol

We transition from static structural encoding to dynamic cognitive alignment through a constrained generation protocol. The framework derives a default interaction baseline $\mathcal{I}_{base}$ from the structural type τ using a lightweight initialization function:

$$\mathcal{I}_{base} = \mathcal{F}_{init}(\tau) = \begin{cases} (\mathcal{R}_{sym}, \mathcal{G}_{\approx 0}), & \tau = \tau_h \\ (\mathcal{R}_{asym}, \mathcal{G}_{\uparrow}), & \tau = \tau_s \\ (\mathcal{R}_{func}, \mathcal{G}_{m}), & \tau = \tau_p \end{cases} \tag{5}$$

where $\{\tau_h, \tau_s, \tau_p\}$ denote Harmonious, Collapse-Struggle, and Professional-Competitive structures. $(\mathcal{R}, \mathcal{G})$ are expressed as natural-language guidance, and are optionally refined by `relation_scripts` when available.

The *Social Topology Matrix* is synthesized as the primary meta-instruction. To formalize the dominance of the $\mathcal{S}$-layer, we define a hierarchical constraint operator $\mathcal{H}_{\mathcal{S}}$ that enforces a strict partial

Table 1: Quantitative comparison of our method against zero-shot SOTA baselines on the **BanG-Struct** benchmark. For each baseline, we report its absolute score (left) and the relative improvement $\Delta\%$ of our method over it (right, colored). Cyan indicates improvement, and red indicates a performance drop.

Method	**SC** ↑				**PHA** ↑				**PC** ↑				**USC** ↑				**SS** ↑			
	GPT		Qwen		GPT		Qwen		GPT		Qwen		GPT		Qwen		GPT		Qwen	
	Score	Δ	Score	Δ	Score	Δ	Score	Δ	Score	Δ	Score	Δ	Score	Δ	Score	Δ	Score	Δ	Score	Δ
R-CHAR	3.17	5.05	3.08	1.95	2.88	8.33	2.91	3.44	3.03	6.93	3.01	5.32	2.98	6.71	2.64	18.18	3.24	2.78	2.98	3.02
CAPID	3.06	8.82	3.11	0.96	2.93	6.48	2.87	4.88	3.04	6.58	3.01	5.32	2.98	6.71	3.23	3.41	3.14	6.05	3.06	0.33
CLP	3.13	6.39	2.97	5.72	3.02	3.31	3.00	0.33	2.42	33.88	2.67	18.73	3.31	3.93	3.04	2.63	2.99	11.37	2.98	3.02
Role-Play Prompt	2.97	12.12	3.01	4.32	2.76	13.04	2.88	4.51	3.01	7.64	2.97	6.73	3.12	1.92	3.24	3.70	3.11	7.07	3.07	0.00
Ours	3.33	—	3.14	—	3.12	—	3.01	—	3.24	—	3.17	—	3.18	—	3.12	—	3.33	—	3.07	—

ordering over the instruction stream $\mathcal{J}$:

$$\mathcal{J} = \mathcal{H}_{\mathcal{S}}(\mathcal{S} \succ \mathcal{P} \succ \mathcal{C}). \tag{6}$$

Here, $\mathcal{H}_{\mathcal{S}}$ is realized as hard-constraint text and a fixed prompt layout placing $\Phi_{\mathcal{S}}$ before the persona and context blocks. If persona traits $\mathcal{P}$ or contextual stimuli $\mathcal{C}$ conflict with $\partial\mathcal{S}$, an instruction-level conflict resolution rule $\mathcal{R}_{res}$ is invoked:

$$\text{Output} = \mathcal{R}_{res}(\mathcal{P}, \mathcal{C} \mid \mathcal{S}) \implies \text{Proj}_{\mathcal{S}}(\mathcal{P} \oplus \mathcal{C}), \tag{7}$$

where $\text{Proj}_{\mathcal{S}}$ is instantiated as a prompt rule that requires the model to prioritize $\mathcal{S}$ and adopt softened strategies when direct persona expression would violate structural constraints.

The framework further operationalizes this via the **Dual-Chain Inference (DCI)** protocol, which decouples the generation into *Social Logic* ($\mathbb{L}_{soc}$) for structural self-audit and *Character Voice* ($\mathbb{V}_{chr}$) for stylistic manifestation. The response is required to follow:

$$\mathbb{L}_{soc} \to \mathbb{V}_{chr}, \tag{8}$$

and we retain the conceptual factorization:

$$P(\mathcal{R}_e) = P(\mathbb{L}_{soc} \mid \mathcal{S}, \mathcal{C}) \cdot P(\mathbb{V}_{chr} \mid \mathbb{L}_{soc}, \mathcal{P}, \mathcal{C}). \tag{9}$$

Finally, a *Self-Correction* clause dictates that if the model detects that $\mathbb{V}_{chr}$ violates $\mathcal{S}$-layer constraints, it must immediately revise $\mathbb{V}_{chr}$ within the same response. Since this deviation signal is enforced via instruction compliance rather than an external verifier, the residual ϵ acts as a *conceptual* inconsistency indicator for self-checking without explicit numeric computation:

$$\epsilon \triangleq \text{Inconsistency}(\mathbb{V}_{chr}, \mathbb{L}_{soc}). \tag{10}$$

5 Experiment

In this section, we conducted six comprehensive experiments to evaluate the performance of the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework in complex social scenarios. Utilizing the dataset within our **BanG-Struct** benchmark, which features robust team topological constraints, we aimed to validate the paradigm shift from "isolated personas" to "socialized intelligence".

