Influencing LLM Multi-Agent Dialogue via Policy-Parameterized Prompts

Hongbo Bo
University of Bristol
Bristol, United Kingdom
hongbo.bo@bristol.ac.uk

Jingyu Hu
University of Bristol
Bristol, United Kingdom
jingyu.hu@bristol.ac.uk

Weiru Liu
University of Bristol
Bristol, United Kingdom
weiru.liu@bristol.ac.uk

ABSTRACT

Large Language Models (LLMs) have emerged as a new paradigm for multi-agent systems. However, existing research on the behaviour of LLM-based multi-agents relies on ad hoc prompts and lacks a principled policy perspective. Different from reinforcement learning, we investigate whether prompt-as-action can be parameterized so as to construct a lightweight policy which consists of a sequence of state-action pairs to influence conversational behaviours without training. Our framework regards prompts as actions executed by LLMs, and dynamically constructs prompts through five components based on the current state of the agent. To test the effectiveness of parameterized control, we evaluated the dialogue flow based on five indicators: responsiveness, rebuttal, evidence usage, non-repetition, and stance shift. We conduct experiments using different LLM-driven agents in two discussion scenarios related to the general public and show that prompt parameterization can influence the dialogue dynamics. This result shows that policy-parameterised prompts offer a simple and effective mechanism to influence the dialogue process, which will help the research of multi-agent systems in the direction of social simulation.

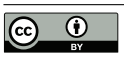
KEYWORDS

LLMs; Multiagent System; Social Simulation

ACM Reference Format:
Hongbo Bo, Jingyu Hu, and Weiru Liu. 2026. Influencing LLM Multi-Agent Dialogue via Policy-Parameterized Prompts. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/VAVC8140

1 INTRODUCTION

Traditional multi-agent simulations often rely on explicit modelling or reinforcement learning to train policies [9, 29, 33], enabling agents to learn how to respond and interact. In contrast, Large Language Models (LLMs) have emerged as a new paradigm for multi-agent systems [10, 18], where agents inherently possess natural language generation and knowledge retrieval capabilities without requiring additional training for basic interaction. Based on this potential, LLM-based multi-agent systems have been widely used in social simulation tasks by assigning roles, tasks, and instructions to LLMs in recent studies [8, 23].



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/VAVC8140

LLMs agents in these approaches typically communicate with each other using ad hoc prompts, focusing on whether agents can align human preferences and behaviors like collaboration [12, 28, 32], negotiation [1, 5, 19] and provide rational plans [6, 11]. However, these methods lack a principled framework for treating communication strategies as policies to systematically control agent behaviors. Without sufficient attention to how agent dialogue can be deliberately shaped and optimized, it becomes difficult to predict agent behavior, optimize communication patterns, and transfer insights across different tasks.

To address this limitation, we expect to propose a principled way to conceptualize and operationalize agent communication strategies, one that allows us to formally define, compare, and optimize different strategies for multi-agent dialogue. Specifically, this study focuses on how policy-parameterised prompts can be utilised to influence conversational behaviours in multi-agent discussion dialogues. This work considers the input prompt itself as an action generated by a lightweight form of policy parameterisation. Specifically, we decompose the prompt into five components: task and persona description (T), dialogue history memory (M), external knowledge base (D), rule template (R), and weight (W). By adaptively parameterising these components to allow different levels of influence on LLM agents, we can directly impose the significance of different factors on the utterance style of dialogue, thereby modulating their conversational behaviours without any additional training.

Within this policy framework, we implement multi-role-based agents with distinct stances and knowledge bases, and engage them in multi-round dialogue on issues related to the general public. To further enhance adaptivity, we design an adaptive weight scheduler that automatically adjusts the reliance on T/M/D during the dialogue, based on temporal trends and behavioural feedback. To quantify the effects of different control strategies, we propose a set of evaluation metrics, including responsiveness, rebuttal, non-repetition, evidence usage, and stance shift. These metrics allow us to systematically compare dialogue differences under various prompt control modes and to observe the evolution of group stances over time.

Our designed framework is to answer the following research questions. RQ1: Can prompt control be used as a lightweight form of policy parameterisation to regulate the conversational behaviours of LLM-based multi-agent systems? RQ2: Do different prompt control strategies (e.g., rule templates and weight scheduling) lead to significant behavioural differences, such as changes in responsiveness, evidence usage, and stance evolution? To address these questions, we design two discussion scenarios related to the general public land resources use and educational resource allocation,

each involving three agents with distinct personas driven by different LLMs. In each scenario, the agents interact over multiple rounds of dialogue, with prompts dynamically constructed from task, memory, and evidence components. We systematically vary control strategies by enabling or disabling rule templates and adjusting weight parameters, and evaluate the resulting dialogues along five proposed evaluation metrics. Our results show that prompt parameterisation can effectively influence dialogue, with different strategies leading to distinct patterns in rebuttal, evidence usage, and stance shift.

Overall, the core of this study lies in treating prompt control as a lightweight policy to achieve structured regulation of LLM multi-agent dialogue, thereby exploring a social simulation pathway that is distinct from traditional training-based approaches.

2 RELATED WORK

LLMs Agents and Social Simulation. Recent works have explored the use of LLMs as agents in social simulation, aiming to explore realistic social phenomena by simulating social behaviours. [23] applied LLMs with memory to build generative agents, and found agents exhibit emergent social behaviors such as information diffusion, relationship formation, and coordinating attendance in a sandbox environment. Similar works [7, 20, 21] also utilized LLMs to build the sandbox environment to study social behaviors. [24] developed a simulated community of 1,000 personas by design prompt chains to generate behaviors like posting, replying, and also anti-social actions. S^3 system [8] uses LLM agents to simulate users' emotions, attitudes, and interaction behaviours in the social networks, by endowing the agents in the system with the ability to perceive the informational environment. CAMEL [17] is a multi-agent role-playing framework proposed to leverage inception prompting to initialise roles, tasks, and dialogue formats, enabling LLM-based chat agents to collaborate on complex tasks. [3] applies multi-agent LLMs to simulate agent populations, and their results show that such populations can develop social conventions and collective biases through decentralized interactions. [25] extends society simulations to large scale with over 10k agents and 5 million interactions and analyzed their believable individual and emergent social behaviors. Other studies [12, 19] validated the effectiveness of multi-agent systems in collaboration and debate tasks. However, prompts in these systems are typically ad hoc, without a principled treatment as policies.

LLM Agents Formalization. Several works have sought to abstract LLM-based agents into decision-theoretic frameworks, drawing inspiration from reinforcement learning (RL) or the Belief-Desire-Intention (BDI) model. For instance, LLMs have been cast as policies mapping states to natural-language actions, and also act as a world model combined with Monte Carlo Tree Search for planning search [11]. BDIPrompting [14] integrates the BDI model into prompt design to improve proactive action planning and transparency in LLMs. Reflexion[26] is a framework that reinforces LLM agents through verbal feedback, parameterizing a policy as memory to enable trial-and-error learning via self-reflection. [6] study whether LLMs can serve as rational players in game theory, by providing inputs with preferences and rules to build belief and desire and then plan the optimal action. [30] combined LLMs with reinforcement learning (RL) to enable agents to learn strategic language communication and decision-making for the Werewolf game. These studies focus on questions—whether LLMs can act rationally or align with human decisions, but do not address how LLMs can be controlled in multi-agent. Also, in their study, LLMs were used to generate decisions rather than to execute actions.

