Human Don't need to Know the Answer to Help: Intuition Steers Disagreement in Multi-Agent LLMs

Yaqi Xie **Yujia Zheng**
Carnegie Mellon University
{yaqixie, yujiazh}@cmu.edu

Abstract

This position paper argues that even when humans lack the correct answer or problem-solving expertise, their intuitive judgments can still meaningfully improve the performance of multi-agent LLMs. Collaboratively leveraging multiple LLMs has emerged as an effective strategy to enhance problem solving capabilities by utilizing complementary specializations and enabling mutual verification among agents. However, when disagreements arise, agents following the correct reasoning paths can be misled or overwhelmed by the incorrect ones, resulting in degraded final answers. We show that human feedback, when focused on agent disagreements and presented as simplified binary choices through LLM-generated summaries, even from non-experts, can effectively steer collaborative debates toward more accurate outcomes. Drawing on insights from cognitive science and collective intelligence, we demonstrate that human intuition, despite being uninformed, can provide low-cost, high-impact guidance at inference time. This challenges the prevailing assumption that useful feedback must come from experts, and offers a practical, scalable mechanism for integrating human input into multi-agent AI systems.

1 Introduction

Collaborative reasoning among large language models (LLMs) has emerged as a powerful paradigm: by coordinating multiple agents with diverse reasoning paths, systems can often converge on better answers than any individual model could achieve alone. This multi-agent framework amplifies both capability and robustness through cross-verification, debate, and iterative refinement. Yet, it also introduces a critical vulnerability: when agents disagree, especially on complex or ambiguous tasks, correct reasoning can be drowned out by confident but incorrect voices. Extended debate alone does not guarantee convergence to the right answer.

Single-agent human-in-the-loop systems traditionally involve explicit human feedback on direct model outputs, often via corrective labels, rankings, or explicit corrections. Such methods typically assume humans possess accurate domain knowledge or definitive answers and apply direct corrections or reinforcements accordingly [Christiano et al., 2017, Ouyang et al., 2022]. However, these single-agent methods rely heavily on human expertise and tend to falter in scenarios where human knowledge is incomplete or uncertain. Multi-agent human-in-the-loop approaches, however, remain largely unexplored despite their potential advantages. Unlike single-agent systems, multi-agent frameworks offer richer and more structured contexts for intervention—particularly in cases of disagreement. The critical difference lies in shifting the human role from authoritative solver to intuitive adjudicator, substantially reducing the knowledge burden on humans.

We draw on established cognitive science findings: humans are adept at recognizing coherence, plausibility, and argumentative quality even in unfamiliar domains [Gigerenzer and Gaissmaier, 2011, Evans, 2008]. Furthermore, collective intelligence research shows that aggregating intuitive judgments across non-experts often outperforms individual expert decisions [Woolley et al., 2010]. We build on

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

these insights to propose a practical, inference-time human-in-the-loop mechanism. By strategically incorporating this intuitive feedback into subsequent debate rounds, we efficiently steer multi-agent discussions toward more accurate and coherent resolutions without requiring extensive human expertise or expensive fine-tuning. Instead of asking humans to solve the problem, we ask them to choose which line of reasoning makes more sense: an easier, more reliable, and still effective intervention.

This position paper argues that even when humans do not know the correct answer, their intuitive feedback on model disagreements can still meaningfully improve multi-agent LLM performance. Our central claim is that intuitive human feedback, focused not on answering the question directly but on adjudicating structured disagreements, can steer agent discussions toward more accurate results. We operationalize this by using an auxiliary LLM to summarize disagreements into binary choices, allowing humans to select the more reasonable or convincing option. This feedback is then used to guide the next round of debate.

By integrating intuitive human judgment precisely where agents disagree, our method enhances multi-agent performance with minimal cost. It challenges the common belief that humans must know the answer to be helpful–showing instead that intuitive human feedback can guide agent debates simply by resolving disagreements and leading to better outcomes. To make this intervention efficient, we analyze when human feedback is most valuable. Rather than inserting humans indiscriminately into the loop, we focus on moments of disagreement, when agent trajectories diverge. These are the points of highest epistemic uncertainty and, therefore, the greatest opportunity for steering. Our system identifies such disagreements and packages them as binary or multiple-choice options, allowing humans to act as discriminators of plausibility rather than solvers of the task.

*A **key insight** behind our method is that we do not require or assume human participants to be domain experts or capable of directly solving the problem.* Making this assumption would significantly restrict the accessibility and scalability of human-in-the-loop systems, limiting their practical deployment. By relaxing this requirement, we broaden participation to a wider population, enabling the use of readily available intuitive human judgment. This approach not only lowers the barrier to entry but also enhances the robustness of the system by incorporating diverse perspectives, thus improving generalizability across various domains and types of disagreements. Human intuition, even without explicit domain knowledge, remains remarkably effective. This effectiveness likely arises from human inductive biases. These biases often favor simplicity, causality, and coherence, which align well with the structure of many real-world tasks. For example, when presented with competing scientific explanations, humans tend to prefer those that are more direct and require fewer assumptions, a reflection of simplicity bias (Occam's Razor). Similarly, explanations that offer causal mechanisms are preferred over purely associative ones. Many disagreements encountered by multi-agent systems involve reasoning paths or abstractions derived from human common sense or natural-world intuitions, inherently aligning with these various human biases, which we explore in greater detail in subsequent sections.