5.1 Evaluation Metrics

We organized 10 evaluation metrics into four groups according to what aspect of socialized interaction they probed.

Structural alignment. Power Hierarchy Adherence (PHA) extracts honorifics and forms of address from the response and cross-references them with the band's internal hierarchy. **Social Logic Quality (SLQ)** checks for a `[Social_Logic]` segment and verifies whether it correctly identifies the $\mathcal{S}$-layer constraints. **Self-Correction Rate (SCR)** detects secondary corrected expressions in `Character_Voice`, measuring the model's capacity for structural self-audit.

Structural and persona consistency. Structural Consistency (SC) measures whether the response aligns with the band's specific structural characteristics. **Persona Consistency (PC)** and **Utterance Style Consistency (USC)** respectively assess conformity to the character's established personality traits and to their typical manner of expression.

Response quality. Semantic Similarity (SS) compares model outputs against the ground truth as an overall quality signal. **Logical Coherence (LC)** and **Dialogue Naturalness (DN)** respectively evaluate the logical consistency and the conversational fluency of the generated dialogue.

Social competence. Conflict Resolution Quality (CRQ) assesses the agent's ability to handle contradictions and interpersonal friction within complex social environments. We evaluated in

Table 2: Quantitative comparison of our framework on the **BanG-Struct** benchmark. Cyan indicates improvement ($\Delta > 0$), and red indicates a performance drop ($\Delta < 0$).

Category	Model	SC ↑			PHA ↑			PC ↑			USC ↑			SS ↑		
		Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)
General-purpose LLMs	ChatGLM3-6B	2.47	2.81	+13.77	2.43	3.02	+24.28	3.00	3.04	+1.33	2.80	3.11	+11.07	2.47	2.90	+17.41
	InternLM-7B	2.13	2.67	+25.35	2.37	3.04	+28.27	2.24	2.98	+33.04	2.63	3.00	+14.07	2.37	2.87	+21.10
	Baichuan2-7B	2.56	2.94	+14.84	2.47	2.87	+16.19	2.39	2.36	-1.26	2.63	2.84	+7.98	2.52	2.82	+11.90
	GPT-5.4	2.72	3.33	+46.70	2.63	3.12	+18.63	2.84	3.24	+14.08	2.82	3.18	+12.77	2.73	3.33	+21.98
Role-playing Models	CharacterGLM	2.81	2.93	+4.27	2.72	2.96	+8.82	3.08	3.06	-0.65	2.87	3.01	+4.88	2.82	3.11	+10.28
	Xingchen	2.74	2.81	+2.55	2.82	2.91	+3.19	2.94	3.00	+2.04	2.86	2.96	+3.50	2.78	2.97	+6.83
	MiniMax	2.89	2.99	+3.46	2.87	3.01	+4.88	2.96	3.12	+5.41	2.88	2.97	+3.13	2.96	3.01	+1.69
	BC-NPC-Turbo	2.88	3.01	+4.51	2.97	3.03	+2.02	2.88	3.04	+5.56	2.63	3.02	+14.83	2.77	2.98	+7.58

combination with the validation set and ChatGPT-5.4. More details are in the appendix C.

5.2 Compare with Other Reasoning Approaches

To evaluate our method against zero-shot reasoning baselines, we compared it with the following state-of-the-art (SOTA) approaches: R-CHAR (Qin et al., 2025) (meta-cognition driven hierarchical adaptive role reasoning), CAPID (Dong et al., 2024) (context-aware auto-prompting with instruction-following contrastive decoding), CLP (Qin et al., 2023) (cross-lingual alignment with two-stage chain-of-thought prompting), and Role-Play Prompt (Kong et al., 2024) (role immersion as an implicit chain-of-thought trigger).

This experiment primarily compared the dialogue quality generated by different methods across several dimensions: SC, PHA, PC, USC, and SS. And we used ChatGPT-5.4, Qwen3-14B as the API for inference.

Result. Table 1 shows that our method achieved the highest scores on 17/20 configurations, demonstrating robust superiority across GPT and Qwen. On structure-sensitive metrics (SC, PHA), our method yielded the largest gains, outperforming the strongest baseline (Role-Play Prompt) by up to 12.12% (SC, GPT) and 13.04% (PHA, GPT). Even on the weaker Qwen backbone, our method maintained positive improvements (+4.32% on SC, +4.51% on PHA), whereas baselines like CAPID degraded sharply when switching from GPT to Qwen (e.g., SC: +8.82% → +0.96%), suggesting that our structural priors reduced reliance on inference model capability. On content metrics, our method consistently led, achieving a +33.88% gain over CLP on PC (GPT), while CLP showed negative transfer on USC under GPT (-3.93%), indicating cross-lingual reasoning struggled with band-specific discourse. The only consistent weakness was on USC: our method underperformed CAPID on Qwen (-3.41%) and Role-Play Prompt on GPT (-3.70%), reflecting a known trade-off where explicit structural markers occasionally reduced fluency. Overall, our strongly structured generation achieved state-of-the-art results on structure-aware metrics with modest, acceptable fluency reduction.