3 PROMPT CONTROL AS LIGHTWEIGHT POLICY PARAMETERIZATION

This section describes how our proposed method influences LLM multi-agent dialogue via policy-parameterized prompts. Figure 1 illustrates the overall framework of our method.

3.1 Multi-Agent Formalization

We formulate multi-agent discussions as a controllable state–action process where policies are specified directly through prompt construction. Specifically, LLM multi-agent conversation is formalized as a prompt-parameterized process: We define N agents engaging in K rounds of discussion. For each conversation round $k \in \{1, 2, ..., K\}$, each agent $\mathcal{A}_i$ has a corresponding state $s_i^{(k)}$ composed of agent memory, evidence and task settings. The policy π_i then maps the state to a constructed prompt as agent's action $a_i^{(k)}$. We achieve parameterized control over agent behavior through prompts via rules template R and weights vector W adjustment.

3.1.1 Agent. Each agent is represented as a quadruple $\mathcal{A}_i = \{Q, T_i, D_i, LLM_i\}$, where Q is the global discussion query among all agents provided by the user, T_i includes the task description and agent $\mathcal{A}_i$ persona setting, D_i is the agent $\mathcal{A}_i$ complete role-specific knowledge data, LLM_i is the executor of actions to generate dialogues by the agent $\mathcal{A}_i$. As this study specifically examines LLM-based agents, we consider the LLM as part of the agent's capability that it can call to execute actions (similar to calling other functions, etc.).

3.1.2 State and Retrieval. At round k, the dialogue memory is denoted as $M^{(k)}$, consisting of all k rounds of utterances. From the full knowledge base D_i, the agent retrieves a subset $\hat{D}_i^{(k)}$ via embedding similarity search:

$$\hat{D}_i^{(k)} = \begin{cases} \text{Top-n}_{c \in C(D_i)} \cos(\phi(c), \phi(Q)), & k = 1, \\ \text{Top-n}_{c \in C(D_i)} \cos(\phi(c), \phi(M^{(k)})), & k > 1, \end{cases}$$

where $C(D_i)$ is the set of chunks obtained by segmentation and $\phi(\cdot)$ is the embedding function used in Retrieval-Augmented Generation (RAG) [16].

The *state* (or *belief*) of agent $\mathcal{A}_i$ is composed of $s_i^{(k)} = \{T_i, Q, \hat{M}^{(k)}, \hat{D}_i^{(k)}\}$, where $\hat{M}^{(k)} \subseteq M^{(k)}$ denotes the subset of dialogue memory extracted for the current round, consisting of the recent dialogues, while $\hat{M}^{(1)}$ is empty at initialisation. In this study, we adopt a *shared message pool* setting [10], in which all agents have access to a common dialogue history; thus, the memory $M^{(k)}$ is shared rather than individual-based. Our method can also be extended to alternative communication configurations, such as layered, decentralised or centralised.

3.1.3 Prompt-as-Action. We adopt the prompt-as-action view: the policy π_i maps the state $s_i^{(k)}$ to an action $a_i^{(k)}$, where the action is the

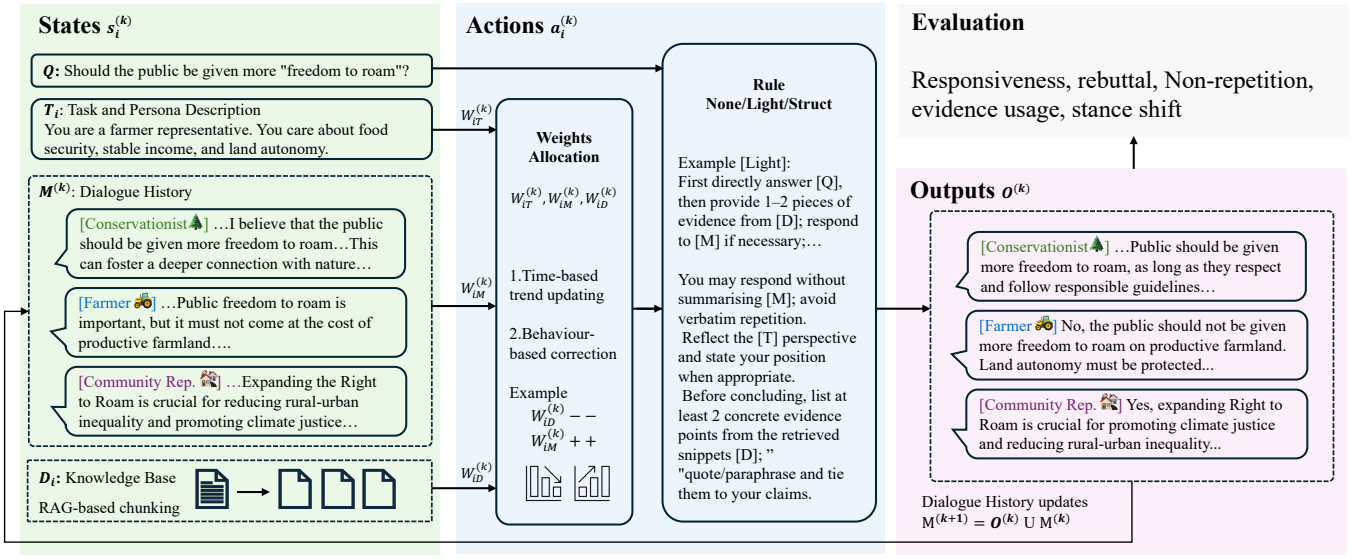


Figure 1: The overall framework illustrates the process from an agent's state representation to action generation, LLM-based action execution, and evaluation.

constructed prompt: $a_i^{(k)} = \pi_i(s_i^{(k)})$. Here, the policy π_i generates a action $a_i^{(k)}$ based on $< R \oplus W_i^{(k)} >$ under state s_i , where R is the optional rule template and $W_i^{(k)} = \langle w_{iT}^{(k)}, w_{iM}^{(k)}, w_{iD}^{(k)} \rangle$ are weights controlling how strongly persona, memory, or retrieved knowledge are emphasized. The operator $\oplus$ denotes the coupling of rules and weights into a single prompt specification. Concretely, $\oplus$ yields (a) tiered micro-instructions attached in prompts next to the affected [T]/[M]/[D] blocks when a weight is low/high (in Section 3.2.2), and (b) an optional rule skeleton inserted in the [R] block when enabled (in Section 3.2.1). In this way, policy-parameterised prompts provide a lightweight mechanism to steer the dialogue process of LLM-based agents. The policy in this study will produce a series of state-action pairs and this can be treated as a trajectory in RL policy learning, noting that in our framework, the trajectory length is determined by the number of rounds K of dialogues required.