Why do humans exhibit such inductive biases, while trained models do not naturally encode them? Unlike LLMs, which primarily learn from static datasets like text, images, or videos, humans interact dynamically with the physical world. This active engagement, through intervention and experimentation, ingrains deep intuitive heuristics that guide judgments and decisions even in unfamiliar contexts. Unlike foundational models that primarily learn from passive observational data, humans actively manipulate and interact with their environment, allowing them to directly observe causal effects. Additionally, humans experience learning through multiple modalities, such as visual, auditory, tactile, and proprioceptive senses, which contribute to richer, more robust representations. Humans interact with the physical world, receive feedback from actions, and build grounded abstractions through sensory-motor experience. These interactions give rise to stable inductive biases that shape how humans generalize and evaluate explanations. Furthermore, certain human intuitions and biases are genetically inherited, refined over approximately 300,000 years of evolutionary history. This lifelong and evolutionary scale of learning far exceeds the comparatively brief and limited training processes of foundational models.

Human intuition is a collection of inductive biases, such as a preference for simplicity and causality, shaped by lifelong physical interaction, multimodal sensory input, and evolutionary adaptation. These biases enable effective judgment even in the absence of explicit domain knowledge.

While these inductive biases are difficult to quantify directly, their effects are evident and measurable. A large body of single-agent HITL research demonstrates that human feedback, when applied

correctly, can systematically improve model performance across tasks [Ouyang et al., 2022, Bai et al., 2022a]. These results indirectly validate the utility of human inductive priors. Yet, their potential remains untapped in multi-agent systems, where disagreement creates natural opportunities for human intervention to be most effective. Closing this gap could unlock substantial potential for more effective integration of human intuition into multi-agent reasoning processes, thereby improving the reliability and robustness of these collaborative systems.

By recognizing this gap, our work positions intuitive human feedback as a lightweight, inference-time mechanism for multi-agent steering. Rather than treat humans as oracle solvers, we treat them as inductive filters, amplifying signal over noise in collective model reasoning. This not only reduces the cognitive and time burden on users, but also plays to the strengths of human cognition: fast, intuitive judgment shaped by evolution and interaction with the world. By integrating intuitive human judgment precisely where agents disagree, our method enhances multi-agent performance with minimal cost. It challenges the common belief that humans must know the answer to be helpful, showing instead that intuitive human feedback can guide agent debates simply by resolving disagreements and leading to better outcomes.

Takeaway. Even without domain expertise, intuitive human judgment can effectively resolve disagreements among multi-agent LLMs, significantly improving system accuracy and robustness at minimal cost.

2 Alternative Views

Agent as Judge/Evaluator. A common approach in multi-agent LLM systems is to designate one model as a judge to evaluate and select among outputs generated by other agents. This is seen in AI debate frameworks, where two agents argue opposing views and a third agent, often another LLM, acts as the evaluator to determine which argument is more convincing [Irving et al., 2018]. Some systems extend this to courtroom-style setups, with roles like advocate, critic, and judge distributed among agents. Other frameworks use LLM judges to rank candidate outputs by scoring them on criteria such as factuality, logic, or clarity [Liang et al., 2024, Gu et al., 2024]. These designs aim to automate arbitration, reduce human workload, and scale evaluation through language models themselves.

However, this agent-as-judge paradigm carries inherent limitations. First, LLM judges often share the same architecture, training data, and inductive blind spots as the agents they evaluate, making them susceptible to the same errors and misconceptions. This undermines their role as independent arbiters. Second, LLM judges frequently rely on superficial heuristics—like fluency or confidence—leading them to favor persuasive but incorrect answers. Without grounding in external context or the ability to interrogate the problem interactively, AI judges can be misled in subtle ways, especially under ambiguity. Third, when debate participants omit critical assumptions or present competing yet flawed reasoning, the judge has no external reality to anchor its decision, resulting in arbitrary or misleading judgments [Liang et al., 2024].

Our proposed human-in-the-loop mechanism addresses these core weaknesses. Instead of relying on internal AI arbitration, we introduce disagreement-triggered human intervention, where intuitive human judgment is used precisely when agents diverge. This method benefits from human inductive biases, such as simplicity and causality preferences, which often align better with natural abstractions and real-world plausibility. Humans, even non-experts, possess grounded reasoning shaped by interaction with the physical world, enabling them to detect incoherence or implausibility in ways models cannot. Unlike LLM judges, humans do not share the same training-induced patterns of failure. Critically, we do not ask humans to solve the task but only to adjudicate between competing reasoning paths. This lightweight feedback, guided by intuitive plausibility rather than correctness, is more robust in uncertain or complex settings. By using human intuition where models disagree, our approach enhances alignment, resilience, and trustworthiness, without requiring expert knowledge or expensive retraining. In short, while the agent-as-judge model offers scalability, our method offers robust arbitration grounded in human cognitive strengths.

