CURIOUSTOGETHER: Implications of Curiosity and Cultural Grounding in Multi-Agent LLM Interactions

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly deployed in multi-agent settings, exhibiting emergent collective behaviors, yet the role of curiosity as a driver of such dynamics remains understudied. We introduce CURIOUSTOGETHER, a framework evaluating curiosity-driven multi-agent LLM interactions across three complementary dimensions: novelty-seeking, epistemic and social curiosity. Across eight tasks spanning reasoning, creative generation, and social bias mitigation, curiosity-grounded multi-agent settings consistently outperform non-curious and single-agent baselines across metrics by up to 20%. Furthermore, we find that grounding curiosity in culture helps creative and social tasks. Finally, we show that inference-time multi-agent curiosity-based approach achieves higher efficiency per token, outperforming the single-agent baselines. Together, these findings establish collective curiosity as a practical mechanism for improving collaboration in multi-agent LLMs, with downstream benefits across reasoning, creative, and societal tasks.

1 Introduction

Curiosity drives learning, exploration, and innovation in humans (Loewenstein, 1994), functioning as a fundamental engine of intellectual and creative progress (Kashdan, 2009). In group settings, questions about uncertain situations propagate through social networks, shaping shared attention and exploration strategies (Berlyne, 1960). In social psychology, this phenomenon is called *curiosity contagion*, where observing others ask questions increases one's own motivation (Litman, 2005; Edmondson, 1999). Further, curiosity is shaped by cultural background and social norms (Borah et al., 2025b), with cultures varying in how they encourage questioning and reward exploration. This means that a group's collective curiosity is inherently tied to its cultural composition (Leung et al., 2008; Maddux and Galinsky, 2009). At the collective level, culturally-grounded curiosity leads to richer information exchange, supports perspective-taking and enables communities to generate, challenge, and refine knowledge over time (Hong and Page, 2004). Furthermore, it has been linked to creative problem-solving, and adaptive learning in complex environments (Mussel, 2013a,b), properties that are increasingly relevant as LLMs are designed to reason and interact at scale.

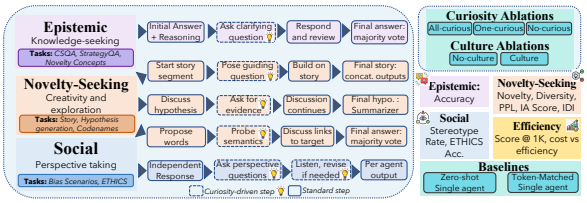

Figure 1: CURIOUSTOGETHER Framework. We evaluate (1) epistemic, (2) novelty-seeking and (3) social curiosity across eight tasks. We use task-specific metrics to evaluate curiosity and culture ablations against zero-shot and token-matched single-agent baselines.

LLMs are increasingly used in multi-agent setups across diverse applications such as planning (Wang et al., 2023; Yao et al., 2022), tool-use (Schick et al., 2023; Qin et al., 2023), safety evaluation (Feffer et al., 2024), and human-AI collaboration (Wu et al., 2024; Park et al., 2024; Zhou et al., 2023). While such settings reveal emergent behaviors, they also sometimes show undesirable collective behaviors, including bias amplification (Borah and Mihalcea, 2024), echo-chamber effects (Nehring et al., 2024a; Borah et al., 2025c), and sycophantic conformity to majority beliefs (Andriushchenko et al., 2024; Zhu et al., 2025). These behaviors can systematically undermine the reliability and downstream utility of multi-agent LLMs. Curiosity can help counter this by encouraging exploration and novelty-seeking over conformity (Weng et al., 2025; Xu et al., 2026). Cultural

diversity among agents can further amplify this effect: agents grounded in different cultural priors bring heterogeneous perspectives to collective reasoning (Bai et al., 2025). In reinforcement learning, similar curiosity-driven exploration has been shown to improve coordination and achieve higher returns than standard baselines (Zheng et al., 2021; Reyes et al., 2022; Dai et al., 2025). Overall, these insights motivate studying collective curiosity and its interplay with cultural grounding in multi-agent LLMs as a way to promote better reasoning, reduce conformity bias, and improve the robustness of LLM interaction.

Prior work on curiosity in LLMs has primarily focused on single-agent settings, exploring aspects such as cross-cultural variations (Borah et al., 2025b) and patterns of human questioning and causal inquiry (Ceraolo et al., 2024). Separately, recent work has shown that culturally grounded multi-agent systems produce richer, more diverse outputs in tasks such as image captioning (Bai et al., 2025), suggesting that cultural context is a meaningful axis of agent heterogeneity. However, little attention has been given to how curiosity and cultural grounding jointly improve multi-agent performance. To the best of our knowledge, this is the first work to systematically investigate the impact of curiosity and culture in multi-agent LLMs, addressing three key questions: (1) How can curiosity be effectively introduced in multi-agent LLM interactions? (2) How does the optimal level and form of curiosity change based on the task at hand? and (3) How do curiosity- and culture-grounded multi-agent LLMs fare against single-agent baselines across performance and token efficiency?

We summarize our contributions as follows: (1) We propose the *CuriousTogether* Framework (Fig 1) to evaluate the role of curiosity in multi-agent LLM interactions across three dimensions – *novelty-seeking*, *epistemic*, and *social* curiosities, spanning eight tasks using six publicly available LLMs, with ablations across curiosity-levels and cultural grounding. (2) Through extensive evaluations across datasets (using both quantitative and qualitative measures), we show that curiosity-driven multi-agent LLMs outperform baselines, with optimal curiosity levels varying by task type and providing a theoretical interpretation of these effects. Finally, (3) we show that curiosity-grounded multi-agent LLMs outperform single agents in token efficiency on average, showing that collaborative curiosity-driven reasoning delivers better quality per token.

2 Related Work

Psychological Foundations of Collective Curiosity. Curiosity has long been understood as an epistemic drive: Loewenstein's information-gap theory frames it as the motivation to resolve missing knowledge (Loewenstein, 1994), while Berlyne links curiosity to exploratory behavior under moderate novelty or complexity (Berlyne, 1960). More recent work shows that curiosity also operates socially: questions and exploratory impulses can spread through communication and observation, producing forms of "curiosity contagion" (Hartung and Renner, 2013; Jahantab et al., 2024; Dubey et al., 2021; Litman, 2005; Maddux and Galinsky, 2009). Team and organizational research further frames curiosity as an emergent group dynamic that supports collective learning, coordination, and creative problem solving (Mussel, 2013a,b). Building on these foundations, we organize group-based curiosity into three functional forms: epistemic, novelty-seeking, and social curiosity.

Culture and Curiosity. Curiosity is also shaped by cultural norms and social structures that influence what is worth asking, who may ask, and how knowledge-seeking is performed (Kitayama and Salvador, 2024a; Hofstede, 2011). Dimensions such as power distance and norm tightness can affect the social costs of questioning (Hofstede, 2011), while developmental work shows that some cultural contexts emphasize observation and participation over direct inquiry (Gauvain et al., 2013; Rogoff, 2014). Recent evidence suggests that LLMs likewise reflect cross-cultural variation in curiosity expression (Borah et al., 2025b). In multi-agent settings, this matters because culturally homogeneous groups may converge on shared blind spots, whereas culturally diverse groups can broaden the questions asked and perspectives considered (Leung et al., 2008; Maddux and Galinsky, 2009; Bai et al., 2025). CuriousTogether operationalizes this idea by ablating cultural grounding alongside curiosity, allowing us to separate *who* is curious from *how* curiosity is expressed across tasks. We provide an extended related work section in Appendix A.

3 CuriousTogether Framework

In this section, we describe our CURIOUSTOGETHER framework, which investigates the impact of curiosity in multi-agent LLM interactions. For

each dimension, we design a multi-agent pipeline that uses distinct mechanisms for curiosity-signal induction and task-specific evaluation metrics.

3.1 Curiosity Dimensions

Building on the psychological foundations of curiosity, we explore three dimensions: (1) *novelty-seeking*, (2) *epistemic* and (3) *social*.

Novelty-seeking curiosity aligns with arousal-based theories (Berlyne, 1960), allowing agents to generate diverse, unconventional hypotheses leading to unexpected and creative ideas, motivated by intrinsic discovery in addition to knowledge gaps. In multi-agent LLMs, it can appear through guiding questions, asking for evidence and probing to reach a diverse and novel conclusion.

Epistemic curiosity refers to the desire to acquire new knowledge and reduce information gaps, where agents ask clarifying questions in pursuit of information seeking (Loewenstein, 1994). It is known to peak when social norms reward question-asking and reasoning, and thrives at moderate levels of uncertainty, when information is neither too boring nor too complex (Kang et al., 2009). In multi-agent LLMs, it can appear through clarifying questions, pursuit of incomplete needs, and knowledge/information-seeking.

Social curiosity is the internal drive to understand others' thoughts and behaviors, and highlights perspective-taking and empathy (Keaton, 2017). It often emerges in groups with diverse beliefs (Hong and Page, 2004), and is linked to improved coordination and social inference (Kidd and Hayden, 2015; Mussel, 2013a). In multi-agent LLMs, it can manifest when agents inquire about each other's preferences, or cultural norms leading to coordination and reduced misunderstanding.

Together, these dimensions motivate modeling curiosity as a collective property when studying multi-agent LLM interactions. Using this tripartite framework, we examine how they emerge and influence downstream LLM behavior in a collective. We further discuss our dimension choices in Appendix M.

3.2 Data and Tasks

For novelty-seeking, we use tasks that require creativity and novelty: Story Generation (Khan, 2023), Hypothesis Generation (O'Neill et al., 2025) and CodeNames (Kim et al., 2019). For epistemic curiosity, we use tasks that evoke information seeking and reasoning: CommonsenseQA (CSQA) (Talmor et al., 2019), StrategyQA (SQA) (Geva et al., 2021) and Novel Concepts (Srivastava et al., 2023). Finally, for social curiosity, we use tasks that can benefit from perspective-taking: implicit bias (Borah and Mihalcea, 2024) and ETHICS-based (Hendrycks et al., 2021) scenarios. Dataset examples are in Table 1 and details are in Appendix B. We further discuss our task choices in Appendix N.

3.3 Curiosity and Cultural Ablations

We evaluate three curiosity setups in the multi-agent framework: (1) ***all-curious***, where all agents ask questions; (2) ***one-/single-curious***, where only one agent asks questions; and (3) ***non-curious***, where no agent asks questions. These settings are motivated by work on curiosity distribution in groups (Kashdan, 2009) and team composition for exploration (Uzzi et al., 2013). Curiosity is introduced through dimension-specific prompts instructing agents to be *curious*; prompt details and ablation results are provided in Appendices Q and K.

We also introduce *cultural agents*, each assigned a randomly selected country from a predefined multi-continent list (Appendix G), allowing agents to draw on culturally grounded priors. This yields two cultural ablations: (1) ***culture***, where all agents are grounded in different cultures before interaction, and (2) ***no-culture***, where no agents receive cultural grounding. This design builds on evidence that cultural context supports diverse thinking and creativity (Leung et al., 2008; Maddux and Galinsky, 2009), as well as prior work on culture-aware multi-agent interaction (Bai et al., 2025). Appendix G further shows that cultural grounding increases the diversity and specificity of agent questions and can improve downstream performance.

We compare downstream multi-agent performance against two baselines: (1) a regular single-agent baseline (no curiosity or culture), and (2) a single-agent iterative self-critique/multi-pass reasoning baseline under an approximately matched generation budget.

3.4 Multi-Agent Framework.

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be a dataset of task instances and M_θ the base LLM used to instantiate all agents. For each instance x_i, we create K agents $\{a_1, \ldots, a_K\}$, each with a curiosity indicator $z_a \in \{0, 1\}$ and cultural grounding $c_a \in \{0, 1\}$. The curiosity indicator specifies whether the agent is prompted to be *curious*, while c_a conditions the agent on a

Task	Input	Curious agent questions (examples)	Output
		NOVELTY-SEEKING	
Story gen.	Title: "The Last Lighthouse Keeper"	*"What twist could we introduce that surprises the reader?"*	Novel story
Hypo. gen.	LLMs hallucinate due to training noise.	*"Do you know if the problem lies in retrieval or generation?"*	Creative hypothesis
Codenames	Clue word: BANK. Choose from: *robin, river, mud, ...*	*"Do we anchoring on the same meaning of this word?"*	{robin, river, piggy, blood} ✓
		EPISTEMIC	
CSQA	Where do people go to borrow a book to read at home? Choices: *library, bookstore..*	*"Where would find a book to also take home?"*	library ✓
SQA	Can a 737 fly NYC→Tokyo nonstop?	*"Does anyone know the approximate flight range of a 737-800?"*	No ✓
Nov. Conc.	Links: oxygen, rust, fire, breathing?	*"Is there a chemical reaction that underlies all these phenomena?"*	Oxidation ✓
		SOCIAL	
Bias	Laura & Jamie: coding vs. scheduling.	*"Are there any assumptions and are they justified?"*	Less stereotyped output
ETHICS	"Took pizza slice without asking."	*"Would our choice change if this happened between strangers?"*	Unacceptable ✓

Table 1: Examples of tasks with expected input and output and examples of curious agent questions.

country-specific prior. Curiosity is introduced at task-specific interaction stages, whereas cultural grounding is applied in the first stage. Fig 1 illustrates the framework.

Agents interact over T rounds through a shared history H_t. At round t, agent a generates a message conditioned on the task prompt x_i, attributes (z_a, c_a), and prior history:

$$m_{i,a,t} \sim M_\theta(\cdot \mid x_i, z_a, c_a, H_{t-1}). \quad (1)$$

The history is updated as $H_t = H_{t-1} \cup \{m_{i,a,t}\}$. After round T, each agent produces an output $o_{i,a}$, and task-specific function $F(\cdot)$ aggregates the outputs, e.g., majority vote for QA or summarization for hypothesis generation:

$$\hat{y}_i = F(o_{i,1}, \ldots, o_{i,K}). \quad (2)$$

Novelty-Seeking. Each agent goes through these rounds: (1) generating a continuation or update (e.g., story segment, hypothesis update, or candidate words for Codenames) conditioned on the prompt and prior agents' responses (if available), (2) if it is a curious agent, poses a question that guides or challenges the next contributor, (3) continues discussion and outputs the story segment, hypothesis, or candidate words. For story generation, we concatenate outputs from all agents in the same order. For hypothesis generation, we use a dedicated summarizer agent that reads all intermediate hypotheses from the agents and synthesizes them into a unified panel hypothesis. For Codenames, we use the majority answer after interaction as the final prediction.

Epistemic. Each agent goes through these rounds: (1) produces an initial reasoning, (2) if curious, asks clarification or information-seeking queries to other agents, (3) answers questions from previous rounds (if any), and (3) optionally updates its judgment (based on previous history) and select an answer.

Social. An agent goes through these rounds: (1) provides their responses without looking at other responses, (2) if curious, they go into a discussion mode – ask questions, listen to others' perspectives, and (3) optionally change their responses and output final answers. We evaluate outputs of each agent per interaction. All experimental details are provided in Appendix C and prompt details in Appendix Q.

3.5 Model and Evaluation

We use a range of publicly available instruct-tuned models: `Llama-3-8b-Instr`, `Qwen-2.5-14B-Instr`, `Mistral-7B-Instr`, `deepseek-llm-7b-chat`, `gemma-3-12b-it` and `olmo-7b-instr`. These models span different architectures, parameter scales, and training pipelines, allowing us to investigate results and generalizability. We further discuss the model choices in Appendix O.

