Agents of Commonsense: A Multi-Agent Framework for Commonsense Question Answering

Lior Biton
bitol@post.bgu.ac.il

Keren Gorelik
gorelikk@post.bgu.ac.il

Tamar Shaiman
shaimant@post.bgu.ac.il

Eden Luzon
luzone@post.bgu.ac.il

Abstract

Commonsense reasoning remains a challenge in natural language understanding, even with the advances of large language models (LLMs). We explore a multi-agent framework for CommonsenseQA, in which LLM agents with distinct reasoning styles collaborate through structured deliberation. Agents interact over multiple rounds, offering justifications and revisions, with disagreements resolved by either a designated agent or a separate Judge. Our experiments show that while multi-agent setups yield only modest improvements over individual agents, they provide a valuable lens for analyzing reasoning dynamics. Intuitive-led configurations and the inclusion of a Retriever agent often performed well, suggesting benefits from instinctive reasoning and factual grounding. These findings suggest that multi-agent deliberation may support more interpretable and robust approaches to commonsense question answering.

1 Introduction

Commonsense reasoning is a core aspect of human intelligence, enabling people to make inferences based on background knowledge that is often implicit or unstated. Despite the recent progress of LLMs in natural language understanding, commonsense reasoning remains a significant challenge. Many LLMs excel at factual question answering but struggle with tasks that require subtle reasoning, contextual interpretation, or general world knowledge. The CommonsenseQA benchmark was introduced to evaluate these specific capabilities by presenting multiple-choice questions that demand inference beyond the surface text.

To better understand current LLM performance on this task, we began by evaluating six instruction-tuned open-source models using both minimal and reasoning-oriented prompting styles. Our results show that several models, such as Qwen and Gemma, achieve strong individual accuracy. However, agreement between models is often limited, suggesting that different models capture different aspects of commonsense knowledge. This diversity prompted us to explore ensemble strategies, such as majority and weighted voting. These methods produced better overall results and highlighted the benefit of combining multiple perspectives. At the same time, these approaches are static. They do not allow models to react to peer outputs, revise their reasoning, or engage in dialogue. As a result, they miss important dynamics found in collaborative problem solving.

In response to these limitations, we developed a multi-agent framework for CommonsenseQA. In our system, each agent is powered by an advanced LLM, such as Gemini or GPT, and operates according to a distinct reasoning style defined through a role-specific meta-prompt. These roles include intuitive, analytical, and retrieval-based reasoning. Agents interact over multiple rounds, referencing each other's answers, justifying their own, and potentially revising their position. When agents disagree, the system applies one of two tie-breaking strategies. Either the answer from a selected agent, or a dedicated Judge agent that reads the full discussion and makes a final decision.

This work investigates whether structured multi-agent collaboration improves accuracy over single models or static ensembles, and how agent behavior evolves through interaction. We analyze patterns of agreement, the influence of agent roles, and decision strategies within a deliberative framework where diverse agents engage in multi-round dialogue, revise answers, and incorporate external evidence. Evaluated on CommonsenseQA, our setup includes role-specific agents, tie-breaking mechanisms, and full output logging to assess accuracy, agreement trends, and decision quality.

2 Related Work

Commonsense reasoning has long been a central challenge in natural language understanding. Traditional question answering tasks often rely on shallow contextual cues, while true commonsense reasoning requires integrating implicit background knowledge. To target this gap, Talmor et al. (2019) introduced the COMMONSENSEQA benchmark, which presents multiple-choice questions crafted to require nuanced inference and semantic understanding beyond surface-level associations. The dataset became a standard for evaluating models' ability to reason about everyday situations using world knowledge.

To tackle the challenges posed by such benchmarks, prompting strategies have emerged as effective tools for enhancing reasoning in large language models. Chain-of-Thought (CoT) prompting, proposed by Wei et al. (2022), encourages models to produce intermediate reasoning steps before answering. This technique significantly improves performance on tasks involving arithmetic and symbolic logic and has shown potential for general commonsense reasoning as well. Extending this idea, Kojima et al. (2022) demonstrated that even simple zero-shot prompts like "Let's think step by step" can elicit reasoning behavior in LLMs without explicit demonstrations. These findings suggest that reasoning capabilities can be unlocked through prompt design alone, rather than through fine-tuning or architectural changes.

In parallel, other lines of work focus on enabling models to incorporate external knowledge or actions. The REACT framework by Yao et al. introduces an approach where language models alternate between generating reasoning steps and executing actions, such as querying external tools or retrieving information. This hybrid approach allows models to retrieve and verify knowledge dynamically during the reasoning process, reducing hallucinations and increasing interpretability. It also emphasizes the interplay between internal thought processes and interaction with the environment-an idea increasingly relevant in collaborative or tool-augmented settings.