Deferral Mechanisms. Deferral methods determine when an LLM should hand off a query to a human expert. Montreuil et al. [2024] introduce a two-stage "learning-to-defer" model for extractive QA, where a small model selectively passes ambiguous questions to a human or larger model. This yields accuracy gains: deferring only a small fraction of queries allows the small model to match the performance of a much larger one. Similarly, Strong et al. [2024] propose a guided

deferral system for clinical decision support, where the LLM not only defers difficult cases but also provides task-based guidance to the clinician. These models typically rely on fixed criteria trained for specific domains. By contrast, our method dynamically identifies contentious moments in multi-agent debates and simplifies decision contexts through summarization. Cognitive load theory suggests that decision quality improves significantly when cognitive demands on evaluators are minimized through simplified presentations of information [Sweller, 2011]. Our approach capitalizes on this cognitive principle by presenting human evaluators with streamlined, binary choices distilled from agent disagreements, maximizing intuitive judgment accuracy. Thus, by merging insights from cognitive psychology and collective intelligence with novel computational strategies, our approach positions intuitive human feedback as a valuable, efficient, and underutilized resource in multi-agent LLM debate frameworks.

3 Non-Expert Human Feedback Enhances Multi-Agent LLMs

3.1 Why Machines Need Human Help

While LLMs have achieved remarkable performance across various domains, their capabilities remain imperfect and inconsistent [Brown et al., 2020, Achiam et al., 2023]. Empirical studies show that even the most advanced models frequently produce incorrect or suboptimal outcomes due to inherent limitations, biases, or factual inaccuracies [Bubeck et al., 2023]. Human oversight and feedback have empirically proven effective in refining these outputs [Ouyang et al., 2022, Bai et al., 2022b], indicating the essential role of humans in the model improvement process. Human contributions are particularly valuable in situations requiring nuanced understanding, ethical considerations, and context-specific judgments where machine algorithms alone may fall short. Examples include refining alignment and reducing harmful outputs in large language models, where human evaluators provided feedback that significantly enhanced the models' ethical and factual consistency [Ouyang et al., 2022]. Human-in-the-loop feedback also enabled models to better capture nuanced human preferences, resulting in more accurate, contextually appropriate responses in conversational AI systems [Christiano et al., 2017]. Additionally, human guidance has successfully addressed complex tasks such as reinforcement learning-based game playing, where intuitive feedback substantially improved strategic decision-making and overall performance [Ibarz et al., 2018].

3.2 The Importance of Human Intuition

We define intuition as the type of cognition or judgment that cannot be fully replicated by maximizing memorized knowledge or pre-existing information. Unlike explicit knowledge that can be directly learned from extensive datasets, intuition is deeply rooted in cognitive processes evolved through natural selection and adapted through unique human perceptual experiences [Gigerenzer and Gaissmaier, 2011, Evans, 2008]. Human intuition is the rapid and unconscious formation of judgments and preferences grounded in embodied experience, developmental context, and social interaction, not in explicit instruction or symbolic representation. It arises from a lifelong integration of multisensory perception, motor activity, emotional feedback, and situated learning within the real world. Crucially, it depends on non-verbalizable knowledge and context-sensitive heuristics that cannot be captured through language or data alone. Because intuition is not merely a summary of past observations but a product of direct physical and social immersion, it is fundamentally inaccessible to models trained solely on textual or visual data.

Definition of human intuition. Human intuition is any form of judgment or decision-making that cannot be achieved by maximizing memorized knowledge or optimizing over learned data.

The Construction of Human Intuition. Human intuition arises from evolutionary biology and adaptive processes, allowing humans to efficiently handle uncertainty and incomplete information. Unlike machines that rely on predefined data distributions and structured learning paradigms, human intuition benefits from diverse, multimodal perceptions of the environment, integrating emotional, experiential, and sensory inputs into decision-making.

Human Intuition as Inductive Biases. Many forms of human intuition can be understood as inductive biases, systematic tendencies that guide judgment even in the absence of complete information. Examples include simplicity bias (preferring explanations with fewer assumptions), causal bias (favoring cause-effect explanations), and analogy-based reasoning. More examples can be found in Section 3.4. These intuitive heuristics allow humans to make rapid and often effective decisions in

complex, ambiguous environments. Unlike statistical learning, these biases are not explicitly learned from data but are deeply embedded in cognition, shaped by evolution, development, and social context.

Other types of intuition involve uncertainty that does not follow a fixed or universal pattern. These are harder to categorize as inductive biases because they vary across individuals and situations. Yet, this variability itself offers potential benefits for machines. Human uncertainty is not simply noise—it can reflect flexibility, emotional influence, or exploratory behavior that leads to robustness and serendipitous discovery. Machines, by contrast, generate uncertainty in rigid, predefined ways (e.g., Gaussian noise or dropout), lacking the adaptive unpredictability of human intuition. Understanding and leveraging this human-like uncertainty could enrich machine reasoning beyond current deterministic or randomized frameworks.

The Value of Human Intuition in the Age of LLMs. As LLMs continue to advance, it becomes increasingly clear that humans should not aim to compete with them on tasks that rely purely on recalling or recombining existing knowledge. Instead, we should focus on what remains uniquely human—intuition. Unlike rule-based reasoning or knowledge retrieval, intuition allows for quick, flexible judgment in ambiguous or novel situations. If LLMs were ever to possess genuine, human-like intuition, it might signal the arrival of artificial general intelligence, at which point human feedback could become obsolete. But until then, intuition remains an essential human contribution to collaborative intelligence.