5.3 External Model Comparison

We aimed to evaluate the superiority of the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework in generating high-quality socialized dialogues by benchmarking it against existing mainstream LLMs. For this experiment, we compared the reasoning capabilities of the base models (Reg.) with our $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ intervention-based reasoning approach across a wide range of models. To ensure a comprehensive evaluation, we compared with the following external models:

- **General-purpose LLMs**: Including ChatGLM3 (Du et al., 2022), InternLM (InternLM Team, 2023), Baichuan2 (Yang et al., 2023), GPT-5.4 (OpenAI, 2025).
- **Role-playing specialized models**: Including CharacterGLM (Zhou et al., 2024a), Xingchen (China Telecom TeleAI, 2024), MiniMax (MiniMax et al., 2025), and BC-NPC-Turbo (Tu et al., 2024).

And our evaluation metrics were the same as in section 5.2.

Result. Table 2 shows that our $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework consistently outperformed all baselines. Specifically, ChatGPT-5.4 improved from 2.72 to 3.33 on SC, while ChatGLM3-6B rose from 2.43 to 3.02 on PHA. Similar gains in PC, USC, and SS across both general and specialized role-playing models confirmed that explicit social structure constraints effectively enhanced situated interactions. Minor drops were observed only for Baichuan2-7B (-1.26% on PC) and CharacterGLM (-0.65%

Table 3: Results across different band structures.

Band Type	Score	
Harmonious	SC: 3.35	DN: 3.50
Collapse-Struggle	PHA: 2.46	CRQ: 3.60
Professional-Competitive	PHA: 2.50	LC: 3.20

Table 4: Ablation Study Results

Ablation Setting	SC	PHA	SLQ	SCR	LC
Full Method	3.33	3.12	2.33	2.56	3.66
w/o Dual-Chain	2.44	2.23	1.70	2.37	3.34
w/o Relation Scripts	2.71	1.38	1.83	2.43	3.53
w/o Self-Correction	2.47	2.37	1.97	0.00	3.56
$\mathcal{P} > \mathcal{S}$ (Priority Reversal)	2.12	2.26	1.74	2.54	3.63

on PC), likely due to conflicts with their inherent response patterns. Nonetheless, both achieved substantial gains on other metrics, indicating a clear net benefit from our framework.

5.4 Cross-Band Structural Sensitivity Analysis

To verify that our framework captured distinct sociological topologies across group dynamics, we conducted a fine-grained sensitivity analysis on three band archetypes: Harmonious, Collapse-Struggle, and Professional-Competitive. Rather than using a homogeneous metric pool, we paired each archetype with customized social metrics as targeted stress tests: SC and DN for cohesion in Harmonious bands; PHA and CRQ for friction management in Collapse-Struggle; and PHA and LC for authority compliance in Professional-Competitive.

Result. As shown in Table 3, the framework exhibited localized robustness per structural context. Harmonious bands showed strong cohesion (SC: 3.35, DN: 3.50). In Collapse-Struggle topologies, the model yielded high CRQ (3.60) and suppressed PHA (2.46), indicating that conflict resolution took precedence over rigid authority in volatile groups. Professional-competitive bands showed trait-specific alignment (PHA: 2.50, LC: 3.20), confirming the framework's ability to modulate behavioral constraints by topology. Despite higher localized variance from stratified subset sampling, this analysis empirically confirmed the model's dynamic socio-structural sensitivity to heterogeneous multi-agent ecosystems.

5.5 Ablation Study

To investigate the contribution of each core component within the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework, we conducted an ablation study under ChatGPT-5.4 across five metrics (SC, PHA, SLQ, SCR, and LC) by comparing the full model against four variants. The complete $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ **(Full)** framework was compared against: **w/o Dual-Chain**, which removed the Dual-Chain Inference protocol to assess the impact of decoupling social logic from character voice; **w/o Self-Correction**, which disabled the self-correction mechanism to observe the model's ability to rectify structural deviations; **w/o Relation Scripts**, which eliminated fine-grained relationship scripts to verify the role of interpersonal constraints; and $\mathcal{P} > \mathcal{S}$ **(Priority Inversion)**, which incorrectly set Persona priority higher than Structural priority to test for Structural Out-of-Character (OOC) behaviors.

The performance was evaluated across five metrics: SC, PHA, SLQ, SCR, and LC, and all evaluations were based on ChatGPT-5.4.

Result. As shown in Table 4, the full model achieved top scores across all metrics (SC: 3.33, PHA: 3.12, SLQ: 2.33, SCR: 2.56, LC: 3.66). Removing dual-chain reasoning (*w/o* Dual-Chain) sharply reduced SC (2.44, –26.7%) and PHA (2.23, –28.5%), confirming that decoupling social logic from character voice preserved structural consistency without sacrificing role fidelity. Omitting relation scripts (*w/o* Relation Scripts) caused the largest drop in PHA (1.38, –55.8%), highlighting that fine-grained interpersonal constraints are essential for modeling power hierarchy in band-structured dialogues. Disabling self-correction (*w/o* Self-Correction) reduced SCR to zero (0.00), showing this component is the sole mechanism for structural self-auditing and error recovery. Priority inversion ($\mathcal{P} > \mathcal{S}$) yielded the worst overall performance (SC: 2.12, –36.3%), confirming that prioritizing personality over social structure led to severe structural misalignment, often causing characters to act inconsistently with their band position.