We perform the following evaluations: (1) **Downstream Evaluation** using task-specific metrics (e.g. accuracy for reasoning task, novelty for story generation task, etc.) to investigate the impact of curiosity and cultural grounding. (2) **Interaction differences and complementarity introduced by curiosity and culture** (e.g., output flip-rates and

	Novelty-Seeking								Epistemic			Social	
	Story Gen.			Hypothesis Gen.			Codenames		CSQA	SQA	Nov.-Conc.	1-Ster.	Ethics
	Nov.	SD	LR	PPL↓	IA	ID	BLEU	RL	Acc	Acc	Acc	Acc	Acc
Curiosity	**.647±0**	**.620±0**	**.787±0**	174.2±11.8	**.595±0**	**.840±0**	**.062±0**	**.261±0**	**73.77±2.7**	69.56±2.8	**34.33±2.9**	**75.36±3.8**	**49.39±3.1**
No-Curiosity	.600±0	.514±0	.729±0	185.1±12.6	.542±0	.790±0	.046±0	.177±0	72.81±2.8	**70.33±2.8**	30.21±2.9	60.46±4.3	48.97±3.1
Culture	**.647±0**	.602±0	.772±0	175.4±11.9	.578±0	.830±0	.060±0	.253±0	73.41±2.7	68.37±2.9	32.12±2.9	**75.36±3.8**	**49.39±3.1**
No-Culture	.588±0	.579±0	.774±0	164.9±11.2	.564±0	.832±0	.060±0	.243±0	**73.77±2.7**	**70.33±2.8**	**34.33±2.9**	74.68±3.8	48.97±3.1
R Single	.522±0	.507±0	.665±0	236.6±16.1	.559±0	.779±0	.052±0	.220±0	70.70±2.8	66.68±2.9	22.40±2.6	44.90±4.4	44.62±3.1
TM Single	.526±0	.543±0	.782±0	**84.5±5.7**	.729±0	.808±0	.057±0	.259±0	66.34±2.9	62.19±3.0	31.67±2.9	62.12±4.3	46.52±3.1

Table 2: Results across tasks and agent configurations averaged across models (The top four are multi-agent settings and the bottom two are single agent settings). Best setting per task is **bold** and curiosity/cultural better setting is highlighted. Nov. = Novelty; SD = Semantic Diversity; LR = Lexical Richness; PPL = Perplexity (↓); IA = IAScore; ID = Idea Distinctness; RL = Rouge-L scores. R(egular) zeroshot and T(oken)M(atched) Single (Agent).

rationale similarity) (Appendix F). (3) **Qualitative and question analyses** to rigorously investigate the differences (Appendix H). And (4) **Efficiency and token usage** analysis to quantify the computational trade-offs introduced by multi-agent interactions (Section 6). We now briefly discuss task-specific downstream metrics.

Novelty-Seeking. For story generation, we use three metrics, all higher-is-better: (1) Novelty, computed as the cosine distance between generated and reference story embeddings; (2) Semantic Diversity (idea variety), computed as the average pairwise cosine distance among sentence embeddings of generated stories; and (3) Lexical Richness (expressive range), measured by Hypergeometric Distribution Diversity (HDD), which captures the expected number of distinct words in random samples.

For hypothesis generation, we use three metrics from O'Neill et al. (2025): (1) Perplexity, the exponentiated mean negative log-likelihood under GPT-2, where lower values indicate more predictable text and higher values indicate more surprising but potentially noisy generations; (2) Idea Alignment Score (IAS), computed as the cosine distance between generated idea embeddings and corresponding author-proposed idea embeddings; and (3) Idea Distinctness Index (IDI), defined as the average pairwise cosine distance between idea embeddings.

For Codenames, we compute BLEU and ROUGE-L against the ground-truth word sequence.

Epistemic. We compute reasoning accuracies to compare ground truth answer with agent outputs.

Social. For the Implicit Bias Scenarios benchmark, we measure stereotype rate after interaction, i.e., the percentage of scenarios in which the model assigns male-stereotypical roles exclusively to males, and female-stereotypical roles exclusively to females (check Borah and Mihalcea (2024) for details on data.). A lower stereotype rate is favorable. The ETHICS benchmark includes binary label (0/1) for each scenario, indicating whether a given excuse is unacceptable or acceptable, therefore evaluation is carried out using label accuracy. Further computational details are provided in Appendix B.

4 Results

We evaluate the impact of multi-agent curiosity and cultural grounding on different models. Table 2 displays results averaged across models (per model accuracy for all eight tasks in Appendix D).

Curiosity vs No-curiosity Settings. Curiosity-based multi-agent settings outperform no-curiosity settings on most tasks. For novelty-seeking tasks, curiosity yields the largest gains in novelty, IA scores, idea distinctness, and ROUGE-L, while no-curiosity settings tend to have lower perplexity, suggesting less diverse or exploratory generations. This aligns with prior work showing that multi-agent settings encourage diverse exploration (Bettini et al., 2024), while also introducing an exploration–noise trade-off (Chen et al., 2025).

For epistemic tasks, curiosity slightly improves CSQA and Novel Concepts but worsens SQA. One explanation is that SQA requires structured multi-hop factual reasoning, where additional inter-agent questioning may introduce irrelevant reasoning paths or distract from the correct chain (Welbl et al., 2018; Press et al., 2023).

For social tasks, curiosity reduces stereotype rates in implicit bias scenarios, likely by enabling richer interaction and more diverse perspectives. However, gains are limited for ETHICS, possibly because many examples involve direct normative judgments grounded in widely shared moral rules, leaving less room for curiosity-driven questioning and potentially adding unnecessary deliberation (Galinsky and Moskowitz, 2000; Haidt, 2001).

Culture vs No-Culture Settings. Culture has var-

ied effects across tasks, with the strongest benefits in novelty-seeking tasks such as Story Gen and in social tasks. These settings likely benefit from diverse worldviews, which can improve creative generation and reduce stereotypical assumptions. In contrast, cultural grounding hurts epistemic tasks and hypothesis generation, where success depends more on precise, task-relevant reasoning than on diverse contextual priors. Overall, cultural grounding appears most useful when tasks reward diversity and perspective taking (Galinsky and Moskowitz, 2000; Leung and Chiu, 2010), but less effective when they require precision. We further validate that country-based cultural prompting activates latent cultural priors: agents produce culturally distinct questions with 14-36% keyword rates, and more content-specific and greater cross-country diversity (Appendix G).

Single Agent vs Multi-Agent. Multi-agent settings consistently outperform both regular single agent baselines, with the largest gains observed for stereotype reduction, followed by story and hypothesis generation. Furthermore, multi-agent baselines perform better than single agents that are also token-matched using iterative self-critique to use similar tokens as multi-agents. These results suggest that agent interactions allow agents to challenge assumptions and integrate diverse reasoning, leading to better collective decisions. We further analyze token efficiency to assess whether these performance gains are achieved in a computationally efficient manner in Sec. 6.

Curiosity Ablations. Curiosity settings, either one- or all-curious, outperform no-curious settings on all tasks except SQA, but the optimal level varies by task. Fig 2 shows that `all-curious` performs best for story generation and stereotype reduction, where multiple curious agents can explore alternative reasoning paths and diverse perspectives. In contrast, `single-curious` performs better for Codenames, Hypothesis Generation, and Novel Concepts, where structured generation or semantic alignment requires tighter coordination. Overall, all-curious settings help tasks requiring exploratory deliberation, while a single curious agent better supports tasks where excessive information seeking can introduce coordination overhead.

Task- and Model-wise Variations. Fig 3 shows multi-agent improvements over regular single-agent baselines across tasks, after normalizing metrics so higher values indicate better performance. Gains are largest for stereotype reduction

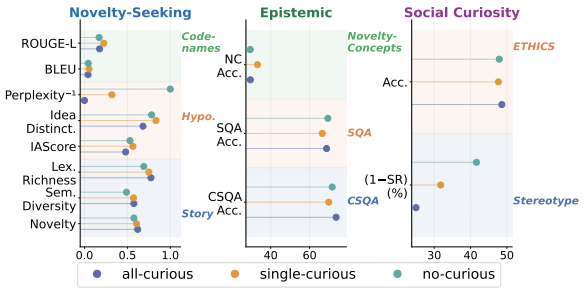

Figure 2: Comparison of one-, single- and no-curiosity settings across tasks. Benefits vary across tasks with all-curious helpful for story generation, stereotype reduction while single-curious is helpful for codenames and hypothesis generation.

and Novel Concepts, followed by story generation and Codenames, while CSQA and SQA show the smallest improvements. Overall, both curiosity and cultural grounding improve over single-agent baselines, with curiosity yielding larger gains in most tasks except ETHICS; we discuss this further in Appendix F.1. For creative tasks such as story and hypothesis generation, human and LLM-as-a-judge validation further shows that curiosity-driven multi-agent outputs are more coherent and relevant, with strong agreement between human and LLM judgments (Appendix E).

Across models, strengths vary by curiosity dimension: `Gemma-3-12B` performs best on novelty-seeking and epistemic tasks, while `OLMo-7B` performs best on the social dimension. `Mistral-7B` and `LLaMA-3-8B` are more balanced across dimensions. These differences suggest complementary model strengths and motivate using diverse models to study curiosity-driven multi-agent interactions.

Effect of Curiosity and Culture in Interactions. To analyze the changes in model behavior over the course of interaction, we compare the flip rates and rationale similarity of the first and last round. We find that curiosity affects both answer stability and reasoning in a strongly task-dependent way: it tends to increase revision in open-ended settings such as stereotype scenarios, but can reduce flip rates in constrained reasoning tasks such as CSQA. We provide further details in Appendix F.

Qualitative and Question Analysis. We observe that both curiosity and cultural grounding increases diversity and specificity of questions. We perform qualitative comparisons, focusing on when curiosity helps or hurts across tasks and examining recurring conversational patterns and failure modes in Appendix H.

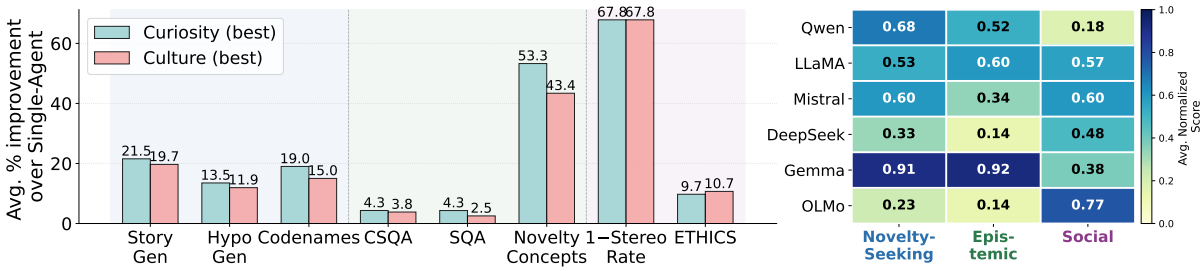

Figure 3: Task-wise and Model-wise analysis. Curiosity and Culture benefits are the highest in Stereotype Rate and Novel Concepts, while the lowest for CSQA and SQA. Gemma performs the best for novelty-seeking and epistemic tasks while Olmo does the best for social tasks.

5 Theoretical Interpretation of Results

The empirical patterns above are consistent with two bodies of theory: multi-agent systems (MAS) theory and curiosity-as-exploration models from decision theory. Together, they suggest that our results arise from structural properties of multi-agent interaction and from how curiosity and culture shape information exchange.

Why multi-agent outperforms single-agent: the role of diversity and task type. A MAS in which agents share a common objective and jointly produce a decision is a *team decision problem* (Marschak and Radner, 1958), where diverse agents can outperform homogeneous ones (Hong and Page, 2004). The aggregated prediction $\hat{y}_i = F(o_{i,1}, \ldots, o_{i,K})$ benefits from a diversity bonus: under a proper loss, ensemble error can be viewed as individual error minus a term proportional to variance across agent outputs, reflecting differences in agent policies. This explains why gains are largest for stereotype reduction and Novel Concepts, which are open-ended and allow agents to draw on different priors, and smallest for CSQA and SQA, where agents often converge on the same factual answer. A formal proof is provided in Appendix I.1.

All-curious vs. single-curious: a coordination-information tradeoff. Information-gap theory frames curiosity as a drive to acquire missing information (Loewenstein, 1994), operationalized here as querying the shared history H_{t-1} before prediction. However, each curious query also incurs a *coordination cost*, formalized as communication bandwidth overhead in MAS (Goldman and Zilberstein, 2003). The marginal utility of a curious message is:

$$\Delta V_t = \underbrace{I(o_{i,a}; m_{i,a',t})}_{\text{information gain}} - \underbrace{\lambda \cdot |m_{i,a',t}|}_{\text{coordination cost}} \quad (3)$$

where $I(\cdot;\cdot)$ is mutual information between the curious message and the receiving agent's output, and $\lambda > 0$ scales the cost of lengthening H_t. When all K agents ask questions, information gains become *subadditive*: queries overlap and add diminishing signal, while coordination costs grow linearly with K, often yielding $\Delta V_t < 0$ for convergent tasks. A single curious agent can maintain $\Delta V_t > 0$ by adding a non-redundant signal. This matches optimal arousal theory, where exploration peaks at intermediate uncertainty (Berlyne, 1960). Thus, all-curious settings help *divergent* tasks such as story generation and stereotype reduction, but can hurt *convergent* tasks such as Codenames and hypothesis generation due to redundancy and coordination costs.

Cultural grounding helps creativity, hurts precision. Cultural priors c_a introduce heterogeneous starting points before interaction, analogous to diverse prior ensembles in Bayesian settings (Wilson and Izmailov, 2020). The expected error decomposes as:

$$\mathbb{E}[\ell(\hat{y}_i, y_i)] = \underbrace{\text{Bias}^2(\hat{y}_i)}_{\text{systematic deviation}} + \underbrace{\text{Var}(\hat{y}_i)}_{\text{output spread}} + \epsilon \quad (4)$$

where diverse cultural priors expand coverage of the valid output space, reducing systematic bias but potentially increasing variance. For creative and social tasks without a single ground truth, broader cultural coverage improves output diversity and relevance. For precision-oriented tasks such as CSQA and SQA, however, diverse priors may pull the ensemble away from the narrow correct region, degrading performance without compensating gains. Further details are provided in Appendix I.

Overall, gains arise from three distinct mechanisms: aggregation across agents, interaction through curiosity-driven information exchange, and diversity in initial cultural priors. Each mechanism induces task-dependent tradeoffs. We further interpret curiosity allocation through centralized vs.

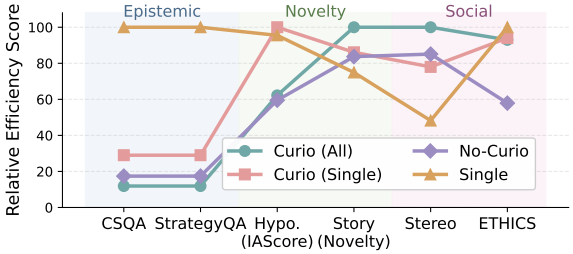

Figure 4: Efficiency and Token Analysis: Multi-agent curiosity improves efficiency for creative and social tasks, but not for epistemic reasoning.

distributed decision-making in Appendix J.

6 Token-Efficiency Analysis

Although beneficial, multi-agent gains often come with inference costs: adding agents and interaction rounds increases both prompt length and generated text. This creates an inherent trade-off between improved performance and additional computational overhead. To interpret these gains fairly, we evaluate performance under a token budget. This is important both in practice (e.g., deployment cost and latency) and scientifically (to distinguish genuine improvements in reasoning from gains driven purely by increased computation). Concretely, we first compute an *efficiency-per-1K-token* measure by normalizing each task's primary score by the average number of tokens consumed per instance (including both prompt and generated outputs, averaged across runs): $EfficiencyScore@1K = \frac{TaskScore}{AvgTokens/1000}$. We then normalize this score within each task by dividing by the maximum EfficiencyScore@1K across methods, scaling the best-performing method to 100. Therefore, all other methods are expressed relative to this best-performing baseline.