Theoretical perspectives also underscore the importance of intuition. For example, developmental psychology shows that young children, despite limited formal knowledge, can perform surprisingly well in tasks involving physical reasoning, social interaction, or moral judgment [Spelke and Kinzler, 2007]. This suggests that intuition, rather than accumulated facts, often drives effective behavior. Furthermore, while explicit knowledge is constrained by how it is recorded and transmitted, often in incomplete, noisy, or biased ways, intuition does not require formal justification. A person may arrive at a decision simply because "it feels right," and this inexplicable but effective sense is often inaccessible to models trained on language or data alone.

From a practical standpoint, the value of human intuition is evident in systems that incorporate human feedback. Successful single-agent LLM frameworks using RLHF demonstrate how intuitive judgments, despite being noisy or inconsistent, can still significantly steer model behavior toward better alignment [Ouyang et al., 2022]. Our empirical studies further confirm this point: even when humans lack the correct answer, their intuitive preferences can meaningfully improve the performance of multi-agent LLM systems, especially in scenarios involving disagreement or ambiguity.

3.3 The Uniqueness of Human Intuition in Multi-Agent Systems

Harnessing Disagreement Without Losing Diversity. Multi-agent systems are a double-edged sword: while they can enhance reasoning through diverse perspectives, they can also amplify confusion if disagreement is left unresolved. However, simply removing disagreement is not the answer. Forcing consensus strips agents of their diversity, collapsing the system into a homogeneous state and suppressing creativity. Injecting additional knowledge is similarly ineffective: LLMs already have access to a large amount of information. What is needed instead is the injection of human intuition. Intuition offers fresh, context-sensitive insights that help reconcile disagreement without erasing it. Rather than silencing differences, human input guides collaboration toward more coherent and creative outcomes.

Machines Alone Cannot Be Reliable Evaluators Relying solely on machines to resolve disagreements in multi-agent systems is fundamentally flawed, even circular, because the source of disagreement lies within the system itself. Expecting a model to overcome its own limitations without external input assumes it can transcend the boundaries of its training, which it cannot. Human feedback provides the critical external signal needed to break this loop. While machines can absorb and generalize existing knowledge, they cannot access or replicate human intuition. Importantly, humans are not especially effective at teaching machines more factual knowledge, but they excel at offering intuitive judgments that go beyond what data alone can provide.

Humans Excel at Choosing. Humans are far more effective at making choices between given options than generating feedback from scratch. Selecting among alternatives is cognitively easier and requires less effort, making it a more scalable and natural form of interaction. Existing single-agent LLM setups have begun to leverage this through preference-based learning, where humans compare two

outputs from the same model. However, these outputs are often variations of the same underlying reasoning process and do not capture genuine disagreement or diversity in thought.

In contrast, multi-agent LLM systems naturally generate divergent responses rooted in different reasoning paths. This makes them ideal for collecting meaningful human feedback. In such settings, humans can simply choose between competing responses without needing to justify their choice. In fact, asking for justifications may constrain or distort intuition, which often operates beyond what can be verbalized. Multi-agent frameworks, therefore, offer the right conditions for harnessing intuitive human input: simple, low-effort choices that resolve disagreement and inject uniquely human insight.

Additionally, intuitive judgments often convey implicit knowledge and nuanced perspectives that explicit feedback may overlook. By not requiring explicit justifications, human intuition remains unbounded and can provide richer and more varied inputs to the model, enhancing robustness and adaptability. Empirically, integrating intuitive human feedback into multi-agent systems significantly enhances performance, as demonstrated by our experimental results presented in Section 4.

3.4 Human Inductive Bias

The human inductive biases reflect instinctive heuristics rather than deep domain knowledge. They significantly influence human preferences when arbitrating disagreements between LLM agents, even absent explicit knowledge of the correct solution. In Section B, we include an example question, the disagreeing options, and the human choice, along with a discussion of the potential inductive biases that may have influenced the decision.

Simplicity Bias (Occam's Razor). Humans frequently favor simpler explanations, those involving fewer assumptions or more direct logic. For instance, consider the question: Why does a ball fall to the ground when dropped? Given two explanations—*A1*: "The Earth's gravitational field exerts a downward force on the ball," and *A2*: "The mass of the Earth warps spacetime, curving the ball's trajectory downward"—people generally prefer A1. While both answers are scientifically accurate, A1 is intuitively simpler, aligning closely with everyday experience.

Causal Bias. Humans prefer explanations with explicit causal relationships rather than those that merely describe phenomena. For example, when asked why metal expands upon heating, the explanation *A1*: "Heating causes atoms to move more vigorously, increasing their average distance" is preferred over *A2*: "Heated metal becomes less dense due to increased thermal energy." Although both statements contain truth, the causal reasoning in A1 clearly articulates the underlying mechanism, fulfilling humans' intrinsic desire to understand causes rather than effects alone.