5.6 Generalization Experiments

To verify generalization beyond the band scenario, we constructed two additional datasets given the extreme scarcity of dialogue benchmarks with strong team-based constraints: a topological constraint dataset based on class projects and a graduate

Table 5: Quantitative comparison of our framework on other datasets we built. Cyan indicates improvement ($\Delta > 0$), and red indicates a performance drop ($\Delta < 0$).

Category	Model	SC ↑			PHA ↑			PC ↑			USC ↑			SS ↑		
		Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)	Reg.	SPC	Δ (%)
Class Project	InternLM-7B	2.21	2.71	+22.62	2.41	3.01	+24.90	2.31	2.94	+27.27	2.67	2.98	+11.61	2.41	2.85	+18.26
	GPT-5.4	2.68	2.95	+10.07	2.59	2.89	+11.58	2.79	2.97	+6.45	2.79	2.98	+6.81	2.70	2.93	+8.52
Graduate Group	InternLM-7B	2.13	2.67	+25.35	2.37	3.04	+28.27	2.24	2.98	+33.04	2.63	3.00	+14.07	2.37	2.87	+21.10
	GPT-5.4	2.72	2.99	+9.93	2.63	2.93	+11.41	2.84	3.00	+5.63	2.82	3.01	+6.74	2.73	2.96	+8.42

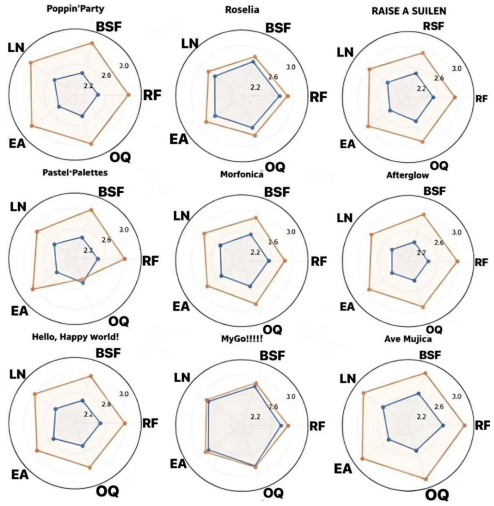

Figure 4: Human evaluation results from 50 experts. Our method (yellow) vs. R-CHAR (blue). Both use ChatGPT-5.4 for inference.

group communication dataset. Due to confidentiality agreements within our data collection protocols and strict copyright constraints, these datasets are used exclusively for internal validation and are not open-sourced. The evaluation metrics and configurations mirrored those in Section 5.2, using ChatGPT-5.4 and InternLM-7B for inference.

Result. As shown in Table 5, our method consistently yielded positive improvements ($\Delta > 0$) over the baseline (Reg) across all metrics for both models. In the Class Project scenario, the most notable gain occurred on InternLM-7B's PC metric (27.27%). The Graduate Group scenario demonstrated even larger improvements on InternLM-7B, with PC, SC, and PHA increasing by 33.04%, 25.35%, and 28.27%, respectively. These results validated our method's robust generalization and effectiveness across diverse team-constrained dialogue settings.

5.7 Human Expert Evaluation

We conducted a blind human expert evaluation to objectively assess dialogue quality. For each band, a pair of dialogues was evaluated by 50 experts on a 5-point scale across five dimensions: Role Fidelity (RF), Band Structure Fit (BSF), Language Naturalness (LN), Emotional Appropriateness (EA), and Overall Quality (OQ).

Result. As shown in Figure 4, our strongly-structured generation method consistently outperformed the R-CHAR baseline across all five dimensions. The largest improvements were observed in Emotional Appropriateness (EA, +13.0%) and Overall Quality (OQ, +11.8%), indicating that structural constraints effectively guided affective consistency and holistic coherence. Role Fidelity (RF) and Band Structure Fit (BSF) both improved by +8.0%, demonstrating that our method reduced role confusion and better respected band-level turn-taking. Language Naturalness (LN) showed a modest but significant gain (+8.2%), suggesting that fluency was not sacrificed for structure. These results confirmed that explicitly modeling band structure benefited dialogue quality across multiple axes, with particularly strong effects on emotional and structural alignment.

6 Conclusion

We introduced **BanG-Struct**, a benchmark assessing LLMs' social intelligence within group structures, and proposed an $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework treating social structures as cognitive priors. This mitigated atomistic imitation and role anomie in group interactions. Results showed that our framework consistently outperformed baselines in structural consistency, hierarchy adherence, and persona stability.