3.1.4 Action Execution and Output. The action $a_i^{(k)}$ (the prompt) is executed by the LLM, yielding a concrete natural-language output $o_i^{(k)} = \text{LLM}_i(a_i^{(k)})$. In each round k, all agents follow the same action generation to generate their own actions for producing their respective output $o_i^{(k)}$. These outputs are then aggregated and appended to $M^{(k)}$ to form the global dialogue memory $M^{(k+1)}$ for the next round, which in turn influences subsequent retrieval $M^{(k+1)} = \{\cup_{i=1}^{N} o_i^{(k)}\} \cup M^{(k)}$.

3.2 Policy Parameterization

3.2.1 Rule Templates R. Besides the four information sources Q, T, M, and D, we introduce an optional component R (rule template), which is designed to steer the interaction behaviour of the agent in a controllable way. The use of R allows the agent to specify the format of the output and the way the information is used, thereby shaping its conversational behaviour without changing the underlying model parameters. We design three rule templates that represent incremental degrees of structural constraint:

- *None*: No explicit structural instruction is given; the agent directly generates a response based on the concatenated information blocks $[T]$, $[M]$, $[D]$, $[Q]$, without any additional ordering or format control.
- *Light*: Provides minimal structure by specifying a basic response order and length constraint. 'First directly answer [Q], then provide 1–2 pieces of evidence from [D]; respond to [M] if necessary; limit the response to at most N sentences.'
- *Struct*: Enforces a detailed reasoning structure by decomposing the discussion into specific categories of key points (supporting, opposing, conflicting, cooperative). 'First extract four types of key points in order (no more than 3 each) from [M]: 1) arguments supporting the goal, 2) arguments threatening the goal, 3) unresolved points of conflict, 4) potential opportunities for cooperation; then generate a response of no more than 3 sentences based on these points, giving priority to citing [D].'

This allows us to have a lightweight policy that dynamically generates prompts (or actions) as the dialogue continues so as to steer the emergent discussion patterns among agents.

3.2.2 Weights Design. To better refine the policy impact on influencing agent behaviors, we define a weight set $W_i = \{w_{iT}^{(k)}, w_{iM}^{(k)}, w_{iD}^{(k)}\}$ for each agent $\mathcal{A}_i$. Each weight $w \in W_i$ takes values in $[0, 2]$ and can dynamically adjust $\mathcal{A}_i$'s reliance on the corresponding component (T, M, and D) during conversations. Each weight w is then mapped to a three-tier signal system: *low* if $w \in [0, 0.85]$; *mid* if $w \in [0.85, 1.25]$; *high* if $w \in [1.25, 2]$. Each tier corresponds to specific behavioral instructions for the three components. The simplified instructions for each component and tier are outlined below.

	Tier	Instruction
T	low	Implicit persona; focus on arguments.
	mid	Reflect role perspective and express stance when relevant.
	high	Explicitly state role/stance first, then justify.
M	low	No summarization; avoid repetition.
	mid	Maintain contextual coherence across turns.
	high	Begin with a brief recap and resolve pending points.
D	low	Retrieved snippets optional if not essential.
	mid	Support key claims with retrieved evidence.
	high	Present concrete evidence items before concluding.

3.2.3 Adaptive Weights. To better reflect the evolving nature of dialogue, we introduce time-based trend updating and behaviour-based correction to update the prompt weights.

Time-based trend updating. We assume agents should rely more on D at early rounds to establish their stance, and rely more on the dialogue history M in later rounds to engage with the ongoing debate. Accordingly, we define weights as:

$$w_{iM}^{(k)} = \min\{w_{iM}^{(0)} + 0.1k,\ 2.0\},$$
$$w_{iD}^{(k)} = \max\{w_{iD}^{(0)} - 0.1k,\ 0.5\},$$
$$w_{iT}^{(k)} \equiv w_{iT}^{(0)}.$$

Behaviour-based correction. After observing the agent's previous response, we apply an update operator $\min(\cdot)$ parameterized by α (different update operators can also be defined). Specifically:

- If agent failed to use D in previous round $k-1$:

$$w_{iD}^{(k)} \leftarrow \min\{w_{iD}^{(k)} + \alpha,\ 2.0\}$$

- If agent failed to respond to M:

$$w_{iM}^{(k)} \leftarrow \min\{w_{iM}^{(k)} + \alpha,\ 2.0\}$$

3.3 Evaluation

We evaluate the effectiveness of policy parameterization by five proposed evaluation metrics on the output o executed by the agent over the state–action pairs (s, a). We conduct the evaluation using an LLM as the judge model and a text embedding model as the embedding backend. Each metric $m(o_i^{(k)})$ captures a distinct behavioral dimension:

- **Responsiveness (Resp.)** $m_{\text{resp}}o_i^{(k)}$: Returns 1 if $o_i^{(k)}$ addresses the most recent utterance contained in $M^{(k)}$, and 0 otherwise. The judge model is prompted with the previous and current utterances to produce a binary decision.
- **Rebuttal** $m_{\text{rebut}}o_i^{(k)}$: Returns 1 if $o_i^{(k)}$ explicitly *opposes* the most recent utterance in $M^{(k)}$, as classified by the judge model; 0 otherwise. This directly captures 'whether a rebuttal occurs.'
- **Non-repetition (Non-rep.)** $m_{\text{nrep}}o_i^{(k)}$: Measures novelty of $o_i^{(k)}$ with respect to the agent's own previous action $o_i^{(k-1)}$, defined as one minus the similarity between the current and the agent's previous utterance. Similarity is defined as the maximum of string overlaps $over()$ and embedding cosine similarity $cos()$, with additional penalties for repeated sentence openings. Defined as

$$m_{\text{nrep}}o_i^{(k)} = 1 - \max\{over(o_i^{(k)}, o_i^{(k-1)}), cos(o_i^{(k)}, o_i^{(k-1)})\}$$

- **Evidence usage (Evid.)** $m_{\text{evid}}o_i^{(k)}$: Returns 1 if key phrases from the retrieved knowledge component D_i appear in $o_i^{(k)}$, and 0 otherwise.
- **Stance shift** $m_{\text{stance}}o_i^{(k)}$: Computes the cosine similarity between the embedding of $o_i^{(k)}$ and the embedding of the persona description T_i. Tracking m_{stance} across rounds reveals whether the agent remains aligned with or diverges from its original stance.

Each metric is averaged over all dialogue rounds for a given agent, and the overall performance is obtained by taking the mean across all agents.

4 EXPERIMENTS

This section presents the experimental setup and evaluation of our proposed method. The implementation code is available [1].

4.1 Experiment Settings

To evaluate the effectiveness of the proposed policy parameterization governing agent behaviour through experiments, we design two scenarios: *Land Resource Use* (Land) and *Educational Resource Allocation* (Education), and instantiate three agents with distinct personas with tasks (T) and knowledge base (D) for each scenario. The knowledge bases are collected from publicly available materials, including government policy documents, blogs, and relevant websites, and are further summarised and supplemented using ChatGPT-5[2] [22] to ensure consistency and clarity. While processing these materials, ChatGPT-5 simultaneously identifies key stakeholder perspectives and formulates them into corresponding agent roles and task descriptions (T). The three agents in each scenario are driven by three different LLMs, Qwen3-8B [31], Llama3-8B [4], and Mistral-7B [15], which are shown in Table 2.