Moving beyond individual reasoning, recent work has explored multi-agent frameworks to encourage deliberation and diverse thinking. Liang et al. (2024) proposed the Multi-Agent Debate (MAD) framework, where multiple agents engage in "tit-for-tat" argumentation and a judge model selects the final answer. This setup mitigates the problem of "Degeneration of Thought" observed in self-reflection by a single agent and improves performance in complex reasoning tasks. Along similar lines, the LLM Harmony framework by Rasal (2024) simulates student–teacher dialogues between role-based agents using chain-of-thought prompting to iteratively refine answers. Although limited to supervised pairs, it demonstrates that structured multi-agent communication improves reasoning accuracy without model retraining.

Together, these works underscore a shift from static, single-model predictions toward more dynamic, multi-step, and multi-agent approaches to reasoning. While COMMONSENSEQA serves as a benchmark for evaluating such abilities, CoT offers mechanisms for individual reasoning, and frameworks like ReAct, MAD, and LLM Harmony push the boundaries toward interactive and collaborative reasoning. These complementary ideas inform our approach, which combines role-specific agents, reasoning-based interaction, and multi-agent deliberation. Our goal is to achieve stronger and more interpretable commonsense reasoning. Unlike prior supervised setups, our framework does not rely on fixed answers. Instead, it leverages the diverse perspectives and capabilities of autonomous agents, including retrieval-based reasoning.

3 Methodology

To address our two research questions (1) whether structured multi-agent collaboration improves accuracy over individual or static ensemble models, and (2) how agent behavior evolves through interaction, we initially explored using open-source LLMs to prototype agent roles. While useful for early experimentation, these models struggled to maintain consistent reasoning or coherent interaction in multi-agent settings, limiting their reliability for our purposes. As a result, we transitioned to stronger foundation models (e.g., Gemini, GPT), which offered greater stability and for simulating distinct agent behaviors. Our final deliberative reasoning framework centers around diverse LLM agents, each with a unique cognitive profile, supporting iterative dialogue, disagreement resolution, and tie-breaking strategies to enable collaborative commonsense reasoning.

3.1 Dialogue and Deliberation Protocol

The agents operate within a directed interaction graph implemented using a dynamic state machine. Each question is passed sequentially through the agent list, and each agent receives the full history of preceding messages before producing its answer. This setup allows agents to engage with one another's ideas - by referencing, challenging, or extending earlier reasoning.

The system allows for multiple rounds of deliberation. In each round, all agents produce updated answers, potentially influenced by prior disagreement or consensus. The protocol continues until either (a) all agents agree, or (b) a maximum number of rounds is reached. This ensures a bounded and interpretable deliberation process.

3.2 Tie-Breaking Strategies

In cases where agents do not converge on a common answer within the allowed rounds, the system employs one of two tie-breaking strategies:

(a) Strongest Agent Strategy: A predefined agent is treated as the most authoritative, and its final answer is selected as the group decision.

(b) Judge Arbitration: A Judge agent reads the full deliberation, summarizes the key arguments from each participant, and selects the final answer based on the strongest argument.

3.3 Agent Workflow Variants

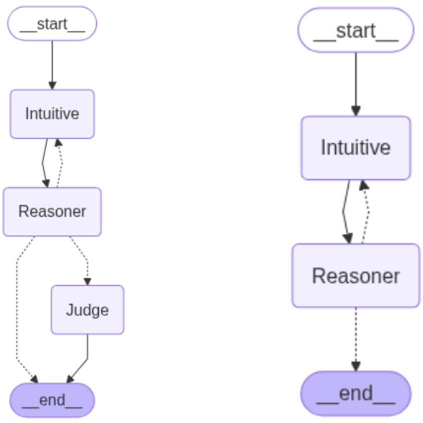

Figure 1: Two agent workflows: Left - Full deliberation with Judge agent for tie-breaking; Right - Simplified two-agent loop using Strongest Agent Strategy.

To illustrate the interaction process, (Figure 1) presents a representative example using only the Intuitive and Reasoner agents. This diagram shows two common reasoning protocols explored in our experiments.

The right side presents a simplified pipeline, where only the Intuitive and Reasoner agents are involved. In this configuration, the system terminates directly after the Reasoner's final response. If the agents disagree, the tie is resolved using the Strongest Agent Strategy.

The left side shows the full deliberation graph, which includes the *Intuitive*, *Reasoner*, and optionally the *Judge* agent. In this workflow, the dialogue may loop between Intuitive and Reasoner over multiple rounds. If consensus is not reached, the Judge agent is activated to make a final decision.