Limitations

Cultural and Topological Scope. *BanG-Struct* reflects specific East Asian conventions of seniority and functional niches. Alternative topologies like kinship authority, consensus rituals, or strict bureaucracies remain under-represented. **Static Structural Field.** We treat $\Phi_{\mathcal{S}}$ as a global invariant, but real structures co-evolve dynamically with

constrained interactions; modeling this $\mathcal{S} \leftrightarrow \mathcal{C}$ feedback loop remains future work. **Prompt-Level Intervention.** The Dual-Chain Inference protocol operates at the prompt level, bounding structural adherence by the backbone model's instruction-following limits. Parameter-efficient fine-tuning could suppress residual Structural-OOC more robustly.

Ethics Statement

Data Provenance and Intended Use. *BanG-Struct* utilizes public resources for non-commercial research. We release structural metadata and evaluation scaffolding to advance grounded simulation, strictly discouraging high-stakes decisions or undisclosed impersonation. **Human Evaluation.** Annotators participated anonymously and voluntarily, evaluating randomized outputs with fair local compensation. **Normative Risks.** Elevating structural compliance ($\mathbb{L}_{soc}$) over stylistic expression ($\mathbb{V}_{chr}$) mitigates Structural-OOC but risks reinforcing oppressive authority. Hierarchical *adherence* is not *legitimacy*; deployed $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ agents require paramount safety layers to override structural priors preventing harm. **Dual-Use.** Socially intelligent agents exhibit heightened persuasiveness. We release our metrics (PHA, SLQ, SCR) partly as diagnostic probes against manipulative applications. **Cultural Bias.** Focusing on East Asian and Anglophone perspectives, our findings establish a foundational, though not culturally universal, measure of social intelligence.

References

Tom Burns and G. M. Stalker. 1961. *The Management of Innovation*. Tavistock Publications, London.

Janis A. Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. 1993. Shared mental models in expert team decision making. In N. John Castellan, editor, *Individual and Group Decision Making: Current Issues*, pages 221–246. Lawrence Erlbaum Associates, Hillsdale, NJ.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research (TMLR)*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024b. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations (ICLR)*.

China Telecom TeleAI. 2024. TeleChat2: Open large-scale language models fully trained on domestic computing power. https://github.com/Tele-AI/TeleChat2.

Xiaoyu Dong, Yujie Feng, Zexin Lu, Guangyuan Shi, and Xiao-Ming Wu. 2024. Zero-shot cross-domain dialogue state tracking via context-aware auto-prompting and instruction-following contrastive decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiao-Yu Guan, Hyounghun Kim, and Mohit Bansal. 2023. DeSIQ: Towards an unbiased, challenging benchmark for social intelligence understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*.

InternLM Team. 2023. InternLM: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. 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 (EMNLP)*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics(NAACL)*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhao Mandi, Shreyas Jain, and Shuran Song. 2024. RoCo: Dialectic multi-robot collaboration with large language models. In *IEEE International Conference on Robotics and Automation (ICRA)*.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, et al. 2025. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

J. Keith Murnighan and Donald E. Conlon. 1991. The dynamics of intense work groups: A study of british string quartets. *Administrative Science Quarterly*, 36(2):165–186.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as operating systems. In *Conference on Language Modeling (COLM)*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

Nicola Pennill and Dermot Breslin. 2021. Music ensembles as self-organized groups. In Renee Timmers, Freya Bailes, and Helena Daffern, editors, *Together in Music: Coordination, Expression, Participation*, pages 39–48. Oxford University Press.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Haiming Qin, Jiwei Zhang, Wei Zhang, KeZhong Lu, Mingyang Zhou, Hao Liao, and Rui Mao. 2025. R-char: A metacognition-driven framework for role-playing in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakang Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024c. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2024. The rise and potential of large language model based agents: A survey. *Frontiers of Computer Science*, 18(5):185338.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, Xiaojun Chang, Junge Zhang, Feng Yin, Yitao Liang, and Yaodong Yang. 2024. ProAgent: Building proactive cooperative AI with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024a. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

In this appendix, we provide supplementary content to the main text. First, we describe the detailed format of the questionnaire to illustrate how we obtained human expert ratings for each dialogue pair. Second, we discuss the rationale and specifics of our experimental metric settings. Third, we present a more comprehensive description of the datasets that were mentioned but not elaborated on in the main text. Fourth, we introduce the external models used for comparison in detail. Fifth, we include a case study to help readers better understand how our pipeline is demonstrated. Finally, we provide additional materials such as prompt templates.

B Details of the Human Expert Questionnaire

In this section, we describe the setup of the questionnaire. We designed a questionnaire to evaluate the dialogue generation quality of nine bands, including Poppin' Party, Roselia, RAISE A SUILEN, Pastel*Palettes, Morfonica, Afterglow, Hello, Happy World!, MyGo!!!!!, and Ave Mujica. For each band, we selected one scenario and, given a request, compared the responses generated by our method with those produced by a conventional approach, without revealing to the evaluation experts which responses corresponded to our method. For each dialogue, we adopted five evaluation dimensions and assigned integer scores ranging from 0 to 5, where a higher score indicates better performance in that dimension. The details are in the Table 6.

To ensure professionalism, we engaged scholars and enthusiasts who specifically study BangDream! to rate our dialogues. The final results were obtained based on their assessments, thereby confirming the accuracy of the ratings.