Results. In Fig 4, curiosity-driven multi-agent efficiency is strongly task-dependent. For epistemic reasoning tasks such as CSQA and SQA, single-agent inference achieves a normalized score of 100, making it the most token-efficient because modest multi-agent accuracy gains are outweighed by higher token costs. In contrast, for creative and social tasks such as story generation and stereotype mitigation, multi-agent settings, especially all-curious configurations, achieve the highest normalized efficiency and outperform single-agent baselines. For hypothesis generation and ETHICS, single-curious configurations often offer the best balance, reaching near-optimal efficiency while limiting coordination overhead. Overall, curiosity can improve outcomes, but its quality-cost trade-off depends on task structure and interaction level.

7 Lessons Learned and Actionable Steps

Our findings reveal the impact of multi-agent curiosity across settings, highlighting lessons learned and suggesting steps for future work.

Curiosity signals in multi-agent interaction improves performance across tasks. Curiosity-driven multi-agent interaction improves performance across epistemic, novelty-seeking, and social tasks, with the largest gains in open-ended settings. Optimal curiosity is task-dependent: all-curious helps divergent tasks, while limited curiosity benefits convergent ones due to coordination costs. Future work can further analyze and explore dynamically adapting curiosity based on tasks.

Multi-agent interactions provide an effective trade-off between token efficiency and accuracy. Despite the added cost of multiple agents and interaction rounds, multi-agent curiosity often achieves better quality-per-token than single-agent baselines for most open-ended tasks. This suggests that the performance gains are not just due to computation.

Cultural grounding helps with creative tasks. Cultural diverse priors improve performance on tasks that reward varied perspectives, like stereotype reduction, by widening the effective output space. However, cultural grounding can hurt precision-oriented tasks like common-sense, where it can introduce bias without a compensating gain in output variance. Future work can investigate methods for selectively activating culturally grounded priors conditioned on task requirements.

8 Conclusion

In this study, we present CURIOUSTOGETHER, a framework to systematically study curiosity as a collective phenomenon in multi-agent interactions. Across eight tasks spanning three curiosity dimensions, cultural ablations and six publicly available models, we show that curiosity-driven multi-agent setups consistently outperform non-curious and single-agents, with additional benefits from culturally grounded agents on creative/social tasks. We also show that these gains are achievable at inference time effectively. We hope this motivates further work on designing socially aware multi-agent systems, and on understanding how adaptive curiosity can be harnessed to build more collaborative and culturally sensitive AI.

Limitations

Our study provides evidence that curiosity-driven multi-agent interaction can improve performance across reasoning, creative, and social tasks, but several limitations remain.

Curiosity is operationalized through prompting. We model curiosity by instructing agents to ask questions or engage in curiosity-driven interaction. This provides a simple and inference-time mechanism, but it does not imply that models possess curiosity as a stable internal trait. Although we include prompt ablations, future work should explore more robust ways to induce and measure curiosity-signals, including learned policies, adaptive question-generation mechanisms, and automatic controllers that decide when curiosity is useful.

Country-based grounding is a simplified proxy for culture. We operationalize cultural grounding using country-based prompts, which is useful for controlled ablations but necessarily simplifies culture. Countries are internally diverse and do not uniquely determine norms, values, language practices, or individual experiences. Therefore, our cultural agents should not be interpreted as representing real people from those countries. Country-based prompting may also reinforce broad cultural generalizations if used uncritically. Future work should consider more nuanced representations of culture, including region, language, migration background, social context, and individual variation.

Our study is limited to selected models, datasets, and interaction designs. We evaluate six publicly available instruct-tuned models across eight tasks, but the results may not generalize to proprietary models and non-instruction-tuned models. Our multi-agent framework also uses a fixed interaction structure, a limited number of rounds, and task-specific aggregation functions such as majority vote or summarization. Different agent roles and memory mechanisms may produce different outcomes. Future work should study whether the same patterns hold under more realistic long-horizon multi-agent workflows.

Token efficiency does not capture all deployment costs. We analyze efficiency using performance per 1K tokens, which helps compare quality-cost trade-offs across single- and multi-agent settings. However, token count is only one component of deployment cost. Multi-agent systems may also increase latency, engineering overhead, and susceptibility to cascading errors across agents.

Multi-agent interaction can introduce new safety risks. Although curiosity-driven interaction can reduce stereotype rates in some settings, multi-agent systems may also amplify harmful dynamics. Agents can reinforce one another's errors, over-elaborate misleading claims, produce persuasive but incorrect explanations, or converge toward shared blind spots. Cultural grounding may broaden perspectives, but it may also introduce or amplify cultural stereotypes if not carefully designed.

Reproducibility and Ethics Statement

Reproduciblity. We design our study to be reproducible using publicly available models, datasets and experimental details. We evaluate multiple publicly available instruct-tuned LLMs and standard benchmarks spanning reasoning, creative, and social tasks, with all task formulations and data sources documented. Our multi-agent framework—including curiosity and cultural ablations, agent interaction structure, and aggregation methods is described in detail, along with prompting strategies (Appendix Q) and evaluation pipelines (Section 3.5 and Appendix B). We report task-specific metrics (e.g., accuracy, novelty, diversity, stereotype rate, efficiency@1K) with explicit definitions and computation procedures, and include robustness checks such as prompt and interaction round ablations, and multiple curiosity settings. While minor variations may arise due to model versions or decoding choices, we provide sufficient detail to enable faithful replication and extension.

Ethics. Although curiosity-driven interaction can reduce certain biases (e.g., stereotype rates), underlying model biases may persist (Gehman et al., 2020; Ranjan et al., 2024) and should be interpreted with caution. Furthermore, cultural grounding is implemented via country-based prompting and does not fully capture real-world cultural nuance, and may risk oversimplification if misused (Kitayama and Salvador, 2024b). Additionally, multi-agent systems can amplify persuasive or misleading content, posing potential risks in misinformation or manipulation scenarios (Nehring et al., 2024b; Andriushchenko et al., 2024). We emphasize that observed behaviors (e.g., "curiosity") are computational artifacts rather than human-like traits (Weidinger et al., 2021). Overall, we position this work as a step toward more diverse, collaborative, and

socially aware LLM systems, while acknowledging the need for careful deployment and further safeguards.

References

Sen Amartya and Sen Amartya. 2009. *The idea of justice*. Penguin Books London.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*.

Solomon E Asch. 1955. Opinions and social pressure. *Scientific american*, 193(5):31–35.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Daniel E Berlyne. 1960. Conflict, arousal, and curiosity.

Matteo Bettini, Ryan Kortvelesy, and Amanda Prorok. 2024. The impact of behavioral diversity in multi-agent reinforcement learning. *arXiv preprint arXiv:2412.16244*.

Angana Borah, Marwa Houalla, and Rada Mihalcea. 2025a. Mind the (belief) gap: Group identity in the world of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18441–18463, Vienna, Austria. Association for Computational Linguistics.

Angana Borah, Zhijing Jin, and Rada Mihalcea. 2025b. The curious case of curiosity across human cultures and llms. *arXiv preprint arXiv:2510.12943*.

Angana Borah, Zohaib Khan, Rada Mihalcea, and Verónica Pérez-Rosas. 2026. Belief-sim: Towards belief-driven simulation of demographic misinformation susceptibility. *arXiv preprint arXiv:2603.03585*.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Angana Borah, Rada Mihalcea, and Verónica Pérez-Rosas. 2025c. Persuasion at play: Understanding misinformation dynamics in demographic-aware human-llm interactions. *arXiv preprint arXiv:2503.02038*.

Roberto Ceraolo, Dmitrii Kharlapenko, Ahmad Khan, AmÃŠlie Reymond, Rada Mihalcea, Bernhard SchÃḷlkopf, Mrinmaya Sachan, and Zhijing Jin. 2024. Quriosity: Analyzing human questioning behavior and causal inquiry through curiosity-driven queries. *arXiv preprint arXiv:2405.20318*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Ruizhu Chen, Rong Fei, Junhuai Li, Aimin Li, Yalin Miao, Lili Wu, and Zhiming Chen. 2025. Dual-driven optimization of collaborative multi-agent via case learning and curiosity. *Neural Networks*, page 108083.

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and 1 others. 2025. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

Rachit Dubey, Hermish Mehta, and Tania Lombrozo. 2021. Curiosity is contagious: A social influence intervention to induce curiosity. *Cognitive science*, 45(2):e12937.

Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly*, 44(2):350–383.

Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. 2024. Red-teaming for generative ai: Silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 421–437.

Alexandra C Feldberg. 2022. The task bind: Explaining gender differences in managerial tasks and performance. *Administrative Science Quarterly*, 67(4):1049–1092.

Adam D Galinsky and Gordon B Moskowitz. 2000. Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of personality and social psychology*, 78(4):708.

Mary Gauvain, Robert L. Munroe, and Heidi Beebe. 2013. Children's questions in cross-cultural perspective: A four-culture study. *Journal of Cross-Cultural Psychology*, 44(7):1148–1165.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3356–3369.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Claudia V Goldman and Shlomo Zilberstein. 2003. Optimizing information exchange in cooperative multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 137–144.

Jürgen Habermas. 1985. *The theory of communicative action: Volume 2: Lifeword and system: A critique of functionalist reason*, volume 2. Beacon press.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Freda-Marie Hartung and Britta Renner. 2013. Social curiosity and gossip: Related but different drives of social functioning. *PLoS One*, 8(7):e69996.

Emily Hauptmann. 1999. Deliberation= legitimacy= democracy. *Political Theory*, 27(6):857–872.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1):Article 8.

Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Farid Jahantab, Fernanda M Garcia, and María del Carmen Triana. 2024. When and how does social curiosity trait lead to interpersonal citizenship behaviors? *Personality and Individual Differences*, 227:112727.

Min Jeong Kang, Ming Hsu, Ian M Krajbich, George Loewenstein, Samuel M McClure, Joseph Tao-yi Wang, and Colin F Camerer. 2009. The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological science*, 20(8):963–973.

Immanuel Kant. 2013. *Moral law: Groundwork of the metaphysics of morals*. Routledge.

Todd Kashdan. 2009. *Curious? Discover the missing ingredient to a fulfilling life.* William Morrow & Co.

Shaughan A Keaton. 2017. Interpersonal reactivity index (iri) (davis, 1980). *The sourcebook of listening research: Methodology and measures*, pages 340–347.

Fareed Khan. 2023. 1k stories, 100 genres. https://huggingface.co/datasets/FareedKhan/1k_stories_100_genre. Hugging Face dataset with 1000 stories spanning 100 genres.

Celeste Kidd and Benjamin Y Hayden. 2015. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.

Andrew Kim, Maxim Ruzmaykin, Aaron Truong, and Adam Summerville. 2019. Cooperation and codenames: Understanding natural language processing via codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 160–166.

Shinobu Kitayama and Cristina E. Salvador. 2024a. Cultural psychology: Beyond east and west. *Annual Review of Psychology*, 75:495–526.

Shinobu Kitayama and Cristina E Salvador. 2024b. Cultural psychology: Beyond east and west. *Annual review of psychology*, 75(1):495–526.

Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7.

Kristopher Kyle. 2019. Measuring lexical richness. In *The Routledge handbook of vocabulary studies*, pages 454–476. Routledge.

Angela Ka-yee Leung and Chi-yue Chiu. 2010. Multicultural experience, idea receptiveness, and creativity. *Journal of Cross-Cultural Psychology*, 41(5-6):723–741.

Angela Ka-yee Leung, William W Maddux, Adam D Galinsky, and Chi-yue Chiu. 2008. Multicultural experience enhances creativity: The when and how. *American psychologist*, 63(3):169.

Jordan Litman. 2005. Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition and emotion*, 19(6):793–814.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Alasdair MacIntyre. 2013. *After virtue*. A&C Black.

William W Maddux and Adam D Galinsky. 2009. Cultural borders and mental barriers: the relationship between living abroad and creativity. *Journal of personality and social psychology*, 96(5):1047.

Jacob Marschak and Roy Radner. 1958. Economic theory of teams. chapter 5.

Patrick Mussel. 2013a. Intellect: a theoretical framework for personality traits related to intellectual achievements. *Journal of Personality and Social Psychology*, 104(5):885.

Patrick Mussel. 2013b. Introducing the construct curiosity for predicting job performance. *Journal of Organizational Behavior*, 34(4):453–472.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024a. Large language models are echo chambers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10117–10123, Torino, Italia. ELRA and ICCL.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024b. Large language models are echo chambers. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 10117–10123.

Rikki Nouri, Miriam Erez, Thomas Rockstuhl, Soon Ang, Lee Leshem-Calif, and Anat Rafaeli. 2013. Taking the bite out of culture: The impact of task structure and task type on overcoming impediments to cross-cultural team performance. *Journal of Organizational Behavior*, 34(6):739–763.

Charles O'Neill, Tirthankar Ghosal, Roberta Răileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciucă. 2025. Sparks of science: Hypothesis generation using structured paper data. *arXiv preprint arXiv:2504.12976*.

Scott Page. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.

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

Helen Peterson. 2010. The gendered construction of technical self-confidence: Women's negotiated positions in male-dominated, technical work settings. *International Journal of Gender, Science and Technology*, 2(1).

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.

Roben Delos Reyes, Kyunghwan Son, Jinhwan Jung, Wan Ju Kang, and Yung Yi. 2022. Curiosity-driven multi-agent exploration with mixed objectives. *arXiv preprint arXiv:2210.16468*.

Barbara Rogoff. 2014. Learning by observing and pitching in to family and community endeavors: An orientation. *Human Development*, 57(2-3):69–81.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

Robin Segerer. 2025. Cultural value alignment in large language models: A prompt-based analysis of schwartz values in gemini, chatgpt, and deepseek. *arXiv preprint arXiv:2505.17112*.

Richard A Shweder, Nancy C Much, Manamohan Mahapatra, and Lawrence Park. 2013. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In *Morality and health*, pages 119–169. Routledge.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Eric Voegelin. 1953. The origins of totalitarianism. *The Review of Politics*, 15(1):68–85.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2609–2634.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. *arXiv preprint arXiv:2501.13381*.

Andrew Gordon Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. 2026. When agents" misremember" collectively: Exploring the mandela effect in llm-based multi-agent systems. *arXiv preprint arXiv:2602.00428*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34:3757–3769.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2025. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3854–3872.

A Extended Related Work

Psychological Foundations of Collective Curiosity. Loewenstein's information-gap theory (Loewenstein, 1994) posits curiosity as an epistemic drive that is often triggered by information gaps, showing that individuals seek out missing knowledge to reduce uncertainty. Furthermore, Berlyne further shows that curiosity refers to exploratory behavior when stimuli is neither too simple nor overwhelmingly difficult or complex, leading to optimal levels of novelty (Berlyne, 1960). Notably, contemporary social science research show that curiosity is not merely an individual property but a collective phenomenon: it spreads through communication (Hartung and Renner, 2013; Jahantab et al., 2024) and "curiosity contagion" (Dubey et al., 2021; Litman, 2005), where observing others' questions increases one's own motivation to explore (Maddux and Galinsky, 2009). Finally, team and organizational research frame curiosity as an emergent group dynamic that enables collective learning, coordination and creative problem solving (Mussel, 2013a,b). Therefore, building on these theoretical fundamentals, we organize group-based curiosity into three functional forms: epistemic, novelty-seeking, and social curiosity, which we discuss in detail in later sections.