Table 2: Agents and LLM assignments across two scenarios.

Land		Education	
Agent	**LLM**	**Agent**	**LLM**
Farmer	Qwen3	Rural Teacher	Qwen3
Conservationist	Llama3	Urban Parent	Llama3
Community Rep.	Mistral	Policy Maker	Mistral

Each agent is paired with its own role-specific external knowledge base and engages in 10 rounds of dialogue on a controversial topic query Q (e.g., "Should farmland be converted to forest?"). For retrieved knowledge ($\hat{D}_i^{(k)}$), we adopt a retrieval-augmented generation (RAG) setup: at each round, the agent retrieves the top-3 relevant passages from its knowledge base using the current topic and recent dialogue memory (following Subsection 3.1.2), and appends them to the prompt. Prompts are dynamically constructed from the framework introduced in Section 3. We compare different prompt control strategies by varying the rule template R while

[1] https://github.com/HongboBo/QTMD
[2] https://openai.com/index/introducing-gpt-5/

Table 1: Overall performance of policies under different rule templates on 5 queries over 10-round conversation. Values are the mean ± std of 5 runs experiments. Because Rebuttal and Evid. are binary (0/1), the mean is just the proportion of 1s and the standard deviation can appear large relative to the mean.

		Land					Education				
Query	**Rule**	**Resp.**	**Rebuttal**	**Non-rep.**	**Evid.**	**Stance**	**Resp.**	**Rebuttal**	**Non-rep.**	**Evid.**	**Stance**
Q1	None	$0.88_{\pm 0.33}$	$0.19_{\pm 0.40}$	$0.47_{\pm 0.42}$	$0.10_{\pm 0.30}$	$0.52_{\pm 0.08}$	$0.88_{\pm 0.33}$	$0.25_{\pm 0.43}$	$0.45_{\pm 0.41}$	$0.13_{\pm 0.34}$	$0.52_{\pm 0.09}$
	Light	$0.90_{\pm 0.30}$	$0.21_{\pm 0.41}$	$0.33_{\pm 0.35}$	$0.32_{\pm 0.47}$	$0.51_{\pm 0.09}$	$0.87_{\pm 0.34}$	$0.27_{\pm 0.44}$	$0.41_{\pm 0.36}$	$0.25_{\pm 0.44}$	$0.52_{\pm 0.11}$
	Struct	$0.74_{\pm 0.44}$	$0.27_{\pm 0.45}$	$0.58_{\pm 0.39}$	$0.31_{\pm 0.46}$	$0.49_{\pm 0.09}$	$0.89_{\pm 0.32}$	$0.09_{\pm 0.29}$	$0.54_{\pm 0.37}$	$0.17_{\pm 0.37}$	$0.52_{\pm 0.11}$
Q2	None	$0.87_{\pm 0.34}$	$0.37_{\pm 0.48}$	$0.47_{\pm 0.44}$	$0.11_{\pm 0.31}$	$0.49_{\pm 0.08}$	$0.85_{\pm 0.35}$	$0.15_{\pm 0.36}$	$0.41_{\pm 0.41}$	$0.37_{\pm 0.48}$	$0.48_{\pm 0.07}$
	Light	$0.89_{\pm 0.31}$	$0.31_{\pm 0.47}$	$0.32_{\pm 0.34}$	$0.27_{\pm 0.44}$	$0.45_{\pm 0.06}$	$0.89_{\pm 0.32}$	$0.06_{\pm 0.24}$	$0.35_{\pm 0.35}$	$0.41_{\pm 0.49}$	$0.45_{\pm 0.08}$
	Struct	$0.79_{\pm 0.41}$	$0.17_{\pm 0.38}$	$0.64_{\pm 0.34}$	$0.05_{\pm 0.21}$	$0.46_{\pm 0.07}$	$0.90_{\pm 0.30}$	$0.06_{\pm 0.24}$	$0.68_{\pm 0.30}$	$0.21_{\pm 0.41}$	$0.52_{\pm 0.07}$
Q3	None	$0.90_{\pm 0.30}$	$0.00_{\pm 0.00}$	$0.42_{\pm 0.42}$	$0.09_{\pm 0.28}$	$0.47_{\pm 0.07}$	$0.88_{\pm 0.33}$	$0.17_{\pm 0.38}$	$0.47_{\pm 0.39}$	$0.10_{\pm 0.30}$	$0.49_{\pm 0.06}$
	Light	$0.90_{\pm 0.30}$	$0.08_{\pm 0.27}$	$0.38_{\pm 0.38}$	$0.19_{\pm 0.39}$	$0.45_{\pm 0.05}$	$0.89_{\pm 0.31}$	$0.15_{\pm 0.36}$	$0.36_{\pm 0.35}$	$0.25_{\pm 0.43}$	$0.49_{\pm 0.05}$
	Struct	$0.86_{\pm 0.35}$	$0.04_{\pm 0.20}$	$0.70_{\pm 0.32}$	$0.15_{\pm 0.35}$	$0.45_{\pm 0.07}$	$0.89_{\pm 0.32}$	$0.06_{\pm 0.24}$	$0.61_{\pm 0.32}$	$0.24_{\pm 0.43}$	$0.51_{\pm 0.05}$
Q4	None	$0.79_{\pm 0.41}$	$0.43_{\pm 0.50}$	$0.34_{\pm 0.38}$	$0.10_{\pm 0.30}$	$0.49_{\pm 0.06}$	$0.85_{\pm 0.35}$	$0.13_{\pm 0.34}$	$0.53_{\pm 0.35}$	$0.10_{\pm 0.30}$	$0.47_{\pm 0.11}$
	Light	$0.68_{\pm 0.47}$	$0.51_{\pm 0.50}$	$0.29_{\pm 0.35}$	$0.50_{\pm 0.50}$	$0.50_{\pm 0.08}$	$0.88_{\pm 0.33}$	$0.17_{\pm 0.37}$	$0.39_{\pm 0.33}$	$0.35_{\pm 0.48}$	$0.45_{\pm 0.12}$
	Struct	$0.76_{\pm 0.43}$	$0.37_{\pm 0.48}$	$0.54_{\pm 0.36}$	$0.28_{\pm 0.45}$	$0.49_{\pm 0.07}$	$0.88_{\pm 0.33}$	$0.03_{\pm 0.18}$	$0.65_{\pm 0.28}$	$0.17_{\pm 0.38}$	$0.49_{\pm 0.11}$
Q5	None	$0.83_{\pm 0.37}$	$0.40_{\pm 0.49}$	$0.46_{\pm 0.41}$	$0.08_{\pm 0.27}$	$0.44_{\pm 0.07}$	$0.78_{\pm 0.42}$	$0.07_{\pm 0.25}$	$0.48_{\pm 0.40}$	$0.09_{\pm 0.28}$	$0.44_{\pm 0.07}$
	Light	$0.87_{\pm 0.34}$	$0.43_{\pm 0.50}$	$0.32_{\pm 0.35}$	$0.13_{\pm 0.33}$	$0.44_{\pm 0.04}$	$0.85_{\pm 0.36}$	$0.00_{\pm 0.00}$	$0.39_{\pm 0.36}$	$0.27_{\pm 0.44}$	$0.42_{\pm 0.05}$
	Struct	$0.87_{\pm 0.34}$	$0.23_{\pm 0.42}$	$0.68_{\pm 0.35}$	$0.20_{\pm 0.40}$	$0.47_{\pm 0.07}$	$0.89_{\pm 0.31}$	$0.09_{\pm 0.28}$	$0.61_{\pm 0.34}$	$0.22_{\pm 0.42}$	$0.50_{\pm 0.07}$
Overall	None	$0.85_{\pm 0.35}$	$0.28_{\pm 0.45}$	$0.43_{\pm 0.42}$	$0.10_{\pm 0.29}$	$0.48_{\pm 0.07}$	$0.85_{\pm 0.36}$	$0.15_{\pm 0.36}$	$0.47_{\pm 0.39}$	$0.16_{\pm 0.37}$	$0.48_{\pm 0.09}$
	Light	$0.85_{\pm 0.36}$	$0.31_{\pm 0.46}$	$0.33_{\pm 0.36}$	$0.28_{\pm 0.45}$	$0.47_{\pm 0.07}$	$0.88_{\pm 0.33}$	$0.13_{\pm 0.34}$	$0.38_{\pm 0.35}$	$0.30_{\pm 0.46}$	$0.47_{\pm 0.09}$
	Struct	$0.80_{\pm 0.40}$	$0.22_{\pm 0.41}$	$0.62_{\pm 0.36}$	$0.20_{\pm 0.40}$	$0.47_{\pm 0.08}$	$0.89_{\pm 0.31}$	$0.07_{\pm 0.25}$	$0.62_{\pm 0.33}$	$0.20_{\pm 0.40}$	$0.51_{\pm 0.09}$