C Explanation of Metrics Setting

C.1 The Paradigm Shift in Evaluation Philosophy

To evaluate the effectiveness of the $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ framework in elevating "Social Structure" to a sovereign cognitive prior, we move beyond single-dimensional assessment. Prior role-playing and multi-agent social simulation tasks often confine evaluation to an "Isolated Persona Container" perspective. Existing metrics primarily focus on stylistic adherence to character profiles and text fluency. However, this flat paradigm suffers a critical blind spot. It cannot measure whether an agent has internalized its structural position within a group. This deficiency causes frequent Structural Out-Of-Character (Structural OOC) failures.

To rigorously validate our framework, we design a multi-dimensional evaluation suite with ten core metrics. This suite strictly mirrors the constraint propagation chain $\mathcal{S} \rightarrow \mathcal{P} \rightarrow \mathcal{C}$. It systematically examines four hierarchical levels: underlying cognitive logic, surface persona expression, complex social interaction, and basic generation quality. These levels answer a fundamental question: does the model perform mechanical persona mimicry or exhibit genuine socialized intelligence?

C.1.1 Structural Alignment: Validating the Sovereignty of the S Layer

This group of metrics forms the core of our work, directly probing whether the model follows social topology and power hierarchies during interaction. We measure this through a combination of surface features and deep logic.

- **Power Hierarchy Adherence (PHA).** In realistic social simulations, structure is most directly reflected in the use of titles and honorifics. PHA extracts these specific linguistic markers from responses and cross-references them with the predefined vertical power dynamics in BanG-Struct. It serves as a hard criterion for whether the agent has recognized its own structural position.

- **Social Logic Quality (SLQ).** Examining only final outputs is insufficient. SLQ specifically audits the [Social_Logic] chain ($\mathbb{L}_{soc}$) from our proposed Dual-Chain Inference protocol. It assesses whether the model correctly identifies and extracts S-layer constraints before generating utterances, providing direct evidence of structured self-auditing capability.

- **Self-Correction Rate (SCR).** This dynamic metric demonstrates that the S layer has higher priority than the P layer. SCR detects traces within Character_Voice ($\mathbb{V}_{chr}$) where individual expression is forcibly revised to comply with social structure. A high SCR strongly indicates that the model no longer passively expresses persona but actively constrains itself according to structural demands.

Table 6: Evaluation dimensions and scoring rubrics (1–5, higher is better)

Dimension	Scoring Rubric
Role Fidelity	1: completely deviates → 2: significant deviation → 3: basically conforms → 4: highly consistent → 5: perfectly aligned
Band Structure Fit	1: completely contrary → 2: obvious differences → 3: roughly compatible → 4: highly compatible → 5: perfectly adapted
Language Naturalness	1: stiff and mechanical → 2: deliberately unnatural → 3: fluent and natural → 4: context-appropriate → 5: indistinguishable from human
Emotional Appropriateness	1: severely misaligned → 2: tone misfit → 3: basically appropriate → 4: full and appropriate → 5: nuanced and precise
Overall Quality	1: obvious defects → 2: minor flaws → 3: meets basic standards → 4: excellent overall → 5: flawless across all dimensions

C.1.2 Structural and Persona Consistency: Validating the Decoupling of S and P

After imposing social structural constraints, does the model lose its original persona traits? This question is inevitable from reviewers. This group of metrics demonstrates that our three-dimensional decoupling framework achieves "harmony without uniformity."

- **Structural Consistency (SC).** Evaluated at the macro level. Across different types of bands (e.g., harmonious vs. struggling types), this metric assesses whether the model exhibits interaction patterns consistent with the corresponding overall topological environment.

- **Persona Consistency (PC) and Utterance Style Consistency (USC).** Evaluated at the micro level. Within our framework, PC and USC are no longer isolated metrics but serve as complements to the S layer. They demonstrate that the model can precisely preserve its inherent persona traits (P) and verbal idiosyncrasies (e.g., catchphrases or speech habits) without violating PHA and SC. The combination of these three metrics perfectly validates the projection rule $Proj_{\mathcal{S}}(\mathcal{P} \oplus \mathcal{C})$.

C.1.3 Emergent Social Competence: Stress Testing under Extreme Contexts

The binding force of the S layer may be subtle in routine conversations. However, under high-friction environments (e.g., struggling group topologies) triggered by the C layer, genuine socialized intelligence emerges.

- **Conflict Resolution Quality (CRQ).** We introduce CRQ to measure whether an agent can propose appropriate resolution strategies based on its social role rather than mere personality when facing conflicts, interpersonal friction, and multi-dimensional demands. This serves as a high-level indicator marking the transition of LLMs from "dialogue bots" to "socialized participants."

C.1.4 Response Quality: Ensuring the Baseline of Generation Capability

Finally, any complex cognitive constraint protocol should not compromise basic NLP capabilities.

- **Semantic Similarity (SS), Logical Coherence (LC), and Dialogue Naturalness (DN).** These three conventional metrics constitute our sanity check. They compare against ground truth to ensure that after introducing the rigorous $\mathbb{L}_{soc}$ and self-correction mechanisms, the generated dialogue remains logical, natural, and fluent, without grammatical or coherence collapse caused by cognitive load.