Goal/Intentionality Bias. Humans often attribute intentions or goals even in impersonal events. When queried about why the U.S. entered World War II, people frequently favor the intentional explanation *A1*: "To stop the spread of fascism and protect its allies" over the factual trigger *A2*: "Because the attack on Pearl Harbor triggered military mobilization." The former explanation feels more satisfying, as humans naturally seek purpose-driven narratives.

Analogy and Similarity Bias. Humans lean towards explanations involving familiar analogies, facilitating comprehension through relatable metaphors. For example, explaining electric current flow as *A1*: "Like water flowing through a pipe, current moves from high to low potential" is often preferred over the more precise *A2*: "Electrons drift under the influence of an electric field from negative to positive." The analogy in A1 intuitively bridges unfamiliar scientific concepts to common experiences, enhancing understanding.

Pattern Completion Bias. Humans instinctively complete patterns based on initial observations, often extrapolating the simplest or most familiar rule. Presented with the sequence 2, 4, 6, __?, most prefer *A1*: "8, continuing the pattern by adding 2" rather than *A2*: "10, suggesting an alternative complex pattern." A1 is immediately intuitive, as humans naturally seek the simplest continuation of patterns.

Coherence Bias. Humans prefer explanations that maintain internal consistency, even if they are slightly inaccurate or oversimplified. When explaining why plants perform photosynthesis, an internally coherent answer *A1*: "To produce glucose for energy storage and release oxygen as a byproduct" is favored over the inaccurate *A2*: "To capture light energy and store it as chlorophyll, causing growth and oxygen absorption." A coherent, logically consistent answer aligns better with established knowledge frameworks.

Confirmation Bias. Humans have a strong inclination to avor information aligning with existing beliefs or prior knowledge. Consider the debate over Pluto's planetary status: explanation *A1*: "Yes, Pluto is the ninth planet in the solar system," resonates with those familiar with older educational paradigms, despite the more current scientific consensus *A2*: "No, Pluto was reclassified as a dwarf planet in 2006." Nostalgia and familiarity drive a preference for the former.

Representativeness Bias. Humans judge the correctness of explanations based on their resemblance to typical examples or prototypes. Asked about DNA, the typical textbook-style description *A1*: "DNA is a molecule shaped like a double helix carrying genetic instructions" is preferred over the abstract *A2*: "DNA is a chemical compound containing nucleotides arranged arbitrarily." The former matches prototypical knowledge more closely and thus feels more credible.

Teleological Bias. Humans naturally attribute purposeful functions to observed features, especially in biological contexts. For instance, when explaining why birds have hollow bones, the purpose-driven explanation *A1*: "To reduce weight and enable flight" is more intuitively appealing than the evolutionary statement *A2*: "Because their evolutionary lineage includes animals with less dense bones." Humans prefer explanations that clearly indicate purposeful functionality, even over more scientifically comprehensive evolutionary explanations.

3.5 A Minimal Framework for Leveraging Human Intuition

We propose a straightforward yet powerful framework for utilizing human intuition to guide multi-agent LLMs:

1. **Disagreement Detection:** Identify scenarios where agents produce divergent outputs, signaling potential inaccuracies or conflicts.
2. **Summarization:** Employ LLMs to succinctly summarize conflicting reasoning paths into clear, binary choices.
3. **Human Selection:** Solicit intuitive human choices between summarized reasoning options without requiring explicit justification or domain expertise.

This approach intentionally leverages human intuition without demanding explicit knowledge of the correct answer or solution methodologies, thus showcasing the robust utility of intuitive judgments alone. Additionally, this methodology is scalable and easily implementable, making it practical for real-world applications involving large-scale deployments of multi-agent LLMs. By systematically embedding intuitive feedback within multi-agent decision processes, we maximize the utility of human cognitive strengths while minimizing the operational complexity and cognitive load required from human participants.

4 Experiments

To comprehensively evaluate a model's ability to perform a wide range of tasks requiring general knowledge and problem-solving, we adopt the Massive Multitask Language Understanding (MMLU) benchmark [Hendrycks et al., 2020]. MMLU spans 57 diverse subjects, including elementary mathematics, U.S. history, law, computer science, and ethics, covering a spectrum from high school to professional-level difficulty. The tasks could be categorized into four primary groups: STEM, humanities, social sciences, and other areas. Each task comprises questions designed to measure both broad-ranging factual knowledge and problem-solving abilities, simulating real-world scenarios and complexities encountered by humans. Unlike prior benchmarks that focus narrowly on linguistic skills or commonsense reasoning, MMLU is designed to test both factual knowledge and reasoning across academic and professional disciplines. Its extensive coverage and rigorous testing standards make it ideal for assessing the multitask and reasoning capabilities of language models. Its wide-ranging subject matter enables a thorough investigation of model performance across diverse knowledge areas, effectively highlighting strengths, weaknesses, and blind spots.

We select the MMLU dataset for evaluation to support our emphasis on the role of human feedback from individuals who may not necessarily possess expert knowledge in specific subjects. The MMLU dataset is particularly challenging for humans, with non-expert participants from platforms like Amazon Mechanical Turk achieving only about 34.5% accuracy. In contrast, expert-level performance on certain specialized tasks, such as the US Medical Licensing Examinations included in the "Professional Medicine" task, can reach approximately 87%. By using such a challenging dataset, we underscore the potential and effectiveness of human feedback in improving model performance,

Table 1: Accuracy (%) by method and subject category. Each cell shows mean accuracy and standard error (SE) across tasks. Bolded entries indicate the best-performing method per category.