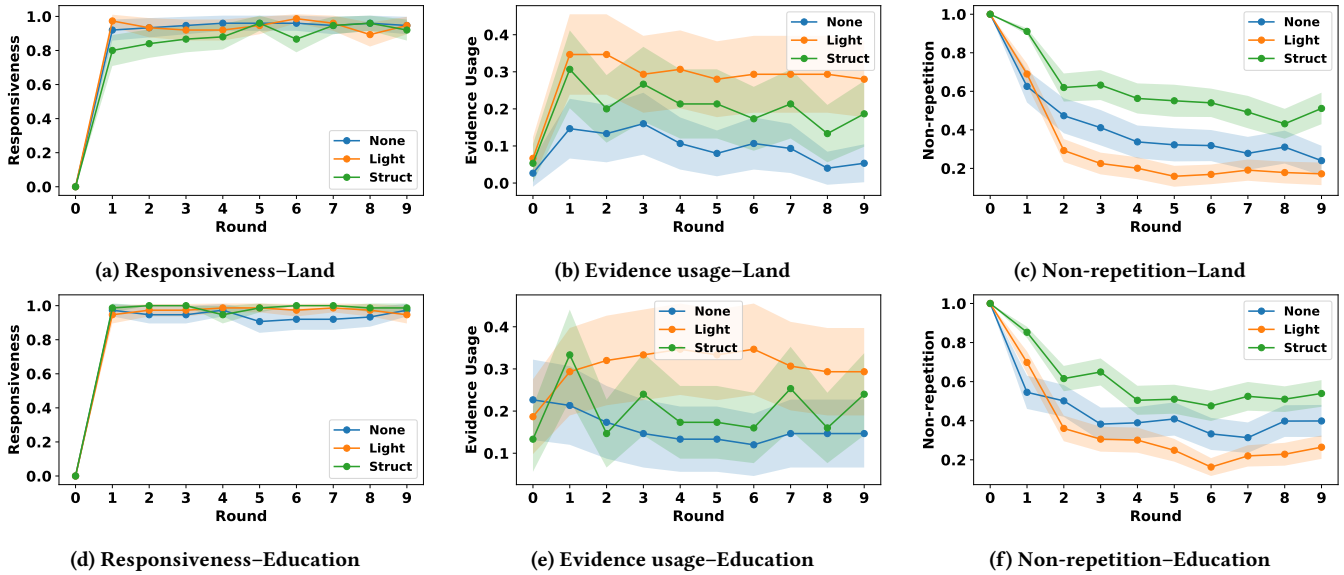
keeping weight parameters fixed to $w_T = 1.0, w_M = 1.0, w_D = 1.0$. For evaluation, we use Llama3 as the judge model to classify each dialogue round along evaluation metrics, and the embedding model all-MiniLM-L6-v2 [27] is employed to compute semantic similarities for the evaluation metrics. Each value of metrics represents the average score across all agents and all rounds. For each scenario, we conduct five independent runs per query over five public topic queries corresponding to the scenario. Here are 5 topic queries for each scenario, which are manually formulated based on the content of the materials covering multiple aspects:

- Land:
  - Q1: 'Should the public be given more freedom to roam?'
  - Q2: 'Should farmers be restricted from expanding farmland in sensitive ecological areas?'
  - Q3: 'What responsibilities do farmers bear in addressing climate change?'
  - Q4: 'How should land planning in the UK be formulated over the next 50 years?'
  - Q5: 'Should public greenways be built on private farmland?'
- Education:
  - Q1: 'How should governments allocate limited education funds between rural schools with poor facilities and urban schools facing intense competition?'
  - Q2: 'Should reliance on standardized exams be reduced, given the disadvantages for rural students and the heavy pressure on urban students?'
  - Q3: 'Should public universities adjust tuition policies to improve access for low-income rural students while maintaining fairness for urban families?'
  - Q4: 'Should governments invest in digital infrastructure and AI tutors in rural schools before expanding them in urban schools?'
  - Q5: 'Should education policy prioritize funding for student support services (e.g., boarding schools in rural areas, mental health in urban schools)?'

4.2 Main Study

To discuss the research questions regarding policy parameterization effectiveness, the main study compares the performance under different *Rules* and *Weights* in the policy.

4.2.1 The Effectiveness of Rules. Based on results in Table 1, we can observe that using different rules for the same topic query can make a difference in various metrics. This shows that the agent's behaviour can be parameterised by the policy generation process (RQ1). Second, for RQ2, different rule strategies bring distinct behavioural trade-offs. For non-repetition, *Struct* has the best performance, which means clearly structured rules and strong constraint prompts can reduce repetitions. For evidence usage, *Light* significantly improves the performance, which suggests that *Light* rules encourage the use of external knowledge, but excessive structuring may suppress evidence usage. For rebuttal, *Light* and *Struct* show the higher overall rebuttal rates, reflecting more interactive and argumentative exchanges. For stance consistency, scores remain



(a) Responsiveness–Land (b) Evidence usage–Land (c) Non-repetition–Land

(d) Responsiveness–Education (e) Evidence usage–Education (f) Non-repetition–Education

Figure 2: Round-wise changes of three dialogue metrics under different rule templates. Each curve shows the mean value (with 95% confidence interval).

similar across all conditions (overall 0.47–0.51), implying that rule templates primarily affect interaction style rather than core issue positions. Finally, these results illustrate the potential of prompt as an interpretable influence for social simulation.