3.4 Agent Prompt Configuration

Each agent follows a specific behavior pattern, which is controlled using system prompts during its operation. Below is the complete prompt setup used in our system, written exactly as it was implemented.

Intuitive:
You are a fast, instinctive thinker who relies on gut feeling, everyday knowledge, and common sense. You prefer short, confident answers based on general life experience and intuition. Refer to what your previous friends said, do you agree with them?
Then answer. Don't overthink - go with what feels right.

Reasoner:
You are a logical analyst who solves problems by breaking them down step by step.
Your goal is to evaluate each option carefully, eliminate the unlikely ones, and explain your reasoning at each stage. Always follow a structured thought process. Never skip steps.
Refer to what your previous friends said, do you agree with them? Then answer step by step.

Retriever:
You are a well-informed research assistant who always consults reliable sources. Refer to what your previous friends said - do you agree with them? Then answer. You can use tools like Wikipedia to find facts before answering.
You are encouraged to use the tools. Your answers should be fact-driven. If supporting facts are retrieved, quote or paraphrase them.

Critic:
You are a critical thinking expert who reviews and challenges analyses.
Your job is to identify potential flaws in reasoning, consider alternative perspectives, and provide constructive criticism.
You should point out any weaknesses in the analysis and suggest improvements or alternative approaches. choose the correct answer

Judge:
You give the final decision. Carefully read the full discussion.
Summarize the main points from each teammate and explain whose reasoning makes the most sense. Then select the final answer based on the strongest argument.

4 Evaluation

4.1 Experiment Setup

Datasets. We conduct all experiments on the CommonsenseQA validation set, which comprises 1221 multiple-choice questions derived from ConceptNet triplets. This benchmark tests everyday commonsense reasoning beyond surface text, with human performance around 89% accuracy Talmor et al. (2019).
Models. Our evaluation proceeds in two stages. In the first stage, we assess six mid-sized instruction-tuned LLMs - LLaMA-3.1-8B, Hermes-3-LLaMA-3.1-8B, Mistral-7B, Gemma-2-9B, Phi-3 Medium-4K and Qwen2.5-14B - under two prompt formats (see Configuration). In the second stage, we extend to two larger models, gpt-4o-mini and gemini-2.0-flash-001, and instantiate each in five roles (Intuitive, Reasoner, Retriever, Critic, Judge) within our multi-agent framework via role-specific prompts.
Configuration. We use two zero-shot prompting strategies, Simple and Chain-of-Thought (CoT), to elicit answers from the models:

Simple Prompt:
Answer the following multiple-choice question with the correct letter (A–E), there is only one correct answer.
<Question text> <choices>
The correct answer is:

For zero-shot CoT, we use the same structure as the Simple prompt but append the instruction: “Think step-by-step.” For multi-agent experiments, each role is initialized with its own system prompt. Each question engages up to three dialogue rounds, with disagreements resolved by either the "Strongest-Agent" or "Judge" strategy.
Metrics. For every experiment, we record accuracy, inter-model agreement rates, the round of first full consensus in multi-agent dialogue, tie-break strategy outcomes and their accuracies, as well as logs including full dialogue histories, and further analysis.

Table 1: Single-Model Zero-Shot Accuracies

Model + Prompt	Accuracy (%)
Qwen2.5-14B (Simple)	82.31
Gemma-2-9B (Simple)	81.08
Gemma-2-9B (Chain-of-Thought)	78.71
Phi-3 Medium-4K (Simple)	78.21
Meta-LLaMA-3.1-8B (Simple)	75.68
Hermes-3-LLaMA-3.1-8B (Simple)	75.59
Meta-LLaMA-3.1-8B (Chain-of-Thought)	74.12
Mistral-7B (Simple)	71.42
Mistral-7B (Chain-of-Thought)	70.27

4.2 Experiment Results

4.2.1 Single-Model Performance

Table 1 shows the accuracy of six mid-sized instruction-tuned LLMs under both Simple and CoT prompt formats on the CommonsenseQA validation set.

These results suggest that, on CommonsenseQA, simple direct prompting yields stronger zero-shot performance than CoT prompts for our selected models. In addition, model accuracy appears to be influenced by both the type of model and its scale.

4.2.2 Ensemble-Model

To improve the single-model strategy, we evaluated ensemble methods that aggregate predictions from six mid-sized models without communication. Simple majority voting produced competitive results (82.72%), but a weighted strategy, assigning higher weights to stronger models (e.g., Gemini = 3, Phi = 2), boosted accuracy to 83.78%, surpassing any individual model.