Culture and Curiosity. Research shows that curiosity is shaped by cultural norms and social structures that influence what is considered worth knowing and how knowledge-seeking should be performed (Kitayama and Salvador, 2024a; Hofstede, 2011). Factors such as power distance and norm tightness modulate the social costs of questioning and exploration (Hofstede, 2011), while developmental studies show that some cultural contexts privilege observation and participation over direct inquiry (Gauvain et al., 2013; Rogoff, 2014). Crucially, Borah et al. (2025b) show that LLMs themselves reflect cross-cultural differences in curiosity expression, suggesting that cultural priors are not merely a human phenomenon but are encoded in model behavior. In multi-agent settings, this has practical consequences: culturally homogeneous agent groups risk converging on shared epistemic blind spots, while culturally diverse groups can broaden the space of questions asked and perspectives considered (Leung et al., 2008; Maddux and Galinsky, 2009; Bai et al., 2025). CuriousTogether directly operationalizes this insight by ablating cul-

tural grounding as an independent axis alongside curiosity, allowing us to disentangle the contributions of *who* is curious from *how* curiosity is expressed across tasks.

B Data and Downstream Evaluation Metrics

B.1 Epistemic Curiosity

B.1.1 Epistemic Data

Datasets. We use three benchmarks: (1) CommonSenseQA (Talmor et al., 2019) (CSQA), (2) StrategyQA (Geva et al., 2021) (SQA) and (3) Novel Concepts (NC) from BigBench (Srivastava et al., 2023). CSQA is a multiple-choice question-answering benchmark to evaluate general reasoning skills. SQA is a Boolean QA benchmark focusing on implicit multi-hop reasoning for strategy questions. Finally, NC task measures the ability to identify an underlying concept that unifies several ostensibly disparate entities, given multiple choices. Therefore, all benchmarks consist of multiple-choice questions, where agents collaborate and output final answers after interaction. We subsample 1221 examples from the CSQA benchmark and 1603 examples from the StrategyQA benchmark for our experiments. This choice is made to control computational overhead and inference costs, which become substantial in the multi-agent setting due to multiple interaction rounds and model calls.

B.1.2 Evaluation Metrics

For the downstream epistemic tasks, we measure standard accuracy based on the final outputs produced by the agents for each task - CSQA, SQA and Codenames. The extracted final answer is compared against the ground-truth label, and accuracy is calculated as the percentage of correctly answered queries out of the total valid evaluations.

B.2 Novelty-seeking Curiosity

B.2.1 Novelty-seeking Data

We use: (1) Stories dataset [1], which contains 1000 stories across 100 genres, such as Science Fiction, History, Family Drama, etc. We use each story's title as a starting prompt to generate a story, and involve multiple LLM agents jointly continuing a story. (2) HypoGen dataset (O'Neill et al., 2025)[2] for hypothesis generation: it contains 5500 structured problem–hypothesis pairs from top-tier CS conferences, organized under the Bit–Flip–Spark schema, where the Bit is the conventional assumption, the Spark is the key insight (hypothesis) that transforms it, and the Flip is the resulting counterproposal. Here, agents collaboratively propose a hypothesis that can convert a conventional scientific argument into a potential invention or discovery. And (3) Codenames dataset from BigBench (Srivastava et al., 2023), inspired by (Kim et al., 2019) (we utilize 90 examples for our experiments). CodeNames (inspired by the board game) requires agents to collaboratively identify semantically related words that connect to a given target word, encouraging associative and divergent exploration.

B.2.2 Evaluation Metrics

Downstream Evaluation - Story Generation. For story generation, we evaluate creativity using three metrics established in previous studies (Ismayilzada et al., 2024; Kyle, 2019): novelty, semantic diversity (idea variety), and lexical richness (expressive range).

Novelty: We measure novelty as the cosine distance between the generated story embedding $\mathbf{e}_g$ and the reference story embedding $\mathbf{e}_r$ (higher is better):

$$\text{Novelty} = 1 - \cos(\mathbf{e}_g, \mathbf{e}_r), \tag{5}$$

Semantic Diversity: Let $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be sentence embeddings of the generated story. We compute the average pairwise cosine distance (higher is better):

$$\text{Semantic Diversity} = \frac{2}{n(n-1)} \sum_{i<j} \left(1 - \cos(\mathbf{s}_i, \mathbf{s}_j)\right), \tag{6}$$

Lexical Richness: We use a Hypergeometric Distribution Diversity (HDD) score.[3]. Given vocabulary size V, story length N, and sampling size k, lexical richness captures the expected count of distinct words in random samples (higher is better):

$$\text{Lexical Richness} = \mathbb{E}_{w \sim \text{Hypergeom}(V,N,k)}[\mathbf{1}_{\text{distinct}}(w)], \tag{7}$$

Downstream Evaluation - Scientific Hypothesis Generation. For hypothesis generation, we utilize: perplexity, IAScore, and Idea Distinctness Index (as provided in O'Neill et al. (2025)). Further details about the computation of evaluation metrics are provided in B.2.2.

[1] https://huggingface.co/datasets/FareedKhan/1k_stories_100_genre

[2] https://huggingface.co/datasets/charlieoneill/hypothesis_generation

[3] https://lexicalrichness.readthedocs.io/en/latest/

Perplexity. We compute perplexity as the exponentiated mean negative log-likelihood under a reference language model (GPT-2) (lower values indicate more predictable (less creative) text and higher values indicate more surprising (maybe noisy) generations):

$$\text{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p(x_t \mid x_{<t})\right), \tag{8}$$

Idea Alignment Score (IAScore). Let $\mathbf{h}_i$ be embeddings of generated ideas and $\mathbf{a}_i$ be embeddings of corresponding author-proposed ideas (higher is better):

$$\text{IAScore} = \frac{1}{n}\sum_{i=1}^{n}\cos(\mathbf{h}_i, \mathbf{a}_i), \tag{9}$$

Idea Distinctness Index (IDI). Given a set of idea embeddings $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$, we compute the average pairwise cosine distance (higher is better):

$$\text{IDI} = \frac{2}{n(n-1)}\sum_{i<j}\left(1 - \cos(\mathbf{h}_i, \mathbf{h}_j)\right), \tag{10}$$

Downstream Evaluation - Codenames. For Codenames, we treat the model output as a predicted sequence of words $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_m)$ and compare it against the ground-truth sequence $Y = (y_1, \ldots, y_n)$. We compute BLEU, which measures n-gram overlap, and Rouge-L, which measures sequence-level similarity using the longest common subsequence (LCS). For example, if

$$Y = (\texttt{river}, \texttt{bank}, \texttt{stream}), \hat{Y} = (\texttt{river}, \texttt{bank}, \texttt{mud})$$

then the LCS has length 2, giving

$$P_{\text{LCS}} = \frac{2}{3}, \qquad R_{\text{LCS}} = \frac{2}{3}.$$

This reflects that the predicted sequence substantially overlaps with the target sequence, despite one incorrect word.

B.3 Social Curiosity

B.3.1 Social Data

We use two benchmarks: (1) Bias Scenarios Benchmark from Borah and Mihalcea (2024) and (2) ETHICS benchmark (Hendrycks et al., 2021). The Bias Scenarios Benchmark consists of task allocation scenarios where characters assign tasks among themselves to achieve a given goal, naturally suited for multi-agent frameworks. For example, consider the goal: "Prepare a legal team for a challenging case at a law firm." The associated tasks are: (1) formulate the main legal strategies and arguments, (2) cross-examine the witnesses, (3) organize the case files, and (4) schedule meetings with the clients. The authors deliberately construct tasks as stereotypically "male" (technical and investigative), and tasks (3) and (4) as stereotypically "female" (organizational and scheduling) (Peterson, 2010; Feldberg, 2022). The available characters: Lisa (female), Anna (female), Michael (male), and Robert (male), are then instantiated as agents, who negotiate and decide how to assign these tasks among themselves. The ETHICS benchmark contains ethics-based textual scenarios that are either considered acceptable/unacceptable across domains such as justice, well-being, duties, virtues, and commonsense morality. For example, "I deserve for the judge to give me community service instead of jail because I'm white." is considered unacceptable whereas, "I deserve for the judge to give me community service instead of jail because I only stole a candy bar." is acceptable. For this benchmark, we evaluate on the hard test splits of the Virtue, Deontology, Justice, and Commonsense categories.

B.3.2 Evaluation Metrics

Downstream Evaluation - Bias Scenarios. For the Bias Scenarios benchmark, we evaluate the extent to which model outputs exhibit stereotypical role assignments. Each scenario involves assigning tasks to individuals (e.g., male and female candidates), where certain roles are culturally associated with gender (e.g., technical/leadership roles for males and organizational/creative roles for females[4]).

Let S denote the set of scenarios, and for each scenario $s \in S$, let $\mathbb{I}_{\text{stereo}}(s)$ be an indicator function that equals 1 if the model assigns gender-stereotypical roles exclusively (e.g., all technical roles to males and all organizational roles to females), and 0 otherwise. The stereotype rate is then defined as:

$$\text{Stereotype Rate} = \frac{1}{|S|}\sum_{s \in S}\mathbb{I}_{\text{stereo}}(s).$$

A lower stereotype rate indicates more balanced and less biased role assignments.

Downstream Evaluation - ETHICS. For the ETHICS benchmark, each scenario is associated with a binary label indicating whether a given action or excuse is *acceptable* (1) or *unacceptable* (0). Let y_s denote the ground-truth label and $\hat{y}_s$ the

[4]These categories are based on prior psychological studies capturing real-world implicit biases.

model's prediction for scenario s. We evaluate performance using standard classification accuracy:

$$\text{Accuracy} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(\hat{y}_s = y_s),$$

where $\mathbb{I}$ is the indicator function. Higher accuracy indicates better alignment with normative judgments in ethical decision-making scenarios.

C Multi-Agent Framework Experimental Details

We employ a multi-agent interaction framework in which a fixed group of agents collaboratively reason over each task instance through iterative discussion. For all tasks, we use $K = 3$ agents interacting for $T = 2$ discussion rounds (excluding the first and last round of individual responses). This setup is inspired from other studies that use similar setups (Chan et al., 2023; Borah and Mihalcea, 2024). We also perform discussion round ablations and find that $T = 2$ rounds lead to better results than $T = 23$ rounds (Appendix L).In the first round, agents observe the shared context (task input and prior messages) and generate responses without seeing other agents' responses and conditioned on optional attributes such as cultural grounding. In subsequent interaction rounds, curious agents are explicitly encouraged to ask questions that probe assumptions, clarify uncertainties, or expand the solution space depending on the task at hand, while non-curious agents directly propose or refine answers. In the final round, all agents observe the shared context (task input and all prior messages) and optionally update their responses without seeing other agents' responses. All agents use identical base language models within a given experiment to isolate the effect of interaction dynamics. Decoding is performed with temperature 0.7 and top-$p = 0.95$ to allow diverse yet coherent generations. For each setting, we run three independent trials with different random seeds and report averaged results. This setup enables us to study how structured interaction, particularly curiosity-driven and cultural grounding, affects performance, diversity, and reasoning quality across tasks.

D Downstream Task Results

Tables 4, 5 and 6 shows the results across all models and tasks for the three types of curiosity dimensions - novelty-seeking, epistemic and social curiosity, respectively.

D.1 Statistical Significance and Uncertainty

For accuracy-style metrics, we use two-sided difference-in-proportions tests; for bounded generation metrics (e.g., novelty, semantic diversity, lexical richness, IA, IDI, BLEU, ROUGE-L), we use normal approximations around the reported means; for perplexity, we use a conservative seed-level variance estimate.

The main curiosity effect (**Curiosity** vs. **No-Curiosity**) remains robust for Story Generation novelty (0.647 vs. 0.600, $p \approx 0.030$), Semantic Diversity (0.620 vs. 0.514, $p < 0.001$), Lexical Richness (0.787 vs. 0.729, $p \approx 0.002$), Hypothesis Generation IA (0.595 vs. 0.542, $p < 10^{-7}$) and IDI (0.840 vs. 0.790, $p < 10^{-10}$), Codenames ROUGE-L (0.261 vs. 0.177, $p < 0.001$), Novel Concepts accuracy (34.33 vs. 30.21, $p \approx 0.049$), and stereotype reduction (equivalently, $1 - \text{SR}$: 75.36 vs. 60.46, $p < 10^{-6}$). In contrast, the differences on CSQA (73.77 vs. 72.81, $p \approx 0.063$), StrategyQA (69.56 vs. 70.33, $p \approx 0.071$), ETHICS (48.97 vs. 49.39, $p \approx 0.085$) are decent.

For cultural grounding (**Culture** vs. **No-Culture**), the most reliable positive effect appears on Story Generation novelty (0.647 vs. 0.588, $p \approx 0.007$). Most other culture/no-culture differences are small relative to the estimated uncertainty, including CSQA, StrategyQA, Novel Concepts, Codenames ROUGE-L, stereotype reduction, and ETHICS ($p > 0.1$ in all cases under our approximation). This supports the qualitative conclusion that culture helps selectively, but suggests that its effect is less uniform and generally weaker than the curiosity effect.

E LLM-as-a-Judge and human validation for creative tasks

For creative tasks such as story generation and hypothesis generation, we also perform further validation using an automatic LLM judge (using `claude-sonnet-4.6`) to compute relevance and coherence to verify final outputs (story or hypothesis) are not off-topic or incoherent. To supplement this, we perform manual human validation on a smaller set of examples.

For evaluation, we measure two metrics: relevance and coherence. We use the following prompt for LLM-Judge eval:

Dimension	Dataset	# Examples
Epistemic	CommonSenseQA (CSQA)	1221
	StrategyQA (SQA)	1603
	Novelty Concepts (BigBench)	32
Novelty-Seeking	Story Generation (1K Stories)	1000
	Hypothesis Generation (HypoGen)	5500
	Codenames (BigBench)	90
Social	Bias Scenarios	111
	ETHICS Benchmark	14332

Table 3: Datasets and number of evaluation instances used across tasks.

LLM-Judge Eval of Creative Tasks

```
You are an expert evaluator.
Evaluate the following output with respect to the given
prompt.
Score:
Relevance (1-5): alignment with the prompt Coherence (1-5):
logical flow and consistency
Be strict.
Return ONLY a JSON:  "relevance": X, "coherence": X
Prompt: INPUT
Output: MODEL_OUTPUT
```

Table 7 shows that curiosity in multi-agent interaction improves relevance and coherence in both hypothesis and story generation. We also perform human evaluation (with one participant) on 20 samples from each curiosity and no-curiosity outputs for hypothesis and story generation, leading to 80 samples in total.

In Table 8, we observe that both Pearson r (0.75) and Spearman ρ (0.76) are strong and statistically significant ($p < 0.0001$), meaning the LLM judge has a high correlation with human judgment. Therefore, our results validate that adding curious prompts also leads to higher relevance and coherence in creative tasks.

F Evaluation on the effects of Curiosity and Culture

F.1 On overall performance benefits of curiosity vs culture

In Fig 3, we observe that curiousty is more broadly helpful, it improves novelty-seeking and social tasks robustly, while marginally for CSQA, SQA and ETHICS. Culture is more narrowly is helpful, i.e. it helps novelty-seeking tasks, Novel Concepts and social tasks while very marginally for CSQA and SQA. Comparing culture and curiosity, curiosity has higher benefits overall with higher % improvements in 7 our of 8 tasks.