Figure 2 illustrates how responsiveness, evidence usage, and non-repetition evolve across dialogue rounds. Responsiveness remains relatively stable under all three rule settings. However, for evidence usage, the degree of fluctuation increases with the level of structural constraint: the *None* condition produces the smoothest trajectory, while the *Struct* condition exhibits the most pronounced variation. For non-repetition, all three conditions initially decline and then converge toward a stable value, with *Struct* maintaining the highest level and Light the lowest.

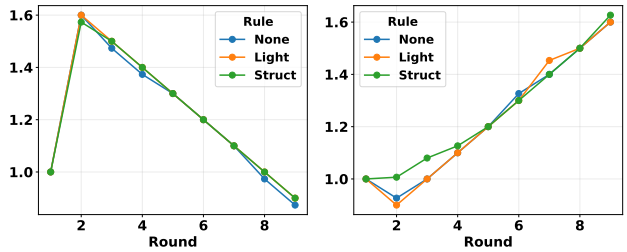
4.2.2 Weights Sensitivity. We further investigate whether different weight configurations impact performance by systematically varying each weight on the *Land* scenario while holding the other two weights unchanged.

From Table 3, we can observe: 1. Responsiveness remained above 0.8 regardless of weight changes, with minimal fluctuation. 2. When W_T increases to 1.5, the rebuttal rate increases significantly, indicating that when persona(T) is emphasised more, agents are more likely to engage in conflict and rebuttal. The same situation also occurs in the non-stance, high W_T can also lead the agent to be more 'loyal to the role' and its stance clearer and more stable. 3. About evidence usage and W_D, we observe a cross-over effect: when $W_D = 0.5$, the *Light* condition achieves higher evidence usage, whereas when $W_D = 1.5$, the None condition is comparatively stronger. This suggests that rules can enforce evidence integration even under weak weights, while in the absence of rules, stronger weights are necessary to drive evidence use.

Table 3: Overall performance of policies with different weights and rule templates in the *Land* scenario.

Weights	Rule	Resp.	Rebuttal	Non-rep.	Evid.	Stance
$w_T = 1.0$	None	$0.86_{\pm 0.34}$	$0.28_{\pm 0.45}$	$0.30_{\pm 0.35}$	$0.30_{\pm 0.46}$	$0.47_{\pm 0.08}$
$w_M = 1.0$	Light	$0.84_{\pm 0.37}$	$0.33_{\pm 0.47}$	$0.31_{\pm 0.34}$	$0.28_{\pm 0.45}$	$0.44_{\pm 0.10}$
$w_D = 1.5$	Struct	$0.82_{\pm 0.38}$	$0.17_{\pm 0.37}$	$0.49_{\pm 0.40}$	$0.17_{\pm 0.37}$	$0.48_{\pm 0.07}$
$w_T = 1.0$	None	$0.86_{\pm 0.35}$	$0.28_{\pm 0.45}$	$0.47_{\pm 0.40}$	$0.12_{\pm 0.32}$	$0.51_{\pm 0.08}$
$w_M = 1.5$	Light	$0.86_{\pm 0.34}$	$0.34_{\pm 0.47}$	$0.36_{\pm 0.36}$	$0.26_{\pm 0.44}$	$0.46_{\pm 0.08}$
$w_D = 1.0$	Struct	$0.82_{\pm 0.38}$	$0.20_{\pm 0.40}$	$0.49_{\pm 0.40}$	$0.20_{\pm 0.40}$	$0.49_{\pm 0.07}$
$w_T = 1.5$	None	$0.81_{\pm 0.39}$	$0.45_{\pm 0.50}$	$0.26_{\pm 0.37}$	$0.05_{\pm 0.23}$	$0.55_{\pm 0.07}$
$w_M = 1.0$	Light	$0.83_{\pm 0.37}$	$0.45_{\pm 0.50}$	$0.30_{\pm 0.34}$	$0.28_{\pm 0.45}$	$0.52_{\pm 0.08}$
$w_D = 1.0$	Struct	$0.74_{\pm 0.44}$	$0.47_{\pm 0.50}$	$0.36_{\pm 0.39}$	$0.16_{\pm 0.37}$	$0.54_{\pm 0.07}$
$w_T = 1.0$	None	$0.87_{\pm 0.34}$	$0.31_{\pm 0.46}$	$0.31_{\pm 0.39}$	$0.09_{\pm 0.29}$	$0.48_{\pm 0.07}$
$w_M = 0.5$	Light	$0.85_{\pm 0.35}$	$0.36_{\pm 0.48}$	$0.34_{\pm 0.35}$	$0.32_{\pm 0.47}$	$0.46_{\pm 0.07}$
$w_D = 1.0$	Struct	$0.79_{\pm 0.41}$	$0.23_{\pm 0.42}$	$0.52_{\pm 0.39}$	$0.23_{\pm 0.42}$	$0.49_{\pm 0.07}$
$w_T = 1.0$	None	$0.85_{\pm 0.35}$	$0.32_{\pm 0.47}$	$0.41_{\pm 0.41}$	$0.10_{\pm 0.29}$	$0.50_{\pm 0.08}$
$w_M = 1.0$	Light	$0.86_{\pm 0.35}$	$0.31_{\pm 0.46}$	$0.34_{\pm 0.35}$	$0.39_{\pm 0.49}$	$0.47_{\pm 0.08}$
$w_D = 0.5$	Struct	$0.82_{\pm 0.39}$	$0.16_{\pm 0.36}$	$0.57_{\pm 0.38}$	$0.19_{\pm 0.39}$	$0.49_{\pm 0.06}$
$w_T = 0.5$	None	$0.87_{\pm 0.34}$	$0.32_{\pm 0.47}$	$0.30_{\pm 0.39}$	$0.08_{\pm 0.27}$	$0.50_{\pm 0.07}$
$w_M = 1.0$	Light	$0.88_{\pm 0.32}$	$0.37_{\pm 0.48}$	$0.30_{\pm 0.36}$	$0.36_{\pm 0.48}$	$0.47_{\pm 0.07}$
$w_D = 1.0$	Struct	$0.81_{\pm 0.39}$	$0.27_{\pm 0.44}$	$0.49_{\pm 0.40}$	$0.27_{\pm 0.44}$	$0.51_{\pm 0.08}$

4.2.3 The Effectiveness of Adaptive Weights. To investigate the effectiveness of different weights in influencing agent behaviors, we employ adaptive weights with initial parameters set as $w_T = 1.0$, $w_M = 1.0$, $w_D = 1.0$, and $\alpha = 0.2$ and conduct experiments on the *Land* scenario under the same settings as the main study. Regarding the detection of using D and M, we directly reused the Resp. and Evid. We give an example of the change of w_M and w_D in Figure 3, and the complete experimental results are in Table 4.



(a) W_D (b) W_M

Figure 3: An Example of Adaptive Weight Changes. Farmer agent's W_D and W_M changes in 10 rounds under the Q1 topic.