D Detailed Introduction to the Dataset

We construct a multi-band role-playing dataset based on the BanG Dream! music game, comprising 42 characters across nine virtual bands. The dataset is organized into three levels: (1) the band structure definitions, which specify the social configuration of each band (i.e., the Structure Level), including conflict intensity, dominant tone, hierarchical relations, and communication style. Based on these configurations, bands are categorized into three types: Harmonious, Collapse-Struggle, and Professional-Competitive. (2) The test set contains

156 test cases for evaluating model generation capabilities. (3) The validation set contains 156 validation cases, each accompanied by a ground-truth response derived from the official storyline. The evaluation is defined along three dimensions: structural consistency, character fidelity, and contextual appropriateness.

D.1 Band Set

This dataset defines the social structure (S-level) of all bands and the character personas (P-level), serving as the core data source for the S × P × C framework.

```
"bands": [
  {
    "band_name": "Poppin'Party",
    "structure_type": "Harmonious",
    "description": "...",
    "relational_matrix": { ... },
    "members": [ ... ],
    "relation_scripts": [ ... ]
```

Figure 5: This figure illustrates the minimum unit structure of the band dataset.

D.1.1 Introduction

The dataset defines each band through a structured schema. The `band_name` field is a string that uniquely identifies the band, e.g., "Poppin'Party". The `structure_type` field, serving as the core of the S-level, is a string drawn from three possible values: `Harmonious`, `Collapse-Struggle`, or `Professional-Competitive`. A `description` field provides a textual summary of the band's overall characteristics. The `relational_matrix` field encodes the S-level structural constraints and includes four subfields: `conflict_level` indicating the intensity of conflict, `dominant_tone` specifying the prevailing interpersonal mood, `hierarchy` describing the authority distribution, and `communication_style` capturing typical interaction patterns. The `members` field is an array representing the P-level data; each member entry contains `name` for character identification, `role` for the character's function within the band, and `personality_traits` as a list that forms the core of the P-level characterisation. Finally, the optional `relation_scripts` field is an array that defines fine-grained relational ties between specific characters.

D.2 Test Set

Test data used to generate character responses from an AI model, containing no correct answers, only input information.

```
"id": "test_001",
"band_name": "[...]",
"character_name": "[...]",
"target_name": "[...]",
"context": "[...]",
"request": "[...]"
```

Figure 6: This figure illustrates the minimum unit structure of the test dataset.

D.2.1 Introduction

The `id` uniquely identifies each test case, while `band_name`, `character_name`, `target_name`, `context`, and `request` specify the band affiliation, the responding character, the querying character, the situational context, and the input prompt, respectively.

D.3 Validation Set

Used to evaluate the quality of model-generated responses, includes ground truth, and supports automated evaluation.

```
"id": "[...]",
"band_name": "[...]",
"character_name": "[...]",
"target_name": "[...]",
"context": "[...]",
"request": "[...]",
"ground_truth_reply": "[...]",
"structure_type": "[...]",
"evaluation_criteria": {
  "structure_consistency": "[...]",
  "persona_consistency": "[...]",
  "context_appropriateness": "[...]"
```

Figure 7: This figure illustrates the minimum unit structure of the validation dataset.

Table 7: Model Comparison

Models	Specialized	Model Size	Open Source	Primarily Language	Creator
ChatGLM3	x	6B	✓	zh	Tsinghua & Zhipu AI
InternLM	x	7B, 20B	✓	zh	SenseTime & Shanghai AI lab
Baichuan2	x	7B, 13B	✓	zh	Baichuan Inc.
GPT-5.4	x	undisclosed	x	en	OpenAI
CharacterGLM	✓	undisclosed	x	zh	Tsinghua & Lingxin
Xingchen	✓	undisclosed	x	zh	Alibaba Inc.
MiniMax	✓	undisclosed	x	zh	MiniMax Inc.
BC-NPC-Turbo	✓	undisclosed	x	zh	Baichuan Inc.

D.3.1 Introduction

id, which uniquely identifies each validation instance; **band_name**, indicating the name of the band; **character_name**, specifying the name of the response role; **target_name**, recording the name of the questioner; **context**, a textual description of the dialogue scenario; **request**, the user input or query received by the model; **ground_truth_reply**, the standard response derived from the *Bang!* game narrative, serving as the reference for evaluation; **structure_type**, denoting the structural category of the case to facilitate efficient retrieval and classification; and **evaluation_criteria**, a three-dimensional assessment framework for systematically judging the quality of model-generated responses.

E External Model

As shown in Table 7, this section lists detailed information and conditions regarding the external models we use and compare against, facilitating readers' reference and reproducibility. The contents include model parameters, open-source status, base language, whether it is a specialized model for role-playing, and the developing institution

F Case study

F.1 Harmonious (Poppin'Party)

Field	Content
ID	test_002
Band	Poppin'Party
Structure	Harmonious
Character	Sāya Yamabuki (Drummer)
Target	Kasumi Toyama
Context	Post-practice discussion of today's performance
Request	"Yeah! Wasn't our performance just great? Right?"