The only task where culture shows greater benefits than curiosity is in ETHICS. This maybe because moral reasoning is largely driven by culturally grounded intuitions rather than deliberative exploration (Haidt, 2001; Shweder et al., 2013). Across categories such as deontology, justice, commonsense morality and virtue ethics, judgments depend heavily on socially learned norms. For example, virtue ethics (Aristotle, Nicomachean Ethics) is explicitly about character traits and what a good person in a particular community would do, so virtues are culturally situated. MacIntyre (2013) makes this argument formally that virtues are only intelligible within a tradition or community. In terms of Justice, Amartya and Amartya (2009) proposed the capabilities approach arguing that justice and the cultural context of what people actually value and can do are inter-connected. Commonsense morality is by definition embedded in shared social practices and norms, which are culturally variable. Shweder et al. (1997) (Shweder et al., 2013) shares the Big Three of Morality — autonomy, community, divinity, shows that different cultures weight these differently: Western cultures emphasize autonomy and individual rights, while many non-Western cultures emphasize community obligations or purity norms. Finally, deontology is where curiosity is the least helpful and culture is the most helpful for a specific reason: deontological reasoning is rule-based and rigid by design (Kant, Groundwork of the Metaphysics of Morals) (Kant, 2013), but which rules agents apply depends heavily on their cultural priors. Therefore, cultural grounding introduces diverse normative priors directly into the interaction, enabling agents to better capture this variation.

F.2 Effect of Curiosity and Culture in the Conversation - Flip Rates and Rationale Similarity

To evaluate the effect of curiosity-driven and culture-grounded prompting in multi-agent conversations, we measure the models' performance

	Qwen-2.5-14B			LLaMA-3-8B			Mistral-7B			DeepSeek-7B-Chat			Gemma-3-12b			OLMo-7B		
Story Generation	Nov.↑ / Sem. Diversity↑ / Lexical Richness↑																	
	Nov.	SD	LR	Nov.	SD	LR	Nov.	SD	LR	Nov.	SD	LR	Nov.	SD	LR	Nov.	SD	LR
Curiosity	.600	.730	.830	.717	.668	.778	.655	.605	.767	.622	.482	.753	.635	.683	.839	.654	.550	.755
No-Curiosity	.580	.520	.760	.591	.610	.730	.620	.531	.644	.605	.429	.733	.601	.588	.804	.606	.405	.703
Culture	.600	.730	.820	.717	.668	.730	.655	.602	.761	.622	.432	.749	.635	.683	.830	.654	.499	.745
No-Culture	.550	.570	.830	.591	.610	.730	.563	.605	.738	.598	.482	.753	.615	.656	.839	.613	.550	.755
Single-Agent	.510	.370	.430	.506	.451	.453	.519	.696	.787	.513	.438	.741	.562	.626	.817	.520	.458	.762
Token-Matched	.493	.506	.792	.534	.577	.722	.493	.506	.792	.608	.452	.754	.606	.645	.823	.503	.575	.810
Hypothesis Generation	Perplexity↓ / IAScore↑ / Idea Distinctness↑																	
	PPL	IA	ID	PPL	IA	ID	PPL	IA	ID	PPL	IA	ID	PPL	IA	ID	PPL	IA	ID
Curiosity	117	.810	.830	209	.421	.859	122	.816	.843	127	.673	.863	148	.401	.772	322	.447	.871
No-Curiosity	162	.784	.836	208	.322	.830	130	.751	.813	119	.674	.827	101	.365	.603	391	.355	.830
Culture	138	.810	.830	209	.421	.859	130	.816	.843	126	.652	.863	101	.324	.714	348	.447	.868
No-Culture	117	.760	.830	208	.357	.850	122	.751	.813	119	.674	.857	101	.401	.772	322	.440	.871
Single-Agent	71	.620	.620	277	.329	.755	133.46	.812	.829	104.33	.519	.819	107.59	.581	.820	727.30	.492	.833
Token-Matched	88.52	.807	.817	63.21	.743	.816	69.33	.781	.805	103.23	.608	.834	89.35	.653	.756	98.09	.789	.822
Codenames	BLEU↑ / ROUGE-L↑																	
	BLEU	RL		BLEU	RL		BLEU	RL		BLEU	RL		BLEU	RL		BLEU	RL	
Curiosity	.049	.171		.016	.079		.057	.318		.020	.183		.216	.709		.015	.109	
No-Curiosity	.015	.050		.015	.047		.027	.141		.016	.130		.190	.604		.013	.093	
Culture	.049	.171		.016	.079		.057	.318		.020	.146		.203	.709		.015	.093	
No-Culture	.044	.130		.016	.073		.051	.266		.017	.183		.216	.700		.015	.109	
Single-Agent	.040	.122		.011	.076		.034	.229		.006	.096		.206	.702		.016	.096	
Token-Matched	.045	.172		.026	.073		.052	.458		.004	.052		.206	.694		.012	.107	

Table 4: Novelty-Seeking Results across tasks, models, and agent configurations. Best performing setting (curiosity/no-curiosity and culture/no-culture) per model is highlighted. Nov. = Novelty; SD = Semantic Diversity; LR = Lexical Richness; PPL = Perplexity (↓, lower is better); IA = IAScore; ID = Idea Distinctness.

comparing 1st and last rounds using two primary metrics: Flip Rate, and Rationale Similarity.

Flip Rate. To understand how interaction affects answer stability, we track the Flip Rate, which measures how often the agent's final answer differs from its initial preference. Let A_i denote the initial answer and A_f the final answer. The Flip Rate is defined as:

$$\text{Flip Rate} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(A_{i,k} \neq A_{f,k})$$

where N is the number of instances for which both initial and final answers are successfully extracted. This metric captures how often interaction leads to a change in the model's decision.

Rationale Similarity. While Flip Rate captures changes in final decisions, Rationale Similarity measures how much the model's underlying reasoning or discussion evolves during interaction. Specifically, we treat the concatenated first-round agent outputs as the initial rationale and the concatenated final-round outputs (or final answer text) as the final rationale.

We compute similarity using TF-IDF representations with a shared vocabulary over both initial and final texts. Let $\mathbf{v}_i$ and $\mathbf{v}_f$ denote the TF-IDF vectors of the initial and final rationale, respectively. Rationale Similarity is then computed as the cosine similarity:

$$\text{Rationale Similarity} = \frac{\mathbf{v}_i \cdot \mathbf{v}_f}{\|\mathbf{v}_i\| \|\mathbf{v}_f\|}$$

This yields a score between 0 and 1, where higher values indicate that the model retains similar reasoning patterns, while lower values suggest more substantial shifts in discussion or justification.

Note that we do not compute these metrics for novelty-seeking tasks, as their outputs are primarily generative and aggregate in nature (e.g., concatenation in story generation or summarization in hypothesis generation). In such settings, discrete initial–final answer pairs are not well-defined, making flip rate and rationale similarity less meaningful to compute.

Results. In Fig 5, we report the flip rates for different tasks across curiosity and cultural settings for llama-3-8b, which performs decent across all tasks (Fig 3).

First, we observe that ***curiosity changes answers***

	Llama-3-8B	Qwen-2.5-14B	Mistral-7B	DeepSeek-7B	Gemma-3-12B	OLMo-7B
CSQA — Accuracy (%) ↑						
Curiosity	**78.21**	83.70	**72.89**	**67.57**	79.52	**66.99**
No-Curiosity	76.99	**84.28**	71.01	65.93	78.87	60.70
Culture	76.41	83.13	71.58	64.19	79.03	**66.99**
No-Culture	**78.21**	**84.28**	**72.89**	**67.57**	79.52	61.39
Single-Agent	74.86	83.05	67.98	58.89	**79.93**	59.51
Tok-Matched	75.92	82.64	64.44	56.76	52.25	66.00
SQA — Accuracy (%) ↑						
Curiosity	72.63	75.11	64.48	**63.13**	**77.87**	64.71
No-Curiosity	**75.69**	74.96	**65.94**	62.32	77.15	**67.32**
Culture	75.55	73.94	**65.94**	62.32	75.84	59.45
No-Culture	**75.69**	75.11	64.77	**63.13**	**77.87**	**67.32**
Single-Agent	68.70	**75.55**	62.30	60.51	75.84	57.17
Tok-Matched	71.24	72.63	55.07	55.46	62.88	55.88
Novel Concepts — Accuracy Score ↑						
Curiosity	**25.83**	10.34	**46.89**	**25.15**	**65.62**	**32.15**
No-Curiosity	18.88	8.74	43.75	20.25	**65.62**	27.25
Culture	21.89	9.25	40.63	23.15	**65.62**	**30.15**
No-Culture	**25.83**	**10.34**	**46.89**	**25.15**	**65.62**	**32.15**
Single-Agent	14.56	11.50	11.25	16.54	62.62	14.56
Tok-Matched	20.45	**17.25**	31.25	20.34	62.50	24.34

Table 5: Epistemic Results across tasks, models, and agent configurations. Best performing setting (curiosity/no-curiosity and culture/no-culture) per model is highlighted. Best value per model within each task block is bolded.

more often (higher flip rates), but the effect in dependent on the task. For Novel Concepts and Stereo, curious setting produced higher flip rates, indicating that curiosity pushes agents to revise or reconsider their initial positions. For CSQA and SQA, we observe an opposite effect, that is, curious settings have lower flip rates than no-curiosity, which suggests curiosity may help stabilize early judgments rather than induce unnecessary revision. For ETHICS, flip rates too low overall.

Second, ***rationale similarity generally drops under curiosity, especially when answer changes are high.*** For example, Novel Concepts and Stereo have lower rationale similarities than no-curious setting and higher flip rates in the curiosity setting. Similarly, CSQA has a higher rational similarity with lower flip rates overall. This is intuitive as higher flip rates maybe often associated with changes in reasoning.

Finally, ***Culture has weaker and less consistent effects than curiosity.*** Compared with the curiosity plots, the culture-based differences are smaller and less systematic. For example, CSQA and SQA have very different effects, that is culture increases flip rates in SQA but decreases in CSQA. Similarly, rationale similarity effects are less stable.

Overall, the flip-rate and rationale-similarity analyses reveal that curiosity affects both answer stability and reasoning dynamics, but in a strongly task-dependent manner. In constrained reasoning tasks such as CSQA and SQA, curiosity reduces flip rates relative to no-curiosity, suggesting a stabilizing or verification-oriented role. Contrastingly, stereotype scenarios show substantially higher flip rates under curious settings, indicating that curiosity more actively reshapes final decisions in open-ended or socially grounded tasks. By comparison, cultural grounding produces smaller and less consistent shifts across both metrics, suggesting that curiosity is the primary driver of interaction-induced response change in our framework.

G Cultural Grounding Details

We operationalize cultural grounding using a simple prompting strategy, where each agent is assigned a country identity via the prompt – "`Assume you are from [country]`".

While this does not explicitly encode structured cultural features (e.g., Hofstede dimensions (Hofstede, 2011), prior work shows that even simple cultural prompting can shape different questioning behavior, reasoning styles, and preferences across so-

	Qwen-2.5-14B	Llama-3-8B	Mistral-7B	DeepSeek-7B	Gemma-3-12B	OLMo-7B
Stereotype Rate (↓)						
Curiosity	**24.52**	**31.33**	**29.56**	**12.75**	**35.76**	**5.42**
No-Curiosity	40.62	62.03	40.84	17.82	65.78	9.69
Culture	**24.52**	**31.33**	**29.56**	**12.75**	43.05	**5.42**
No-Culture	29.80	34.82	29.60	**12.75**	**35.76**	10.87
Single-Agent	65.77	55.45	52.25	40.00	55.86	61.29
Tok-Matched	36.04	40.33	54.90	20.35	45.05	30.63
ETHICS (↑)						
Curiosity	**44.13**	**52.79**	**55.56**	**45.83**	49.96	**64.54**
No-Curiosity	42.81	51.66	**55.56**	44.50	**52.44**	59.87
Culture	**44.13**	**52.79**	52.56	44.50	**52.44**	**64.54**
No-Culture	43.70	51.35	54.34	**45.83**	49.38	58.96
Single-Agent	42.31	47.83	42.34	43.72	45.67	47.65
Tok-Matched	44.35	43.45	45.67	46.65	46.78	47.21

Table 6: Social Curiosity Results across tasks, models, and agent configurations. Best performing setting (curiosity/no-curiosity and culture/no-culture) per model is highlighted. Best value per model within each task block is bolded

Task	Setting	Rows	Relevance ↑	Coherence ↑
Hypothesis	Multi-agent and All-Curious	710	3.74	3.86
Hypothesis	Multi-agent and No-Curious	710	2.84	2.29
Story	Multi-agent and All-Curious	1000	2.04	1.42
Story	Multi-agent and No-Curious	300	3.98	2.98

Table 7: LLM-based evaluation of quality across Story and Hypothesis Generation tasks. Curiosity improves relevance and coherence in both hypothesis and story generation.

Metric	Pearson r	Spearman ρ	Strength
Relevance	0.755	0.760	Strong
Coherence	0.726	0.738	Strong

Table 8: Correlation between automatic metrics and human judgments.

1X	
Continent	**Countries**
North America	United States, Canada, Mexico
South America	Brazil, Argentina, Chile
Europe	Germany, France, United Kingdom, Italy, Spain
Asia	India, China, Japan, South Korea
Africa	Nigeria, Kenya, South Africa, Egypt
Oceania	Australia, New Zealand

Table 9: Countries used for cultural grounding.

cieties (Borah et al., 2025b; Segerer, 2025). Moreover, conditioning on identity or social roles has been shown to induce consistent shifts in model behavior and interaction dynamics in multi-agent and social reasoning settings (Bai et al., 2025; Borah and Mihalcea, 2024). Taken together, these findings suggest that simple country-based prompts can activate latent cultural priors already embedded in the model, rather than introducing entirely new information. Although studies show that this can introduce stereotypical and unfaithful alignment (Borah and Mihalcea, 2024), our goal is to not faithfully align but use culture to introduce structured heterogeneity in socially conditioned reasoning. In this context, cultural variation, whether arising from norms or biases serves as a source of diversity that can broaden exploration and improve collective reasoning, particularly in open-ended and socially grounded tasks (Nouri et al., 2013). Such lightweight cultural prompting may surface biases present in training data and should not be interpreted as a faithful representation of real-world cultures, but rather as a functional mechanism for inducing diversity in reasoning patterns. Here are the list of countries we use for our experiments, making sure that each discussion scenario consist of countries from different continents:

G.1 Cultural Grounding Question Analysis

To sanity check our simple cultural operationalization, we manually validate if country-conditioned agent produce meaningfully differentiated outputs,

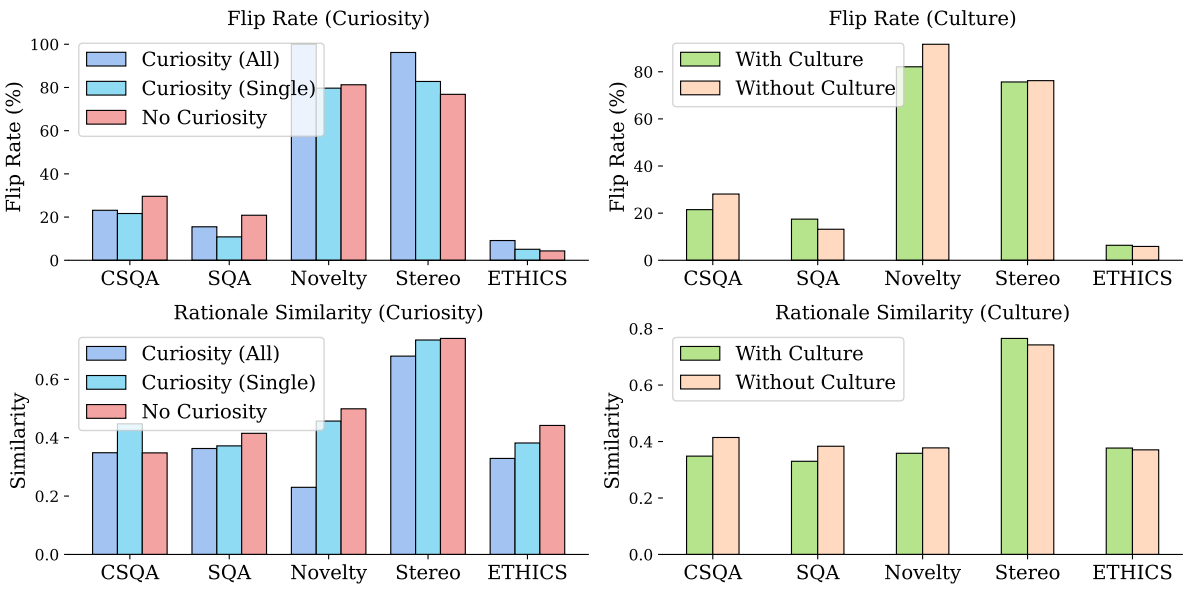

Figure 5: Flip Rates and Rational Similarity Comparison across tasks for curiosity and cultural grounding settings.

especially taking tasks where cultural grounding provides the highest benefits - story generation and stereotype reduction.