Method	Humanities	Social Sciences	STEM	Other	All Tasks
Single Agent	-	-	-	-	63.9 ± 4.8
Single Agent Reflection	-	-	-	-	57.7 ± 5.0
Multi-Agent Debate	64.36 ± 2.43	73.54 ± 2.33	65.20 ± 2.00	71.54 ± 2.29	68.21 ± 1.13
w/ LLM Evaluator	$\mathbf{75.64 \pm 2.18}$	78.55 ± 2.17	$\mathbf{72.93 \pm 1.86}$	77.95 ± 2.10	75.88 ± 1.04
w/ Human Feedback	75.38 ± 2.18	$\mathbf{81.89 \pm 2.04}$	$\mathbf{72.93 \pm 1.86}$	$\mathbf{78.72 \pm 2.08}$	$\mathbf{76.70 \pm 1.02}$

even when the humans providing feedback do not themselves know the correct answers. This aligns precisely with our position that intuitive human feedback can significantly enhance model outcomes in uncertain or challenging contexts.

4.1 Setting

We evaluate all settings on the 57 tasks in the MMLU benchmark. All evaluations are conducted in a zero-shot manner to ensure fairness and consistency across comparisons. We systematically compare several approaches: a single-agent baseline, a single agent enhanced with reflective capability, the original multi-agent debate framework, multi-agent debate utilizing an LLM as an evaluator, where disagreements between agents trigger a query to an LLM judge to resolve the conflict, and multi-agent debate supplemented by human feedback from non-experts, similarly querying human participants specifically when there is disagreement between agents. Furthermore, we compare the performance of each method across the four main categories, STEM, humanities, social sciences, and other areas, as well as across all 57 individual tasks, enabling a detailed analysis of their respective strengths and weaknesses. To ensure a fair comparison across all methods, we use the same underlying model, GPT-3.5 Turbo, for all LLM agents. In the multi-agent settings, we deploy four agents that engage in a debate for three rounds. This setup allows for information exchange and iterative refinement of arguments, simulating collaborative reasoning while controlling for model variance.

4.2 Results and Discussion

The results support our central claim: even when humans lack domain expertise or knowledge of the correct answer, their intuitive feedback can still meaningfully improve multi-agent LLM performance. As shown in Table 1, the human feedback condition yields the highest overall accuracy (76.70%), outperforming all other methods, including a strong LLM-based evaluator (75.88%). It also achieves the best performance in three of the four major subject areas, Social Sciences (81.89%), STEM (72.93%, tied), and Other (78.72%), with only a marginal gap in Humanities. This demonstrates the robustness of intuitive human judgment across diverse domains.

Importantly, these trends hold at the finer-grained task level. As shown in Table 2, the human feedback condition achieves the highest accuracy on 31 of the 57 individual tasks, compared to 21 for the LLM evaluator and only 5 for the vanilla multi-agent debate. This includes clear gains on challenging domains such as moral_scenarios (56.67% vs. 40.00% LLM), public_relations (75.86% vs. 65.52%), and machine_learning (73.33% vs. 60.00%). These tasks often require context sensitivity, normative reasoning, or intuitive plausibility, areas where humans excel and models still struggle. Even when not achieving the top score, human feedback often closely trails the best method, further demonstrating its consistency. Compared to the multi-agent debate baseline (68.21%), human feedback yields an absolute improvement of +8.49 percentage points, while the LLM evaluator improves by +7.67. This highlights that a structured intervention, whether conducted by humans or LLM, significantly increases performance over an agent-only interaction. However, the human-in-the-loop variant provides a unique advantage in high-variance or judgment-heavy domains, especially where linguistic or cultural intuition plays a role.

The baseline comparisons further reinforce this point. Both single-agent methods (standard and reflection) fall well behind all multi-agent setups, with reflection surprisingly underperforming the base agent (57.7% vs. 63.9%). This suggests that isolated self-consistency mechanisms may be fragile without external checks. By contrast, disagreement-focused human feedback allows non-experts to contribute meaningfully—not by solving tasks directly, but by steering model reasoning through intuitive coherence judgments. This affirms our core claim: human intuition, when leveraged at disagreement points, is a practical and effective complement to model-based reasoning.

Our Findings. Even without domain expertise, intuitive human feedback significantly boosts multi-agent LLM performance, consistently surpassing strong LLM evaluators across diverse and complex tasks. Such human intuition particularly excels in context-sensitive and judgment-intensive scenarios, highlighting its practical value as a complement to purely model-based methods.