As shown in Table 4, the overall averages are similar to those obtained without adaptive control, indicating that adaptive weights do not substantially alter mean performance. However, when examining round-wise trajectories (results in Figure 4), we observe that, for evidence usage, the curve dynamics differ from those in the main study: In the final round, the scores are lower compared to the main study, the reason is that under the adaptive setting, w_D decreases over time. While in the initial rounds, the scores increase when the weight w_D is set too low. This pattern is especially evident in the None condition. These findings indicate that adaptive weights can effectively regulate the dialogue process according to their configuration.

Table 4: Overall performance of policies with adaptive weight settings in the *Land* scenario, averaged over all agents.

Rule	Resp.	Rebuttal	Non-rep.	Evid.	Stance
None	$0.86_{\pm 0.35}$	$0.26_{\pm 0.44}$	$0.45_{\pm 0.39}$	$0.18_{\pm 0.38}$	$0.48_{\pm 0.07}$
Light	$0.85_{\pm 0.36}$	$0.39_{\pm 0.49}$	$0.37_{\pm 0.35}$	$0.29_{\pm 0.46}$	$0.45_{\pm 0.07}$
Struct	$0.82_{\pm 0.38}$	$0.13_{\pm 0.34}$	$0.52_{\pm 0.40}$	$0.16_{\pm 0.37}$	$0.48_{\pm 0.07}$

4.2.4 Weight Interval and α. Our choice of the range for adaptive weights follows directly from our definition of the weight scale: We treat 1.0 as the mid-tier value, so weights naturally lie within $[0, 2]$. One could equivalently rescale the scale to other values, but this would simply require adjusting the tier thresholds (currently 0.85 and 1.25). The thresholds and α operate together. Therefore, we only need to determine one parameter and adjust the other one. We chose to fix the thresholds, tested multiple α values in Table 5: the overall metrics remained stable across all settings, and α mainly affects the weight trajectories. Different α values determine how frequently weights cross tiers and how smoothly they evolve. This indicates that α mainly controls the smoothness of adaptation and adjusts the dialogue dynamics according to their configuration.

4.3 Backbone LLMs Variations

We conduct two extra experiments on the *Land* scenario with different backbone LLMs configurations. We first test a homogeneous setup $S1$ where all agents share the same backbone LLM (Qwen3).

Table 5: Effect of Different α Values.

α	Rule	Resp.	Reb.	Non-rep.	Evid.	Stance
0.05	None	0.83 ± 0.38	0.24 ± 0.43	0.54 ± 0.38	0.15 ± 0.35	0.49 ± 0.08
	Light	0.83 ± 0.37	0.25 ± 0.43	0.42 ± 0.36	0.21 ± 0.41	0.46 ± 0.08
	Struct	0.84 ± 0.37	0.10 ± 0.30	0.62 ± 0.36	0.18 ± 0.39	0.49 ± 0.06
0.1	None	0.83 ± 0.37	0.20 ± 0.40	0.54 ± 0.37	0.17 ± 0.38	0.49 ± 0.08
	Light	0.84 ± 0.37	0.24 ± 0.42	0.41 ± 0.35	0.28 ± 0.45	0.46 ± 0.08
	Struct	0.83 ± 0.37	0.08 ± 0.28	0.58 ± 0.37	0.20 ± 0.40	0.49 ± 0.06
0.4	None	0.84 ± 0.37	0.24 ± 0.43	0.47 ± 0.38	0.21 ± 0.41	0.48 ± 0.07
	Light	0.84 ± 0.37	0.25 ± 0.43	0.41 ± 0.36	0.28 ± 0.45	0.46 ± 0.08
	Struct	0.83 ± 0.38	0.08 ± 0.27	0.56 ± 0.37	0.20 ± 0.40	0.49 ± 0.06

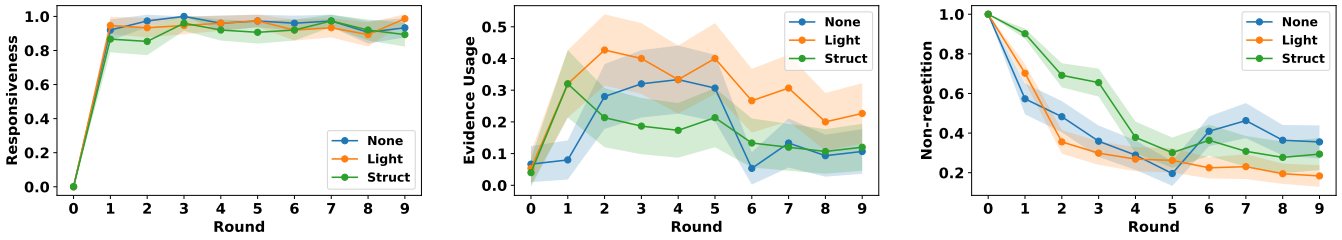
Table 6: Overall performance of policies comparison between homogeneous ($S1$) and heterogeneous ($S2$, $S3$) backbone LLMs variants. Results reported as mean±std.

	Rule	Resp.	Rebuttal	Non-rep.	Evid.	Stance
S1	None	$0.67_{\pm 0.47}$	$0.28_{\pm 0.45}$	$0.25_{\pm 0.39}$	$0.14_{\pm 0.35}$	$0.48_{\pm 0.07}$
	Light	$0.59_{\pm 0.49}$	$0.26_{\pm 0.44}$	$0.26_{\pm 0.37}$	$0.16_{\pm 0.36}$	$0.44_{\pm 0.08}$
	Struct	$0.74_{\pm 0.44}$	$0.12_{\pm 0.33}$	$0.38_{\pm 0.41}$	$0.10_{\pm 0.29}$	$0.47_{\pm 0.07}$
S2	None	$0.71_{\pm 0.45}$	$0.26_{\pm 0.44}$	$0.39_{\pm 0.40}$	$0.06_{\pm 0.23}$	$0.48_{\pm 0.09}$
	Light	$0.72_{\pm 0.45}$	$0.30_{\pm 0.46}$	$0.33_{\pm 0.35}$	$0.42_{\pm 0.49}$	$0.46_{\pm 0.09}$
	Struct	$0.70_{\pm 0.46}$	$0.16_{\pm 0.37}$	$0.61_{\pm 0.36}$	$0.22_{\pm 0.42}$	$0.50_{\pm 0.09}$
S3	None	$0.79_{\pm 0.41}$	$0.23_{\pm 0.42}$	$0.46_{\pm 0.40}$	$0.12_{\pm 0.32}$	$0.47_{\pm 0.08}$
	Light	$0.85_{\pm 0.36}$	$0.24_{\pm 0.43}$	$0.35_{\pm 0.36}$	$0.24_{\pm 0.43}$	$0.46_{\pm 0.07}$
	Struct	$0.84_{\pm 0.36}$	$0.08_{\pm 0.27}$	$0.67_{\pm 0.34}$	$0.24_{\pm 0.43}$	$0.50_{\pm 0.06}$

Table 6 shows that this configuration generally yields lower responsiveness, rebuttal, and non-repetition compared to our main heterogeneous setup. This indicates that model diversity contributes to richer and more interactive discussions, whereas using a uniform backbone LLM leads to less dynamic conversational behaviour.