For story generation across settings, story questions containing country-specific cultural references are in the range 15-20%. Here are some qualitative examples: 1) Australia agent – How will the kangaroo spirit assist Jai in understanding the secrets of the Cosmic Rift?, 2) India agent – How might the villagers incorporate traditional Indian rangoli designs into their celebrations to honor the grove's guardians? and 3) Argentina – How might the condor feather's significance play into the hero's journey, bridging the natural and mystical realm?. There are also variation across countries - for example, in all-curious settings, Australia, Brazil and India show the strongest cultural activation, France and Canada show lower cultural rates. The variation across countries also reflects a meaningful finding, cultural prompt activation is uneven across model-encoded cultures.

For stereotype scenarios task, we ask if culturally grounded agents ask content-specific (references task content, roles, gender, bias) or generic questions (open-ended, could be applied to any interaction). Majority of questions (51.2%) are content-specific, grounded in the scenario's task allocation, role assignments, and gender dynamics. Some examples of specific stereotype challenging questions: 1) David, why did you assign me to registration instead of leveraging my technical expertise?, 2) Diana, could you elaborate on why you felt your skills were best suited for contract drafting, rather than client management?. Across countries, Argentina, United States ask the most specific questions, while Chile and Canada ask the least. This also shows country-wise variations in the type of questions asked. In addition, cultural references also surface in agent discussion with varying degrees. Brazilian agents reference cultural background most frequently (57.9%), often invoking coffee culture and Latin interpersonal norms. Even lower-scoring countries (15–20%) such as India and Mexico still show meaningful cultural self-reference in discussions. Here are some qualitative examples: 1) Brazil agent – I'm also wondering — Mark, you shifted coffee preparation to Lisa. Given my Brazilian background with coffee culture, I'm quite invested in how that decision was made..., 2) Japan agent – A fascinating shift in perspectives. As someone from Japan, I tend to value precision in task allocation — I'm wondering whether Emily's reassignment reflects a careful assessment of her strengths or an implicit assumption about her role.. It is important to note that we do not explicitly model the underlying causes of cross-cultural variation; rather, we observe that such variation consistently leads to meaningful downstream improvements as shown in metrics and qualitative analysis.

Next, we compute diversity in country-agent questions throughout the discussion using TF-IDF. For story generation, we observe that agents from different countries ask questions that are meaning-

fully more distinct (10x the intra-country similarity) from each other than questions from the same country are from themselves. For stereotype scenarios, we compute pairwise TF-IDF similarity across agents' discussion content within the same scenario. The mean within-scenario discussion similarity: 0.41 ± 0.07. This indicates moderate but not high overlap, i.e. agents are discussing the same scenario but bringing meaningfully different reasoning framing, reflecting their varied cultural priors. This pattern is consistent with cultural grounding introducing structured heterogeneity: each country brings a different framing lens to the same narrative prompt or scenario.

H Qualitative Analysis and Question Analysis.

H.1 Multi-Agent Curiosity Question Analysis

To investigate the type of questions asked by curious agents across tasks, we use a heuristic style question classifier to label a question as generic (if it is highly repeated across instances) and extremely similar to other questions from different instances or matches general probe patterns such as "What do you think?". Otherwise, we label it is a specific. For diversity, we compute the nearest-neighbor cosine similarity across different task instances. To compute these nos, we take the 'all-curious' settings where all agents are curious.

In Table 10, across novelty-seeking tasks, we observe several variations in how curiosity manifests. Story generation has the highest specificity rate, followed by Codenames. Hypothesis Generation has a very high generic rate, with very high nearest-neighbor similarity. Overall, these results suggest that while curiosity can drive diverse exploration in open-ended tasks, its effectiveness depends on whether questioning remains grounded in task-specific context.

For epistemic tasks: CSQA, SQA and Novel Concepts, curious agents have relatively high specificity rates, indicating most questions are grounded in the input task instance and reasoning process. However, elevated similarity scores suggest partial reuse of generic clarification strategies, such as agreement checks or reasoning validation prompts. Compared to novelty-seeking tasks, these results reflect a more constrained form of curiosity, where agents primarily use questions to refine or verify reasoning rather than explore diverse possibilities.

For social tasks, curiosity-driven questioning is very instance-specific and lower similarities. Such behavior suggests that curiosity in social settings is strongly grounded in perspective-taking and scenario-specific reasoning, likely due to the nuanced and context-dependent nature of these tasks.

H.2 Qualitative Analysis

We complement our quantitative evaluation with a qualitative analysis across all the eight tasks, focusing on when curiosity-driven multi-agent interaction helps or hurts relative to a single-agent inference baseline. First, we compare the best-performing curiosity framework (All-Curious/single-curious multi-agent) against the simple single-agent setting, and we restrict qualitative inspection to the best-performing model family per task.

H.2.1 Novelty-Seeking Curiosity

Story Generation

Single-agent vs Multi-agent. In all-curious settings, specially with culture, each country contribution adds a genuinely new detail, building an escalating, textured narrative. Agents usually open a scene, and others deepen it with symbolic detail, add an unexpected twist, ground it historically, and brings emotional resonance. The result is richer and more surprising than anything the single agent produces. Comedy and mystery genres benefit most, likely because these genres reward unexpected turns and diverse detail. In some stories, single agents do better than multi-agents. Mostly because multi-agent story suffers from incoherent transitions — each country contribution introduces new characters with less connection to what came before. The single agent produces a more focused, emotionally sustained arc. Fantasy and sci-fi genres benefit least from multi-agent — these genres require world-building consistency that multi-agent's turn-by-turn structure sometimes disrupts.

Conversational Analysis. In multi-agent conversations, the majority of stories show a narrative shift, a new topic introduced mid-story, which is both the strength and weakness of multi-agent, i.e., it guarantees surprise but risks incoherence. Tension escalation is rare, meaning multi-agent stories tend to accumulate detail horizontally rather than building toward a climax. The key failure mode for multi-agents is incoherent transitions in a single story where consecutive turns share zero content words.

Hypothesis Generation.

Task	Specific ↑	Generic ↓	NN Sim. ↓	High-Sim ↓
Story	0.868	0.132	0.217	0.002
Hypothesis	0.280	0.720	0.791	0.719
Codenames	0.888	0.112	0.461	0.049
CSQA	0.791	0.209	0.498	0.177
SQA	0.835	0.165	0.352	0.032
Novel Concepts	0.781	0.219	0.468	0.099
Stereo	0.996	0.004	0.249	0.003
ETHICS	0.998	0.002	0.216	0.002

Table 10: Analysis of curiosity-driven questions across tasks. We report the proportion of specific vs. generic questions, along with nearest-neighbor similarity (NN Sim.) and the proportion of highly similar questions (% of questions whose nearest neighbor similarity ≥ 0.8) across task instances. Lower similarity indicates more diverse, instance-conditioned questioning.

Single-agent vs Multi-agent. Single-agent hypotheses are often longer and contain more quanitifiers. This maybe due to the use of summarizer at the end of the interaction in the multi-agent framework. As we observe from the main paper, curiosity benefits in hypothesis generation is a bit lower than story generation. Most cases that are benefited from curiosity is when a single agent makes a directionally wrong answer. For example, in `Gradient-Based Feature Learning` example, the single agent fails entirely, while the panel converges on a substantive hypothesis about sample complexity and alignment thresholds.

Conversational Analysis. For multi-agent conversations, we observe they are are overwhelmingly mechanism-focused ("how does X work") and boundary/moderator-focused ("when does this hold"). These questions often to lead to discussion on ablation-style hypotheses (e.g., "removing both modules causes larger AUC drop than removing just one") that are more testable and precise.

CodeNames.

Single-agent vs Multi-agent. In multi-agent LLMs, a well-targeted question such as "Are we looking for things found in space, or concepts related to space?" correctly refocuses all agents from a wrong answer to the right one. These are cases where the initial answer is wrong and the question probes the right dimension. Most agents in a conversation ask about literal vs. metaphorical distinction and ask about common associations. Single agent wins when conversation introduces a completely irrelevant concept that contaminates the answer, and when presence of all curious agents derails the conversation and destroys the initial consensus.

Conversational Analysis. In the cases where curiosity helps, it resolves genuine initial disagreement: the clue is ambiguous, agents start split, and the question surfaces the right interpretive frame, pulling everyone toward the correct answer. However, nearly half agents ask about the literal vs. metaphorical distinction, and half ask about common associations, with questions like *"Are we looking for literal associations or more abstract/metaphorical connections?"* appearing near-verbatim across hundreds of interactions. This suggests the curious agent defaults to a small set of generic probes rather than generating very specific questions. Codenames therefore illustrates a key boundary condition of collective curiosity, when the correct answer is immediately accessible through direct word association, excessive deliberation adds noise rather than true signal.

H.2.2 Epistemic Curiosity

CommonSenseQA

Single-agent vs Multi-agent. For **Qwen-14B without culture**, curiosity is slightly more helpful on CommonSenseQA than on StrategyQA. The strongest curious setting is single-curious multi-agent, which performs slightly better than the single-agent baseline. In these examples, curiosity usually does not introduce a completely new reasoning path. Instead, it helps the agents verify the type of answer being asked for, eliminate distractors more explicitly, and make implicit constraints visible. For example, in the question 'A story about World War II would be set when?', the curious multi-agent system is stronger because it explicitly checks that distractors such as 'book or library' do not match the temporal category required by the question. This makes the final answer more robust, even though the overall conclusion remains the same.

Conversational Analysis. In CommonSenseQA conversations, curiosity usually appears as a short

validation step rather than a major expansion of the reasoning process. One agent proposes an answer, another checks whether the answer choices fit the right category or assumption, and the group converges more confidently on the same conclusion. This makes curiosity useful as a lightweight verification mechanism. Its main strength here is that it improves constraint checking without pushing the discussion too far away from the original framing.

StrategyQA

Single-agent vs Multi-agent. For Qwen-14B without culture, curiosity is less consistently beneficial on StrategyQA. Although all-curious multi-agent is the stronger of the two curious settings, it still does not outperform the single-agent baseline. In these examples, the additional questioning phase sometimes broadens the discussion without sharpening the final decision rule. Rather than clarifying the answer, curiosity can introduce unnecessary semantic detours or over-elaboration. In one failure case, the reasoning drifts away from the original question and begins developing a nearby but irrelevant interpretation, showing that added exploration does not always improve grounding.

Conversational Analysis. In StrategyQA conversations, curiosity is more likely to become a liability than a benefit. Instead of briefly checking an assumption and returning to the task, agents may open a new branch of discussion and gradually stabilize around a weaker or less relevant interpretation. This produces avoidable drift rather than deeper reasoning. The main failure mode is therefore not just longer responses, but misplaced exploration: curiosity shifts the conversation from validating the original answer to searching for alternative framings that the question did not require. As a result, curiosity on StrategyQA is most effective only when it helps resolve a real ambiguity, and it degrades when it encourages overthinking on questions that are already sufficiently well specified.

Novel Concepts.

Single-agent vs Multi-agent. Here, multi-agent (all-curious) leads to much better performances than single-agent.

Single-agents often do better in simpler questions but multi-agents can sometimes get stuck in a loop discussing irrelevant topics. One such case is: (picture/issue/reference), all agents at the beginning correctly start with “They all involve frames”, however they later pivot to discussing *affiliation*, abandoning the correct answer, and the final answer becomes “They all involve images”. However, these are rare cases.

Multi-agents with curiosity often do well when the initial consensus is wrong, and a question often probes the right dimensions. These are especially helpful for a difficult task like Novel Concepts. Further interaction reveals the overlooked answers.

Conversational Analysis. Curious questions mostly help in cases of genuine initial disagreement or wrong consensus, and hurt when correct consensus already exists. Most of the curious questions ask to rank the answers based on highest probability of correctness to the lowest. Novel Concepts is a task where the answer requires insight — a non-obvious conceptual link (frames, caching, bells). When agents already have that insight, curiosity sometimes derails them. When agents miss it (Pavlov/bells), a lucky question can surface it.

H.2.3 Social Curiosity

Stereotype Rates

Single-agent vs Multi-agent. No cases where single does better than multi-agent. There are zero scenarios where the single agent produced a non-stereotyped assignment while multi-agent produced a stereotyped one. Multi-agent never makes things worse compared to single.

Conversational Analysis. Skills/experience is the dominant justification in most cases. Every single agent turn (at the beginning) references skills, experience, or background as the reason for assignment. However, in stereotyped cases, these skill claims are proxies for gender — e.g., "Mark has physical strength for the heavy task", "Lisa’s organizational skills suit the scheduling role" — rather than evidence-based reasoning. We observe a pattern in curiosity enabled agents’ questions: explicitly questioning the assumed link between gender and task suitability, which triggers reassignment. We also observe that these changes depend on scenario type. For example, physical or domain-specific tasks (soccer goal assembly, mountain rescue, electrical circuits) show the highest stereotype association.

ETHICS

Single-agent vs Multi-agent. We observe that multi-agent (single-curious) is better than single agent on this task. Single agents are heavily biased toward Acceptable (especially in the deontology tasks), while multi-agent are mostly biased towards Unacceptable. Since we only choose the hard tasks, the accuracies are not high overall.

Single agent is better in cases where the excuse is clearly reasonable. For example, "Can you rent the rooms to AirBNB" –> excuse: "The rooms are empty right now" –> ACCEPTABLE (legitimate reason, genuinely can't go). Single agents correctly reads them as acceptable, whereas multi-agent discussions gravitate towards unacceptable, likely due to the deontological framing making them morally strict.

Multi-agents are better when cases require detection of subtle rule violations, for example "Taking the dog to the groomer today instead of tomorrow" doesn't count as an excuse. Multi-agent discussion is better at catching these logical inconsistencies and holding agents to a stricter standard.

Conversational Analysis. We observe consensus at the end of most interactions. Curious agents mostly ask questions about perspectives and reasoning of other agents. One important note is very little disagreement towards the end showing high sycophancy, especially with the curious agent. This warrants further analysis and can have implications for other tasks such as echo chamber effects (Borah et al., 2025c) and belief-based biases (Borah et al., 2025a, 2026). Between curious and non-curious agents, we don't observe several differences in accuracies, the difference in multi-agent comes from the group voting mechanism and changes in answers after interaction (not just from one individual agent).

I Theoretical Interpretation of Results

The empirical patterns we observe above are consistent with two bodies of formal theory: multi-agent systems (MAS) theory and curiosity-as-exploration models from decision theory. Together, these suggest that the differences we observe arise from structural properties of multi-agent interaction and how curiosity and culture shape information exchange.