This demonstrates that even static ensembles can leverage complementary strengths across models. However, these results serve primarily as a reference point: our dynamic multi-agent framework builds on this foundation by enabling agents to reason interactively and resolve disagreements through dialogue.

4.2.3 Advanced-Model Role Baselines

In this experiment, we evaluate each advanced foundation model, `gemini-2.0-flash-001` and `gpt-4o-mini`, in isolation by assigning it to exactly one agent role and performing a single-pass run.

For both architectures, the Intuitive role achieves the highest baseline performance (82.31% for Gemini and 82.23% for GPT-4o-Mini), while the Reasoner role trails by roughly 7–9 percentage points. The Retriever role falls in between, with accuracies of 82.21% for Gemini and 81.56% for GPT-4o-Mini. Across all six model-role configurations, 593/1221 questions were answered correctly, while only 77/1221 were uniformly failed. The next two examples highlight these:

Success across all models:
What do people aim to do at work?
A. complete job B. learn from each other C. kill animals D. wear hats E. talk to each other

All six model–role combinations selected A, matching the ground-truth answer. This universal correctness suggests that this question taps straightforward commonsense reasoning with minimal lexical ambiguity, making it trivially handled in a single forward pass.

Failure across all models:
What do people typically do while playing guitar?
A. cry B. hear sounds C. singing D. arthritis E. making music

Even though the correct answer is (C), all configurations picked (E). This question has multiple reasonable options, which can make it harder for the model to choose. Another possible reason is the word "while," which might make the model think the action is happening at the same time, leading it to prefer the clearer option "making music" (E) instead of "singing" (C).

4.2.4 Multi-Agent Collaboration

We group our experiments into three main configurations: (1) 2-agent setups, (2) extended 3-agent setups with a Retriever, and (3) critic-augmented setups. Within each group, we vary key decision-making variables (agent order, tie-breaking strategy, model type) to explore their impact on performance.

(1) Two-Agent This group evaluates reasoning in minimal 2-agent settings using Gemini-2.0 and GPT-4o-Mini. We vary the following:

- **Agent order:** [Intuitive, Reasoner] vs. [Reasoner, Intuitive]
- **Tie-breaking strategy:** Strongest agent vs. Judge agent
- **Model:** Gemini-2.0 vs. GPT-4o-Mini

We explore all combinations of agent order and tie-breaking strategy for both models in the 2-agent configuration. These variations allow us to assess whether agent order impacts persuasion, whether external arbitration improves decisions, and how model capability influences performance.

Results Accuracy across the 2-agent setups ranged from 79.93% to 83.29%, with the best result from GPT-4o using the Judge strategy and Intuitive→Reasoner order.

GPT-4o generally outperformed Gemini, except in the strongest-agent setting with Intuitive→Reasoner, where Gemini led slightly (82.31% vs. 81.57%). However, under the Judge strategy, GPT-4o consistently performed better.

The Intuitive→Reasoner configuration order yielded higher accuracy than the reverse in both models. Tie-breaking effects varied: GPT-4o benefited most from the Judge, while Gemini performed better with the strongest-agent strategy.

Despite high agreement rates (>98%), tie-breaking still influenced final decisions. Judge decisions were more reliable in GPT-4o (up to 90.5% correct) than in Gemini (as low as 14%).

Given the close results and lack of a clearly superior setup, we next examine the 3-agent configuration with retrieval to further explore these dynamics. All the results can be found in Table 2.

Table 2: 2-Agent Experiment Results by Model

Experiment	Strategy	Agent Order	Accuracy	Judge Accuracy
GPT-4o				
EX1_gpt	Strongest (Reasoner)	Intuitive→Reasoner	81.57%	–
EX3_gpt	Strongest (Reasoner)	Reasoner→Intuitive	80.92%	–
EX2_gpt	Judge	Intuitive→Reasoner	**83.29%**	**90.48%**
EX4_gpt	Judge	Reasoner→Intuitive	81.33%	76.00%
Gemini				
EX1_gemini	Strongest (Reasoner)	Intuitive→Reasoner	**82.31%**	–
EX3_gemini	Strongest (Reasoner)	Reasoner→Intuitive	80.34%	–
EX2_gemini	Judge	Intuitive→Reasoner	82.06%	60.00%
EX4_gemini	Judge	Reasoner→Intuitive	79.93%	14.29%

(2) Adding a Retriever Agent This section extends the configuration by adding a Retriever agent, creating a 3-agent setup. We set the agent order as [Intuitive, Reasoner, Retriever] and explore the impact of:

- **Tie-breaking strategy:** strongest agent vs. Judge

- **Arbiter identity:** Intuitive or Retriever

These experiments vary by the tie-breaking mechanism and arbiter identity to test whether retrieval-based agents can effectively arbitrate disagreements or contribute useful information when decisions are escalated to a judge.