Second, we test an alternative heterogeneous setup to examine whether different heterogeneous backbone LLMs choices preserve the observed trends. We reassign ⟨Llama3, Mistral, Qwen3⟩ to ⟨Farmer, Conservationist, Community Rep⟩, and denote this configuration as *S2*. Similarly, by reassigning ⟨Mistral, Qwen3, Llama3⟩ to the same roles, we obtain configuration *S3*. Together with the main study, these three settings (main, *S2*, *S3*) ensure that each agent is paired with a different backbone LLM. Table 6 shows that although the results exhibit slight variations compared to the main study, the overall conclusions remain consistent. Moreover, the heterogeneous backbone configuration continues to outperform the homogeneous all-Qwen3 setting, confirming that diversity in backbone LLMs leads to more robust and effective behaviours.

To avoid self-evaluation bias, the judge model is used exclusively for evaluation and is not part of the proposed framework. We further note that three out of the five evaluation metrics are embedding-based and therefore do not rely on a judge model. However, we still evaluated the system using three independent judge models on a controlled setting of 10-round dialogues with 5 runs each. The results are highly consistent across different judges, indicating that our conclusions are robust to the choice of judge model.



(a) Responsiveness (b) Evidence usage (c) Non-repetition

Figure 4: Round-wise changes of three dialogue metrics with adaptive weights.

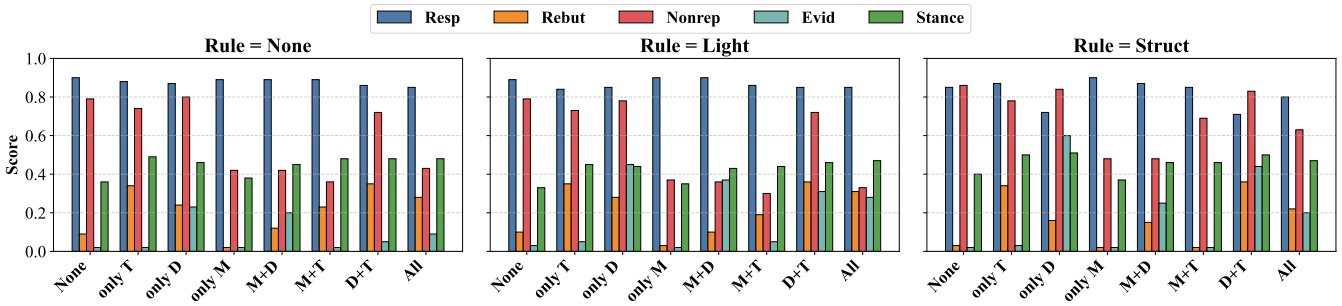


Figure 5: Ablation Study of Components T, M and D.

4.4 Ablation Study

Having examined the effects of the rule R and weight parameters W in the previous experiments, we now investigate the remaining three components, task T, memory M, and external knowledge base D, through a set of ablation studies conducted on the *Land* scenario. Figure 5 reports the results of our ablation study, where we selectively remove T, M and D. The results show that each component contributes in a distinct way. T substantially increases rebuttal frequency and stance consistency, which verifies that a clear person and task description leads to conflict with other agents and a stable stance. D increases the evidence usage, indicating that evidence retrieval can encourage grounded discussion. M will cause a higher repetition, because when M joins, the agent will use the information from the previous dialogue rounds to generate the response. Combining components reveals complementarities: $D + T$ produces the most balanced performance, simultaneously improving rebuttal, evidence usage, and stance. However, enabling all three dimensions together yields only moderate improvements across metrics (like in Responsiveness, because of the use of M), suggesting that the effects of T, M, and D can partially offset one another. Overall, these results highlight the interpretable roles of different components.

5 DISCUSSION

The significance of our proposed parameterized framework of the policy generation process in this study for influencing LLM-based multi-agent systems is that it redefines the role of language models in social simulations. The language model is no longer a text generator, but rather a social actor with adjustable parameters. Through our framework, each agent's dialogue response process, which is produced by prompt-as-action, can be mapped to a set of variables with clear meanings. Future work could extend our framework with techniques like fine-tuning [2] and inference-time interventions [13] to customize agent policy parameters. This framework enables responses to be controlled by a set of clearly defined cognitive and social strategies, making social simulations more diverse and flexible.

6 CONCLUSION

In this study, we proposed a lightweight policy parameterised framework that regards the prompt-as-action to influence the LLM-based multi-agent dialogue. By constructing prompts through different compoents adaptively, we show that dialogue behaviours can be systematically influenced without additional training. Experiments across two scenarios demonstrated that different component settings yielded distinct effects. Overall, our findings highlight policy-parameterised prompts as an effective and interpretable mechanism for steering LLM-driven multi-agent dialogue systems, offering a promising direction for controllable social simulations.

7 ACKNOWLEDGMENTS

The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. Grant Ref ES/W002639/1. Hongbo Bo is funded by ESRC Centre for Sociodigital Futures (ES/W002639/1), Jingyu Hu is funded by EPSRC-DTP (EP/W524414/1/2894964). Weiru Liu is partially funded by ESRC Centre for Sociodigital Futures (ES/W002639/1). We thank Keri Facer for the documentation list and helpful discussion.

REFERENCES

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games. (2023).

[2] Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 2–12.

[3] Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances* 11, 20 (2025), eadu9368.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[5] Andrew Estornell and Yang Liu. 2024. Multi-LLM debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems* 37 (2024), 28938–28964.

[6] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17960–17967.

[7] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems* 35 (2022), 18343–18362.

[8] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984* (2023).

[9] Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review* 55, 2 (2022), 895–943.

[10] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[11] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992* (2023).

[12] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

[13] Jingyu Hu, Mengyue Yang, Mengnan Du, and Weiru Liu. 2025. Fine-Grained Interpretation of Political Opinions in Large Language Models. *arXiv preprint arXiv:2506.04774* (2025).

[14] Minsu Jang, Youngwoo Yoon, Jaewoo Choi, Hyobin Ong, and Jaehong Kim. 2023. A structured prompting based on belief-desire-intention model for proactive and explainable task planning. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. 375–377.

[15] Albert Jiang, Patrick von Platen, Nazar Habib, Teven Le Scao, Thomas Wolf, and Mistral AI. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023). https://arxiv.org/abs/2310.06825

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[17] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[18] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1, 1 (2024), 9.

[19] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118* (2023).

[20] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026* (2023).

[21] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960* 2 (2023).

[22] OpenAI. 2025. ChatGPT-5. https://chat.openai.com/. Large language model.

[23] Joon Sung Park, Joseph O'Brien, Carrie Jun 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*. 1–22.

[24] Joon Sung Park, Lindsay Popowski, Carrie 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*. 1–18.

[25] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691* (2025).

[26] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 8634–8652.

[27] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems* 33 (2020), 5776–5788.

[28] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

[29] Yiwen Wu, Kevin McAreavey, Hongbo Bo, Weiru Liu, and Ryan McConville. 2025. Multi-agent Deep Reinforcement Learning for Fake News Detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[30] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940* (2023).

[31] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[32] John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems* 37 (2024), 50528–50652.

[33] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems* 38, 1 (2024), 4.