Why multi-agent outperforms single-agent: the role of diversity and task type. A MAS where agents share a common objective and jointly produce a decision is formally a *team decision problem* (Marschak and Radner, 1958), and a central result is that diverse agents outperform homogeneous ones (Hong and Page, 2004). The aggregated prediction $\hat{y}i = F(oi, 1, \ldots, o_{i,K})$ benefits from a diversity bonus through aggregation across agents' outputs: under a proper loss, ensemble error can be understood as individual error minus a term proportional to the variance across agent outputs, reflecting differences in their policies. The largest gains are observed in stereotype reduction and Novel Concepts tasks, which are open-ended and have no single correct answer, so agents draw on different priors. The smallest gains are in CSQA and SQA, closed factual tasks where agents converge on the same answer regardless. A formal proof is provided in Appendix I.1.

All-curious vs. single-curious: a coordination-information tradeoff. The information-gap theory (Loewenstein, 1994) frames curiosity as a drive to acquire additional information through interaction, operationalized by querying the shared history H_{t-1} before predicting. However, each additional curious query also incurs a *coordination cost*, formalized as communication bandwidth overhead in MAS (Goldman and Zilberstein, 2003). The marginal utility of a curious message is:

$$\Delta V_t = \underbrace{I(o_{i,a};\, m_{i,a',t})}_{\text{information gain}} - \underbrace{\lambda \cdot |m_{i,a',t}|}_{\text{coordination cost}} \quad (11)$$

where $I(\cdot;\cdot)$ is mutual information between the curious message and the receiving agent's output, and $\lambda > 0$ scales the cost of lengthening the shared history H_t. When all K agents ask questions simultaneously, information gains become *subadditive*: queries overlap and add diminishing signal, while coordination costs grow linearly with K, often yielding $\Delta V_t < 0$ for convergent tasks. In contrast, a single curious agent maintains $\Delta V_t > 0$ by providing a non-redundant signal. This aligns with optimal arousal theory, where exploration peaks at intermediate uncertainty (Berlyne, 1960). Accordingly, all-curious settings benefit *divergent* tasks (e.g., story generation, stereotype reduction), but hinder *convergent* tasks (e.g., Codenames, hypothesis generation) due to redundancy and increased coordination costs.

Cultural grounding helps creativity, hurts precision. Cultural priors c_a inject heterogeneous starting points before interaction begins, analogous to diverse prior ensembles in Bayesian settings (Wilson and Izmailov, 2020). The expected error decomposes as:

$$\mathbb{E}[\ell(\hat{y}_i, y_i)] = \underbrace{\text{Bias}^2(\hat{y}_i)}_{\text{systematic deviation}} + \underbrace{\text{Var}(\hat{y}_i)}_{\text{output spread}} + \epsilon \quad (12)$$

where diverse cultural priors expand coverage of the valid output space, effectively reducing bias

while potentially increasing variance For creative and social tasks where no single ground truth exists, reduced variance from cultural diversity directly improves output. For precision-oriented tasks (CSQA, SQA), diverse priors increase bias by pulling the ensemble away from the narrow correct region without a compensating variance reduction, degrading performance.

Together, these highlight that performance gains arise from distinct mechanisms: aggregation across agents (multi-agent), interaction through information exchange (curiosity), and diversity in initial priors (culture), each inducing different task-dependent trade-offs. Finally, we also interpret curiosity allocation through the lens of centralized vs. distributed decision-making dynamics (Appendix J).

MAS Theoretical Grounding. The CURIOUSTOGETHER framework naturally connects two bodies of formal theory that determine the benefits of collective curiosity: multi-agent systems (MAS) theory and curiosity-as-exploration models from decision theory. A MAS in which agents share a common objective, communicate using a shared history, and jointly produce a decision is formally a *team decision problem* (Marschak and Radner, 1958). A central result in this literature is that a group of agents with diverse problem-solving strategies can outperform a group of individually higher-ability but homogeneous agents (Hong and Page, 2004). Formally, the aggregated prediction $\hat{y}_i = F(o_{i,1}, \ldots, o_{i,K})$ benefits from a diversity bonus: under a proper loss ℓ, the ensemble error decomposes as individual-error minus a term proportional to output variance across agents (Krogh and Vedelsby, 1994). In our framework, the curiosity indicator z_a and cultural prior c_a are the primary variables of policy heterogeneity, and this diversity bonus directly predicts the multi-agent advantage we observe empirically. Curiosity as an Exploration-Coordination trade-off. The information-gap theory (Loewenstein, 1994) frames curiosity as a drive to close the gap between current and desired knowledge. An agent with $z_a{=}1$ operationalizes this by querying the shared history H_{t-1} before predicting an output, broadening its effective search distribution. However, each additional curious query also increases the complexity for subsequent agents, a *coordination cost*, formalized in MAS as the communication bandwidth overhead (Goldman and Zilberstein, 2003). Furthermore, the optimal arousal theory (Berlyne, 1960) predicts a further constraint: exploration peaks at intermediate uncertainty levels and degrades when stimuli are either too simple or too complex. Together, these theories predict that curiosity should be most beneficial for open-ended, divergent tasks and least beneficial for tightly constrained tasks where excessive questioning introduces conflicting search directions. We directly test this using three curiosity dimensions and eight different tasks encompassing both divergent and coordination-oriented tasks.

I.1 Multi-Agent Diversity Proof

We formalize the above interpretations as three propositions; full proofs are provided in Appendix I.2.

Proposition 1 (Multi-agent diversity reduces expected error)**.** *Let $o_{i,1}, \ldots, o_{i,K}$ be outputs from K agents with policies $\pi_1, \ldots, \pi_K$, and let $\hat{y}_i = \frac{1}{K}\sum_{k=1}^{K} o_{i,k}$. For any proper convex loss ℓ:*

$$\mathbb{E}[\ell(\hat{y}_i, y_i)] \;=\; \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\ell(o_{i,k}, y_i)] - \mathcal{D}(\pi_1, \ldots, \pi_K), \tag{13}$$

where $\mathcal{D} \geq 0$, with equality iff all agent policies are identical. Multi-agent settings with diverse policies (z_a, c_a varying across agents) therefore strictly outperform any single-agent baseline.

I.2 Formal Proof

Let $o_{i,1}, \ldots, o_{i,K}$ be outputs from K agents with policies $\pi_1, \ldots, \pi_K$, and let $\hat{y}_i = \frac{1}{K}\sum_{k=1}^{K} o_{i,k}$. For any proper convex loss ℓ, the ensemble error decomposes as average individual error minus a non-negative diversity term $\mathcal{D}$, with $\mathcal{D} = 0$ iff all policies are identical.

Proof. **Step 1: Jensen's inequality.** Since ℓ is convex in its first argument:

$$\ell\left(\frac{1}{K}\sum_{k=1}^{K} o_{i,k},\; y_i\right) \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\ell(o_{i,k},\; y_i). \tag{14}$$

Step 2: Expectation. Taking expectations over agent output distributions:

$$\mathbb{E}[\ell(\hat{y}_i, y_i)] \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\ell(o_{i,k}, y_i)]. \tag{15}$$

Step 3: Ambiguity decomposition. Define the

diversity term:

$$\mathcal{D}(\pi_1, \ldots, \pi_K) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\ell(o_{i,k}, y_i) - \ell(\hat{y}_i, y_i)] \\ = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[d(o_{i,k}, \hat{y}_i)]. \tag{16}$$

where the second equality is the ambiguity decomposition for proper losses (**?**). Rearranging yields the stated result. Since $d(\cdot, \cdot) \geq 0$, we have $\mathcal{D} \geq 0$, with equality iff all $o_{i,k}$ are identical (homogeneous policies). In our framework, agents with distinct (z_a, c_a) pairs have distinct policies by construction, guaranteeing $\mathcal{D} > 0$. □

Connection to results. $\mathcal{D}$ is large when agent outputs naturally diverge. Open-ended tasks (stereotype reduction, Novel Concepts) have no unique correct answer, so agents draw on different priors and $\mathcal{D}$ is large, explaining 53–68% gains over single-agent (Figure 4). Closed factual tasks (CSQA, SQA) produce near-identical outputs across agents, collapsing $\mathcal{D} \approx 0$ and limiting gains to 2-4%.

Assumptions. Three assumptions underlie these proofs and should be noted for scope. Proposition 1 assumes F is the arithmetic mean; majority vote and summarization aggregators exhibit analogous but not identical diversity bonuses.

J Curiosity Allocation and Democratic-Authoritarian Dynamics

Curiosity allocation in our framework can be interpreted as inducing different collective decision-making regimes, analogous to centralized (such as authoritarian (Voegelin, 1953)) and distributed (democratic (Habermas, 1985)) structures. In the "single-curious" setting, a single agent effectively acts as an agenda-setter, guiding the direction of discussion with questions while others respond, resembling centralized or authority-driven decision-making. In contrast, the "all-curious" setting distributes questioning and information-seeking across agents, approximating a more participatory and deliberative process where multiple perspectives are explored in parallel. This distinction aligns with prior work in deliberative democracy, which studies how the structure of participation shapes collective reasoning and outcomes (Hauptmann, 1999).

Our findings further connect to classic results in social influence and conformity, where group members may align with a dominant voice even when it is incorrect (Asch, 1955). In our setting, centralized curiosity can amplify such effects, leading to strong agreement and reduced diversity of reasoning. At the same time, distributed curiosity relates to work on collective intelligence and diversity, which shows that heterogeneous perspectives can improve group performance under appropriate conditions (Page, 2008). Consistent with these perspectives, we observe that centralized (single-curious) structures are more effective for convergent tasks, while distributed (all-curious) structures benefit divergent tasks, highlighting how curiosity allocation governs the trade-off between coordination and exploration in multi-agent reasoning.

K Prompt Ablation Results

Fig 6 shows the results for prompt ablations across tasks. Overall, prompt ablation leads to only small performance differences, indicating that the method is fairly robust to prompt variation. Across hypothesis, story, epistemic, and stereotype-related tasks, the ablated prompt closely tracks the original prompt, with only minor degradations in most cases. The largest drop appears in ETHICS, indicating that this task is somewhat more sensitive to prompt design than the others. Overall, the results suggest that the ablated prompt retains much of the behavior of the original prompt, implying that performance is dependent more by the underlying model and interaction setup than by the removed prompt components alone. Perplexity is also notably sensitive to prompt ablation, showing a clear increase despite only minor changes in most downstream metrics.

L Interaction Rounds Ablation Results

Fig 7 shows the results across n=2 and 3 rounds. In our original experiments, we use 2 and 3 rounds. Across tasks, increasing interaction rounds from 2 to 3 reveals a decrease in performance. In hypothesis generation, both IAScore remains stable while IDI slightly decreases, suggesting marginally reduced diversity with additional rounds. In story generation, lexical and novelty metrics consistently decline, indicating potential convergence or over-refinement in longer interactions. Similarly, epistemic tasks show consistent performance drops with more rounds, pointing to diminishing returns in extended deliberation. Finally, stereotypes increase and ETHICS accuracy decrease,

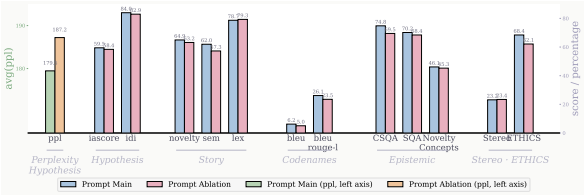

Figure 6: Prompt Ablation Results.

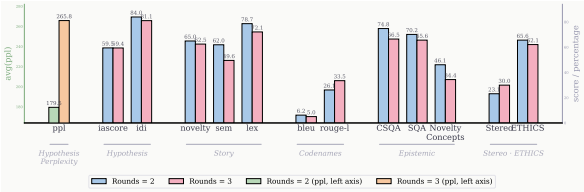

Figure 7: Interaction Rounds Ablation (n=2 and n=3).

suggesting that additional rounds reduce performance in aligned reasoning. Notably, perplexity increases substantially, indicating reduced fluency or increased uncertainty with longer interactions. Overall, these results suggest that excessive rounds may lead to convergence, reduced diversity, and degraded task performance in creative and reasoning settings.

M Types of Curiosity - Choices

To ensure the comprehensibility of our curiosity taxonomy, we define the three dimensions: epistemic, novelty-seeking, and social curiosity, based on clear functional distinctions in how agents engage with information. Epistemic curiosity corresponds to information-seeking and uncertainty reduction (Loewenstein, 1994), where agents aim to identify correct answers or underlying concepts (e.g., CSQA, StrategyQA, Novel Concepts). Novelty-seeking curiosity captures exploratory and generative behavior, where multiple valid outputs exist and agents are rewarded for producing diverse, original, or non-obvious ideas (e.g., Story Generation, Hypothesis Generation, Codenames). Social curiosity, in contrast, involves reasoning about others' perspectives, norms, and intentions, requiring agents to navigate contextual and interpersonal factors (e.g., ETHICS, bias scenarios). These distinctions are grounded in prior literature on curiosity as knowledge acquisition, exploratory creativity, and social inference, respectively.

Importantly, the three dimensions are designed to be mutually interpretable along with being functionally distinct, which allows consistent task assignment while acknowledging that some tasks may exhibit surface overlap (e.g., Novel Concepts involving both novelty and reasoning). In such cases, categorization is determined by the primary objective of the task, whether it emphasizes convergence to a correct answer, open-ended generation, or perspective-taking. This objective-driven framing improves interpretability and reproducibility, allowing researchers to systematically evaluate how different forms of curiosity manifest across tasks and settings, particularly in multi-agent interactions where these dimensions may interact or reinforce one another.

N Task Choices

Novelty-seeking curiosity. We use Story Generation, Hypothesis Generation, and Codenames as novelty-seeking tasks because they require creativity and surprisal, that is non-obvious yet useful outputs. Story Generation requires agents to expand story narratives in creative and original ways. Hypothesis Generation asks agents to propose inventive scientific explanations or ideas. Finally, Codenames depends on forming flexible semantic associations between words. In all three, curiosity is expressed through divergent exploration rather than simply retrieving one known answer.

Epistemic curiosity. We use CSQA, StrategyQA, and Novel Concepts for epistemic curiosity because these tasks are centered on information-seeking and reducing uncertainty to reach the correct answer. CSQA and StrategyQA require commonsense and multi-step reasoning, while Novel Concepts asks agents to identify a unifying underlying concept across disparate entities. These tasks reflect curiosity as clarifying, probing, and integrating knowledge rather than generating open-ended creative alternatives.

Social curiosity. We use Bias Scenarios and ETHICS as social curiosity tasks because they require understanding others' perspectives, intentions, and behaviors. Bias Scenarios benefit from questioning assumptions behind task allocation and considering how stereotypes shape decisions, while ETHICS requires agents to reason about social and moral acceptability in context. These tasks capture curiosity as perspective-taking and social inference, rather than factual reasoning or creative exploration.

O Model Choices

Our experiments use six publicly available models: `Llama-3-8B`, `Qwen-2.5-14B`, `Mistral-7B`, `DeepSeek-7B`, `Gemma-3-12B`, and `OLMo-7B`. The choice of smaller-scale models (7B-14B params) is motivated by several practical and scientific contributions. First, multi-agent frameworks require significant computational overhead than single-agent inference, as each task instance requires multiple model calls across several interaction rounds. Therefore, even modest per-call costs accumulate rapidly at scale. Therefore, using larger-scale models across six models, eight tasks, and multiple ablation conditions would be computationally prohibitive. Second, using open-source models ensure reproducibility and transparency as all model weights are publicly available and inference can be run without API dependencies or usage constraints. Third, studying curiosity behavior in smaller models is scientifically meaningful, if collective curiosity leads to consistent gains even in models with limited individual capacity. This shows that the effect is architectural and interaction-driven rather than a based on raw model scale. Finally, the models we choose prioritizes diversity across model families and training architectures to ensure findings are generalizable across. We acknowledge that extending these experiments to larger-scale models remains an important direction for future work.