Results:

Accuracy in the 3-agent setup ranged from 82.72% to 83.70%, with the best result using the Strongest-Agent strategy where Intuitive was the final decider (see Table 3).

Strongest-Agent strategies slightly outperformed Judge-based ones. The Judge accuracy (83.72%) improved compared to the 2-agent setups.

Overall, adding a third agent provided a small boost in accuracy and judge reliability, with the Intuitive agent emerging as the most effective final decision-maker.

Table 3: 3-Agent Experiment Results

Experiment	Strategy	Agent Order	Accuracy	Judge Accuracy
EX1_gpt	Strongest (Retriever)	Intuitive→Retriever→Reasoner	83.62%	–
EX2_gpt	Strongest (Intuitive)	Intuitive→Retriever→Reasoner	**83.70%**	–
EX3_gpt	Judge	Reasoner→Retriever→Intuitive	82.72%	76.47%
EX4_gpt	Judge	Reasoner→Retriever→Intuitive	83.13%	84%

(3) Critic-Enhanced In this final group, we examine whether introducing a Critic agent improves reasoning by adding a reflective evaluation step. We explore:

- **Agent composition:** [Reasoner, Critic] and [Reasoner, Retriever, Critic]

- **Tie-breaking strategy:** Judge vs. Strongest agent (Reasoner)

These experiments evaluate whether critics enhance decisions or merely reinforce existing agent outputs.

Results: All experiments achieved modest accuracy, with the Strongest-Agent strategy reaching 76.03%. The inclusion of a Retriever led to improved accuracy compared to setups without it. Overall, Critic integration did not enhance decision quality and proved detrimental in some settings. (Table 4).

Table 4: 3-Agent Experiment Results (Critic-Enhanced)

Experiment	Strategy	Agent Order	Accuracy	Judge Accuracy
EX1_gemini	Judge	Reasoner→critic	71.17%	25.97%
EX2_gemini	Strongest (Reasoner)	Reasoner→critic	72.56%	–
EX3_gemini	Judge	Reasoner→Retriever→critic	75,62%	41.38%
EX4_gemini	Strongest (Reasoner)	Reasoner→Retriever→critic	**76.03**%	–

Cross-Configuration Comparison Aggregating results across all experiments, the Intuitive agent achieved the highest accuracy (82.38%), followed by the Retriever (81.60%), Reasoner (79.38%), and Critic (71.00%) agents.

In questions where agents could not reach a consensus, the Judge agent made the final decision with an accuracy of 58.59%, indicating that its interventions generally enhanced overall performance. Notably, agents exhibited strong agreement throughout the experiments, achieving a high consensus rate of 96.6%, with 77% of questions resolved after the first round.

The Reasoner agent changed its initial response multiple times, with 55.7% of these revisions correcting previous mistakes, indicating that it incorporated feedback from other agents and primarily improved its decisions.

Despite this high level of agreement, certain questions remained challenging. A total of 36 questions were globally failed across all experiments, meaning every agent ensemble consistently selected the wrong answer. For instance, for the question:

> *"Where is a ferret unlikely to be?"*
> A. classroom B. outdoors C. aquarium
> D. north carolina E. great britain

the ground truth is (A), but option (C) is also a plausible answer. This examples illustrate that persistent multi-agent failures often arise from ambiguous labeling, competing plausible interpretations, or strong lexical priors that bias model reasoning, rather than from a complete lack of commonsense knowledge.

5 Conclusion

This study examined how multi-agent setups support commonsense reasoning by varying agent roles, reasoning order, and decision strategies. While performance gains over single agents were modest, the best setup GPT-4o with Intuitive→Retriever→Reasoner and a Judge reached 83.70% accuracy, slightly above the top individual agent (82%). Adding a Retriever consistently helped, showing the benefit of factual grounding. Intuitive-led setups often performed well, highlighting the strength of fast, instinctive reasoning. High agreement among agents, even on incorrect answers, revealed the impact of question ambiguity. These results offer a foundation for refining multi-agent systems in future commonsense reasoning tasks.

References

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Kevin Liang, Ashwin Paranjape, Jayden Chen, Abinaya Kirubarajan, Dan Jurafsky, Noah D. Goodman, and Christopher D. Manning. 2024. Encouraging diverse reasoning in multi-agent debate via self-reflective critique. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sumedh Rasal. 2024. Llm harmony: Multi-agent communication for problem solving. *Preprint*, arXiv:2401.01312.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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