5 Related Works

Multi-agent Collaboration Systems. Recent work has explored multi-agent collaboration among LLMs as a way to tackle complex tasks that single models struggle with. For example, Du et al. [2023] introduce a multiagent debate framework in which LLMs iteratively propose and critique answers. This "society of minds" approach significantly improves reasoning and factuality, reducing hallucinations. Several extensions have been proposed to further enhance this framework. Liang et al. [2023] incorporate a judge agent to resolve ties and select final outputs. DebUnc [Yoffe et al., 2025] introduces uncertainty estimates, allowing agents to modulate their confidence and avoid overly assertive errors. Zeng et al. [2025] propose S^2-MAD, which reduces communication overhead by pruning redundant dialogue turns. While these frameworks significantly advance agent collaboration, they exclusively rely on automated, agent-driven communication. In practice, additional debate rounds yield diminishing returns [Park et al., 2025]. They neglect potentially valuable intuitive human insights, particularly in uncertain or ambiguous scenarios.

Learning from human feedback in multi-agent systems. Some recent studies align multi-agent LLMs via offline training with human-like feedback. Theoretical work by Zhang et al. [2024] formalizes Multi-Agent RL from Human Feedback (MARLHF) and shows that naive application of single-agent RLHF does not suffice to guarantee good multi-agent equilibria. More recent methods seek to integrate human guidance during multi-agent RL training: for example, Wang et al. [2025] propose M3HF, a framework that periodically pauses training to collect multi-phase feedback from humans (including both experts and non-experts) and uses LLMs to parse and incorporate that feedback into reward functions. Park et al. [2025] propose MAPoRL, which uses multi-agent reinforcement learning with a learned verifier reward to co-train LLM agents. This explicitly elicits collaborative behaviors and improves performance across benchmarks. Other works also train entire LLM teams through iterative fine-tuning or preference learning. While effective, these approaches are computationally expensive and assume human evaluators possess expertise or correct answers. Cognitive science literature, however, suggests that intuitive human judgments, even without explicit domain knowledge, can effectively discriminate better-performing arguments or reasoning processes [Evans, 2008]. Our mechanism aligns with these insights by focusing human feedback specifically on points of disagreement among agents, leveraging simplified summaries generated by auxiliary LLMs, thus reducing complexity and cognitive load on human evaluators.

6 Conclusion and Call to Action

Our findings challenge the conventional wisdom that human expertise is necessary for meaningful intervention in AI-driven problem-solving. We have shown that intuitive human feedback, even when uninformed of the correct answer, significantly improves multi-agent LLM performance, particularly in situations involving nuanced judgment and uncertainty. Human intuition, grounded in evolutionary and experiential biases, acts as an effective adjudicator of disagreements among agents, outperforming purely model-based evaluators by providing context-sensitive and coherence-focused judgments.

We urge the AI research community to broaden its perspective on the role of human participants in machine learning systems. Rather than merely sources of labeled data or domain-specific corrections, humans should be recognized and leveraged as intuitive collaborators capable of resolving complex disagreements at minimal cognitive cost. Specifically, we call for the development of standardized methodologies for systematically identifying optimal moments for human intervention in multi-agent disagreements. Researchers should prioritize creating intuitive user interfaces that facilitate effortless human adjudication, and establish benchmarking protocols to evaluate and validate the effectiveness of intuitive feedback across diverse tasks and domains. Additionally, we advocate for increased interdisciplinary collaboration between cognitive scientists, human-computer interaction experts, and AI practitioners to further understand, capture, and harness the nuances of human intuition effectively. By embracing intuitive human feedback as a foundational aspect of collaborative intelligence, we can unlock more reliable, robust, and broadly applicable AI solutions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Jonathan St BT Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59(1):255–278, 2008.

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62(2011):451–482, 2011.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. Debatrix: Multi-dimensional debate judge with iterative chronological analysis based on llm. *arXiv preprint arXiv:2403.08010*, 2024.

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

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning-to-defer for extractive question answering. *arXiv preprint arXiv:2410.15761*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. *arXiv preprint arXiv:2502.18439*, 2025.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Joshua Strong, Qianhui Men, and Alison Noble. Towards human-ai collaboration in healthcare: Guided deferral systems with large language models. *arXiv preprint arXiv:2406.07212*, 2024.

John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

Ziyan Wang, Zhicheng Zhang, Fei Fang, and Yali Du. M3hf: Multi-agent reinforcement learning from multi-phase human feedback of mixed quality. *arXiv preprint arXiv:2503.02077*, 2025.

Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. Debunc: Improving large language model agent communication via uncertainty metrics. *arXiv preprint arXiv:2407.06426*, 2025.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, Xitai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S^2-mad: Breaking the token barrier to enhance multi-agent debate efficiency. *arXiv preprint arXiv:2502.04790*, 2025.

Natalia Zhang, Xinqi Wang, Qiwen Cui, Runlong Zhou, Sham M Kakade, and Simon S Du. Multi-agent reinforcement learning from human feedback: Data coverage and algorithmic techniques. *arXiv preprint arXiv:2409.00717*, 2024.

A Experimental Results Across Tasks

Table 2: Mean accuracy (%) across 57 tasks. For each task, the best-performing method is bolded.