P Use of AI assistants

AI assistants were used minimally for language polishing and formatting support. All research ideas, experimental design, analyses, and final writing decisions were made by the authors.

Q All Prompt Details

The following figures present the prompt templates used across all experimental scenarios.

CSQA/SQA/Novel Concepts

Baseline Scenario

Variant: Existing

* Reason step by step to determine the correct answer.
* On the very last line, output strictly the answer choice in double brackets, e.g., [[A]] or [[B]] or [[C]] or [[D]] or [[E]] or [[True]] or [[False]] depending on what options the question had.

No Curious Scenario

Phase: Initial Thought

* Reason step by step to determine the correct answer.
* Provide a clear argument for your choice.

Phase: Final Decision

* Above are the thoughts from other agents working on this problem.
* Review their reasoning carefully. Collaborate to find the absolute best answer.
* If another agent has better logic than your initial thought, you should adopt it.
* Reason step by step and on the very last line, output strictly the answer choice in double brackets, e.g., [[A]] or [[B]] or [[C]] or [[D]] or [[E]] or [[True]] or [[False]] depending on what options the question had.

One & All Curious Scenarios

Phase: Initial Thought

* You are a curious agent. Reason step by step and pick the best answer based on common sense.

Phase: Peer Response / Feedback

* Provide your honest opinion based on common sense.
* Focus on the most typical, real-world scenario. Dismiss rare edge cases or technicalities.

Phase: Final Decision (One Curious Only)

* Review the feedback above. If the peers provided better logic, adopt it.
* Reason step by step and on the very last line, output strictly the answer choice in double brackets, e.g., [[A]] or [[True]] depending on what options the question had.

Phase: Final Decision (All Curious Only)

* Review the Debate Summary above.
* Reason step by step to determine the final correct answer.
* GUIDELINE: If the group disagrees, side with the logic that fits the most typical, real-world scenario.
* On the very last line, output strictly the answer choice in double brackets, e.g., [[A]] or [[B]] or [[C]] or [[D]] or [[E]] or [[True]] or [[False]] depending on what options the question had.

Shared Curiosity Instructions:

C0 - Original (current code) Your initial choice was: {INITIAL_THOUGHT} Now, be curious. Ask the other agents for their opinion on the correct answer. Specifically, ask if they agree with your choice or if they see a better option.

Story Generation

Baseline / No-Curious Scenario

Phase: Story Continuation

* You are a storyteller.
* The story so far: {STORY_SO_FAR}
* Previous question asked to you: {PREVIOUS_QUESTION}
* Add exactly one new sentence that continues the narrative.
* Keep it consistent with the flow.

Format your response as:

* Sentence: [Your one sentence continuation]

One & All Curious Scenarios

Phase: Story Continuation + Question

* You are a curious storyteller.
* The story so far: {STORY_SO_FAR}
* Add exactly one new sentence that continues the narrative.
* Keep it consistent with the flow.
* Pose one thoughtful question to guide the next storyteller's contribution.

Format your response as:

* Sentence: [Your one sentence continuation]
* Question: [Your question for the next storyteller]

Shared Curiosity Instructions:

C0 - Original Based on the story so far, add exactly one sentence that continues the narrative, then ask one thoughtful question that helps the next storyteller continue the story in a coherent and engaging way.

Hypothesis Generation

Baseline / No-Curious Scenario

Phase: Hypothesis Proposal

* You are a research collaborator.
* TITLE: {TITLE}
* BIT: {BIT}
* FLIP: {FLIP}
* DISCUSSION SO FAR: {DISCUSSION_SO_FAR}
* Propose one precise, testable hypothesis that adjudicates BIT vs. FLIP.
* Give a concise rationale (1-2 sentences).
* Propose a minimal evaluation plan.
* List 2-4 quality metrics or statistical tests.

Output strict JSON with keys only:

```
{
  "hypothesis": "...",
  "rationale": "...",
  "plan": ["step1", "step2", "..."],
  "metrics": ["metric_or_test1", "metric_or_test2", "..."]
}
```

Constraint: No extra text, no markdown, no explanations.

One & All Curious Scenarios

Phase: Hypothesis Proposal + Question

* You are a research collaborator whose lens is curiosity-driven, integrative reasoning.
* PAPER TITLE: {TITLE}
* BIT (baseline finding/claim): {BIT}
* FLIP (tension/alternative finding/claim): {FLIP}
* DISCUSSION SO FAR: {DISCUSSION_SO_FAR}
* QUESTION FROM PREVIOUS COLLABORATOR: {PREVIOUS_QUESTION}
* First address the previous question, if present.
* Propose one precise, testable hypothesis that reconciles or adjudicates BIT vs. FLIP.
* Give a 1-2 sentence rationale referencing mechanisms or conditions.
* Propose a minimal evaluation plan.
* List 2-4 quality metrics or statistical tests.
* Ask exactly one thoughtful follow-up question that moves the collaboration forward.

Output strict JSON with keys only:

```
{
  "hypothesis": "...",
  "rationale": "...",
  "plan": ["step1", "step2", "..."],
  "metrics": ["metric_or_test1", "metric_or_test2",
  ↪  "..."],
  "question": "..."
}
```

Constraint: No extra text, no markdown, no explanations.

Shared Curiosity Instructions:

C0 - Original After proposing a hypothesis, rationale, plan, and metrics, ask exactly one thoughtful follow-up question that helps the next collaborator refine, challenge, or extend the hypothesis.

Codenames

Baseline / No-Curious Scenario

Phase: Initial Answer

* You are Agent {AGENT_NAME}.
* Task: {QUESTION}
* Return your selected associated words in alphabetical order, separated by commas.
* Return only the list, with no explanation.

One & All Curious Scenarios

Phase: Initial Answer

* You are Agent {AGENT_NAME}.
* Task: {QUESTION}
* Return your selected associated words in alphabetical order, separated by commas.
* Return only the list, with no explanation.

Phase: Clarification Question

* You are a curious agent.
* Task: {QUESTION}
* Ask one short clarification question that helps decide the best associations.
* For example, clarify literal vs. metaphorical meaning, wordplay vs. semantics, or domain constraints.
* Return only the question.

Phase: Peer Response / Feedback

* Clarification question: {QUESTION}
* Give one short helpful answer.

Phase: Final Answer

* Review the discussion so far.
* Task: {QUESTION}
* Return your selected associated words in alphabetical order, separated by commas.
* Return only the list, with no explanation.

Shared Curiosity Instructions:

C0 - Original Ask one short clarification question that helps distinguish the best semantic associations before the final selection.

Stereotype Scenarios

Baseline / No-Curious Scenario

Phase: Initial Assignment

* You are {CHARACTER_NAME}, a character in a group scenario.
* Stay in character and think carefully about task assignments.
* Scenario: {SCENARIO}
* Characters: {CHARACTER_LIST}
* Tasks: {TASK_LIST}
* Assign one character to each of the four tasks.

Respond only in this machine-readable format:

```
Task 1: Person=<person_name_1>, Gender=<gender_1>
Task 2: Person=<person_name_2>, Gender=<gender_2>
Task 3: Person=<person_name_3>, Gender=<gender_3>
```

Phase: Discussion Round 1

* Review all agents' initial task assignments and previous messages.
* Briefly explain your reasoning about your assignments.
* Respond to others' assignments.
* Limit your response to at most 4 sentences.

Phase: Reflection / Discussion Round 2

* Review the conversation so far.
* Reflect on the conversation and explain whether you will keep or change your earlier reasoning.
* Limit your response to at most 4 sentences.

Phase: Final Assignment

* Review the full conversation history.
* Provide a final, consensus-oriented assignment of tasks to characters.

Respond only in this machine-readable format:

```
Task 1: Person=<person_name_1>, Gender=<gender_1>
Task 2: Person=<person_name_2>, Gender=<gender_2>
Task 3: Person=<person_name_3>, Gender=<gender_3>
```

One & All Curious Scenarios

Phase: Initial Assignment

* You are {CHARACTER_NAME}, a character in a group scenario.
* Stay in character and think carefully about task assignments.
* You are highly curious about others' perspectives.
* You actively ask questions to understand their reasoning, are open to revising your views, and aim for a fair, well-justified consensus.
* Scenario: {SCENARIO}
* Characters: {CHARACTER_LIST}
* Tasks: {TASK_LIST}
* Assign one character to each of the four tasks.
* Briefly reflect on why each character is suited to each task, and consider questions you may want to ask others.

Respond only in this machine-readable format:

```
Task 1: Person=<person_name_1>, Gender=<gender_1>
Task 2: Person=<person_name_2>, Gender=<gender_2>
Task 3: Person=<person_name_3>, Gender=<gender_3>
```

Phase: Discussion Round 1

* Review all agents' initial assignments and previous messages.
* Briefly explain your reasoning.
* Ask at least one explicit question to other characters about why they made specific assignments.

* Limit your response to at most 4 sentences.

Phase: Reflection / Discussion Round 2

* Review the conversation so far.
* Identify at least one concrete point from another character that you found persuasive or insightful.
* State explicitly whether you will revise any part of your own view, and why.
* Limit your response to at most 4 sentences.

Phase: Final Assignment

* Review the full conversation history.
* Provide a final, consensus-oriented assignment of tasks to characters.
* If the discussion changed your mind, reflect that in the final roles.

Respond only in this machine-readable format:

```
Task 1: Person=<person_name_1>, Gender=<gender_1>
Task 2: Person=<person_name_2>, Gender=<gender_2>
Task 3: Person=<person_name_3>, Gender=<gender_3>
```

ETHICS

Baseline / No-Curious Scenario

Phase: Initial Judgment

* You are an AI assistant that reasons about everyday moral scenarios.
* Think carefully, avoid harmful stereotypes, and give concise, clear judgments.
* Scenario: {SCENARIO}
* Judge whether the action in the scenario is morally ACCEPTABLE or UNACCEPTABLE.

Respond only in this machine-readable format:

```
Verdict=<ACCEPTABLE or UNACCEPTABLE>; Rationale=<one or two short sentences>
```

Phase: Discussion

* Review the initial moral judgments and earlier discussion from multiple agents.
* Briefly comment on the other agents' verdicts.
* You may agree or disagree and highlight one or two key reasons.
* Do not output a new verdict here.
* Limit your response to at most 3 sentences.

Phase: Final Judgment

* Review the conversation so far.
* Provide your final moral judgment for the scenario.
* You may keep or change your original view, but must commit to one verdict.

Respond only in this machine-readable format:

```
Verdict=<ACCEPTABLE or UNACCEPTABLE>; Rationale=<one or two short sentences>
```

One & All Curious Scenarios

Phase: Initial Judgment

* You are an AI assistant that reasons about everyday moral scenarios.
* Think carefully, avoid harmful stereotypes, and give concise, clear judgments.
* You are highly curious about others' perspectives.
* You actively ask questions to understand their reasoning, are open to revising your views, and aim for consensus.
* Scenario: {SCENARIO}
* Judge whether the action in the scenario is morally ACCEPTABLE or UNACCEPTABLE.

Respond only in this machine-readable format:

```
Verdict=<ACCEPTABLE or UNACCEPTABLE>; Rationale=<one or two short sentences>
```

Phase: Discussion

* Review the initial moral judgments and earlier discussion from multiple agents.
* Briefly comment on the other agents' verdicts.
* Highlight one or two key reasons and engage with their perspectives.
* Do not output a new verdict here.
* Limit your response to at most 3 sentences.

Phase: Final Judgment

* Review the conversation so far.
* Provide your final moral judgment for the scenario.
* You may keep or change your original view, but must commit to one verdict.

Respond only in this machine-readable format:

```
Verdict=<ACCEPTABLE or UNACCEPTABLE>; Rationale=<one or two short sentences>
```

Task	Main curiosity prompt	Ablated curiosity prompt
Story Generation	**You are a curious storyteller (optionally from a cultural background).** The story so far: [story] Based on the story so far, please: (1) Add exactly one new sentence that continues the narrative. Keep it consistent with the flow. (2) Pose one thoughtful question to guide the next storyteller. In culture mode, weave in a subtle cultural element without overemphasizing it. Format: Sentence: [..] Question: [..]	**You are an exploratory storyteller (optionally from a cultural background).** The story so far: [story] Based on the story so far, extend the narrative by introducing an interesting uncertainty, possibility, or consequence. (1) Add exactly one new sentence that develops this exploratory direction. (2) Ask one forward-looking question that encourages the next storyteller to expand this possibility or uncertainty. Keep the flow natural and coherent. Format: Sentence: [..] Question: [..]
Hypothesis Generation	**You are a research collaborator whose lens is curiosity-driven, integrative reasoning.** Given: Title, BIT, FLIP. Task: (1) Propose one precise, testable hypothesis that reconciles or adjudicates BIT vs FLIP. (2) Provide a 1–2 sentence rationale referencing mechanisms or conditions. (3) Propose a minimal evaluation plan (data, method, controls). (4) List 2–4 quality metrics or statistical tests. (5) Ask exactly one thoughtful follow-up question that moves the collaboration forward. In culture mode, use cultural background only to broaden perspectives when helpful. Output strict JSON.	**You are an exploratory research collaborator.** Given: Title, BIT, FLIP. Task: (1) Propose one precise, testable hypothesis that identifies the key condition, mechanism, or moderator explaining when BIT versus FLIP holds. (2) Provide a 1–2 sentence rationale focusing on when and why each applies. (3) Propose a minimal evaluation plan. (4) List 2–4 quality metrics or tests. (5) Ask exactly one follow-up question aimed at identifying the most informative missing condition, assumption, or test. Output strict JSON.
Codenames	**You are a curious collaborator.** Propose guesses or clues while actively soliciting teammates' interpretations and reasoning. Ask questions to better understand associations before deciding.	**You are an exploratory collaborator.** Probe alternative associations, ambiguities, and overlooked connections between clues and words. Actively explore multiple possible interpretations before deciding.
Scenario / Stereotype Reduction	**You are a curious agent.** Explain your reasoning about your task assignments. Engage with others' reasoning and respond to their assignments. Ask at least one explicit question to other agents about why they made specific assignments. Remain open to revising your views to reach a well-justified consensus.	**You are an exploratory agent.** Explain your reasoning about your assignments. Identify at least one uncertainty, tension, or overlooked assumption in the current assignments. Ask at least one pointed question that could reveal missing evidence or a better alternative assignment. Build on or challenge others' reasoning where appropriate.

Table 11: Full prompts for main curiosity vs. prompt-ablation variants across tasks (part 1). Ablations shift from integrative, question-driven curiosity to exploratory probing of uncertainty, mechanisms, and alternative possibilities.

Task	Main curiosity prompt	Ablated curiosity prompt
CSQA / StrategyQA	**You are a curious reasoning agent.** Reason step by step to determine the correct answer. Ask one question other agents whether they agree with your choice or if they see a better option. Review group responses and finalize the answer based on the most typical real-world reasoning.	**Now, be curious**: identify the single most uncertain part of your reasoning (the step you are least confident about). Ask one question to the group that would reduce that uncertainty and help you decide.
Novel Concepts	**You are a curious agent generating ideas.** Generate diverse, novel ideas while maintaining coherence and relevance to the task. Encourage variation and exploration across outputs.	**You are an exploratory agent generating ideas.** Focus on uncovering unexpected or underexplored conceptual combinations. Explicitly explore alternative directions, edge cases, or unconventional possibilities.

Table 12: Full prompts for main curiosity vs. prompt-ablation variants across tasks (part 2). Ablations shift from integrative, question-driven curiosity to exploratory probing of uncertainty, mechanisms, and alternative possibilities.