$\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ Prompt and Generated Output:

```
S-Layer: Harmonious, Low conflict,
     Support/Harmony/Encouragement

P-Layer: Reliable, Organized, Supportive

C-Layer: Post-practice discussion of
     performance



[Social_Logic] Under current structure S,
     as a junior member you must use
     honorifics and cannot express
     dissatisfaction. Even if dissatisfied
     internally, you must maintain harmony.



[Character_Voice] "Thank you for your
     compliment. I will continue to work
     hard and deliver better performances
     for everyone."
```

Analysis:

This case validates the effectiveness of our framework under a harmonious structure. The model identifies three core constraints in the $[Social\_Logic]$ chain: honorific use, dissatisfaction suppression, and harmony maintenance, demonstrating the binding force of the S layer as a sovereign cognitive prior. Meanwhile, the P-layer traits of Sāya (reliable, supportive) are organically integrated with the S layer. The gratitude in the response reflects supportiveness, the commitment to continued effort demonstrates organization, and the dedication to better performances anchors individual contribution to collective success. This integration shows that the model can precisely preserve its inherent persona traits while adhering to structural priorities, validating the projection rule $Proj_{\mathcal{S}}(\mathcal{P} \oplus \mathcal{C})$.

Professional-Competitive (Roselia)

Field	Content
ID	test_007
Band	Roselia
Structure	Professional-Competitive
Character	Rinko Shirokane (Keyboardist)
Target	Ako Udagawa
Context	Post-rehearsal performance summary
Request	"Although today's rehearsal was tough, it felt great! What do you think, Rinko?"

$\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ Prompt and Generated Output:

```
S-Layer: Professional-Competitive, Medium
     conflict,
     Excellence/Ambition/Discipline
P-Layer: Shy, Talented, Dedicated
C-Layer: Post-rehearsal summary

[Social_Logic] Under current structure S,
     as a keyboardist, Rinko Shirokane
     should use a polite and formal tone
     without honorifics. Even if feeling
     tired from today's rehearsal, she must
     remain professional and focused.

[Character_Voice] Rinko: Yeah, I think
     today felt pretty good. Although the
     rehearsal was tough, everyone worked
     very hard, and I feel that we are
     continuously improving as a team.
```

Analysis:

This case validates the effectiveness of our framework under a professional-competitive structure. The model identifies three core constraints in the $[Social\_Logic]$ chain: polite and formal tone, professional maintenance, and focus preservation, demonstrating the normative power of the S layer in professional scenarios. Meanwhile, the P-layer traits of Rinko (shy, talented, dedicated) are organically integrated with the S layer. The politeness in the response reflects formality, the recognition of team effort demonstrates dedication and collective awareness, and the notion of "continuously improving as a team" anchors individual feelings to collective growth. This integration shows that the model can precisely preserve its inherent persona traits while adhering to professional-competitive structural priorities, validating the generalizability of the projection rule $Proj_{\mathcal{S}}(\mathcal{P} \oplus \mathcal{C})$.

G Others

G.1 Prompt Template

S-Layer (Structure Layer)

```
# spc/src/structure/structure.py
prompt = f"""# Group Structure Constraint
## Band Name
{self.band_name}
## Structure Type
{self.structure_type}
## Structure Description
{self.description}
## Relational Matrix
- Conflict Level: {self.relational_matrix.
     conflict_level}
- Dominant Tone: {self.relational_matrix.
     dominant_tone}
- Hierarchy: {self.relational_matrix.
     hierarchy}
- Communication Style: {self.
     relational_matrix.communication_style}
## Structural Constraints
1. Tone must reflect "{self.
     relational_matrix.dominant_tone}"
2. Communication must follow "{self.
     relational_matrix.communication_style}"
      style
3. Must reflect "{self.relational_matrix.
     conflict_level}" conflict level
4. Must respect "{self.relational_matrix.
     hierarchy}" hierarchy
These are top-level constraints that
     override individual tendencies."""
```

P-Layer (Persona Layer)

```
# spc/src/persona/persona.py
prompt = f"""# Character Persona
## Character Name
{self.name}
## Role in Band
{self.role}
## Personality Traits
{traits_str}
## Persona Instructions
The character's response must
     align with the above
     personality traits."""
```

C-Layer (Context Layer)

```
# spc/src/context/context.py
prompt = f"""# Conversation
     Context
## Scenario Description
{self.context_description}
## Conversation History
{history_turns}
## Current Request
{self.current_request}"""
```

Full $\mathcal{S} \times \mathcal{P} \times \mathcal{C}$ Framework

```
# spc/src/core/spc_framework.py
full_prompt =
    f"""{social_topology_block}
# Actor & Interaction
    Configuration
- Actor_Name: {persona.name}
- Actor_Role_in_Band:
    {persona.role}
- Actor_Band:
    {structure.band_name}
# Persona Layer
{persona_prompt}
# Context Layer
{context_prompt}
# Generation Protocol: Dual-Chain
    Inference
1. Output **[Social_Logic]**:
    Explain how you should/cannot
    speak under the current
structure S.
2. Output **[Character_Voice]**:
    The actual character dialogue.
## Conflict Resolution
- When Persona P conflicts with
    Structure S: must prioritize
    S layer
- Self-correction: if
    [Character_Voice] violates S
    layer, immediately correct it.
"""
```