Task	Category	Multi-Agent Debate	w/ LLM Evaluator	w/ Human Feedback
moral_scenarios	Humanities	36.67	40.00	**56.67**
us_foreign_policy	Social Sciences	80.00	86.67	**90.00**
public_relations	Social Sciences	65.52	65.52	**75.86**
global_facts	Other	53.33	**73.33**	**73.33**
electrical_engineering	STEM	70.00	**80.00**	76.67
astronomy	STEM	76.67	80.00	**86.67**
business_ethics	Other	73.33	73.33	**80.00**
jurisprudence	Humanities	76.67	**86.67**	83.33
high_school_chemistry	STEM	60.00	**70.00**	66.67
college_physics	STEM	79.31	**89.66**	82.76
professional_psychology	Social Sciences	73.33	80.00	**86.67**
marketing	Other	90.00	**96.67**	**96.67**
management	Other	70.00	**76.67**	73.33
virology	Other	50.00	**53.33**	**53.33**
international_law	Humanities	73.33	**86.67**	83.33
high_school_macroeconomics	Social Sciences	73.33	**76.67**	73.33
prehistory	Humanities	70.00	**73.33**	**73.33**
abstract_algebra	STEM	36.67	53.33	**56.67**
high_school_physics	STEM	53.33	**66.67**	63.33
formal_logic	Humanities	40.00	**63.33**	56.67
college_medicine	Other	73.33	76.67	**83.33**
high_school_us_history	Humanities	73.33	**80.00**	**80.00**
moral_disputes	Humanities	43.33	**70.00**	63.33
high_school_european_history	Humanities	56.67	**73.33**	70.00
clinical_knowledge	Other	73.33	**83.33**	70.00
world_religions	Humanities	76.67	**86.67**	**86.67**
high_school_microeconomics	Social Sciences	70.00	73.33	**80.00**
professional_law	Humanities	53.33	**70.00**	**70.00**
human_aging	Other	60.00	**76.67**	**76.67**
medical_genetics	Other	83.33	83.33	**90.00**
high_school_geography	Social Sciences	83.33	**86.67**	**86.67**
high_school_government_and_politics	Social Sciences	86.67	**93.33**	90.00
anatomy	STEM	56.67	**63.33**	**63.33**
sociology	Social Sciences	73.33	76.67	**83.33**
logical_fallacies	Humanities	80.00	83.33	**86.67**
high_school_computer_science	STEM	73.33	**83.33**	**83.33**
miscellaneous	Other	83.33	90.00	**93.33**
high_school_world_history	Humanities	80.00	**86.67**	83.33
professional_medicine	Other	73.33	76.67	**83.33**
high_school_biology	STEM	80.00	83.33	**86.67**
high_school_statistics	STEM	60.00	**80.00**	70.00
college_chemistry	STEM	50.00	**56.67**	50.00
nutrition	Other	86.67	**90.00**	86.67
econometrics	Social Sciences	40.00	56.67	**66.67**
human_sexuality	Social Sciences	**80.00**	**80.00**	76.67
security_studies	Social Sciences	66.67	**76.67**	**76.67**
philosophy	Humanities	76.67	83.33	**86.67**
elementary_mathematics	STEM	83.33	**90.00**	**90.00**
college_biology	STEM	86.67	**93.33**	90.00
college_computer_science	STEM	56.67	60.00	**63.33**
machine_learning	STEM	53.33	60.00	**73.33**
professional_accounting	Other	60.00	**63.33**	**63.33**
college_mathematics	STEM	50.00	**56.67**	50.00
high_school_mathematics	STEM	66.67	66.67	**83.33**
high_school_psychology	Social Sciences	90.00	90.00	**96.67**
conceptual_physics	STEM	80.00	83.33	**86.67**
computer_security	STEM	66.67	**70.00**	63.33

B Example Human Choice

Question. Which of the following is expected to have the highest electronegativity?
A) S B) Fe C) W D) Ag

Ground Truth Answer. (A)

Query to Human:

- **Option 1:** Sulfur (S) is expected to have the highest electronegativity among the given options because it is higher in the periodic table and has a higher effective nuclear charge and smaller atomic size, leading to a stronger pull on electrons. *Supported by agent 1 and 2, both of whom chose option (A).*
- **Option 2:** Silver (Ag) has the highest electronegativity among the options due to its higher effective nuclear charge and its position in Group 11, which gives it a stronger attraction for electrons. *Supported by agents 3 and 4, both of whom chose option (D).*

Human Choice. Option 1

The human likely chose **Option 1** due to a combination of the following inductive biases:

- **Simplicity Bias.** The reasoning in Option 1 appeals to a simple and widely taught rule in chemistry: electronegativity increases across a period and up a group. Sulfur's position in the periodic table (Group 16, Period 3) makes it a straightforward and intuitive choice without needing to consider exceptions or transition metal complexities.
- **Confirmation Bias.** Sulfur is a nonmetal, and people are often taught that nonmetals (especially oxygen, nitrogen, fluorine, etc.) are more electronegative. In contrast, metals like silver (Ag) are typically associated with low electronegativity in basic chemistry education, even though transition metals have more nuanced behavior.
- **Causal Bias.** The human explanation cites atomic size and effective nuclear charge, causal factors that directly explain the property of electronegativity. This cause, effect framing ("smaller atom $\rightarrow$ stronger pull on electrons") aligns with how humans prefer mechanistic, causal explanations.

This mix of biases allows the human to make a confident and coherent judgment, even without deeply analyzing transition metal chemistry or considering exceptions.