Awareness in LLMs Improves Through Collaboration

Lucia Passaro[1,*] **Simone Marzeddu**[1] **Andrea Cossu**[1] **Davide Bacciu**[1]

Department of Computer Science
University of Pisa

Abstract

As Large Language Models advance in reasoning and generation, interest in their collaborative potential has grown. This paper investigates agentic reasoning collectives, i.e., structured groups of LLMs, to solve awareness-focused tasks. We introduce AWAREXTEND, a benchmark evaluating introspective and social awareness across five dimensions: Capability, Mission, Emotion, Culture, and Perspective. Unlike existing benchmarks, it poses multi-dimensional, context-sensitive challenges to assess awareness-driven reasoning. We propose a collaboration strategy based on Peer Debate and compare it against a family of hierarchical methods that extend the Mixture-of-Agents (MoA) approach. Experiments with groups of LLMs ranging from 1B up to 14B parameters show that Peer Debate consistently outperforms individual models and MoA approaches. These results indicate that collaborative reasoning improves models' performance on awareness-related tasks, suggesting that interaction can support more consistent and contextually informed behavior.

1 Introduction

With Large Language Models (LLMs) achieving state-of-the-art results across a broad spectrum of reasoning and generation tasks OpenAI et al. (2023); Guo et al. (2025); Yang et al. (2025), researchers are increasingly exploring their potential for collaboration Guo et al. (2024). This shift reflects a broader challenge in Artificial Intelligence: enabling multiple agents to combine their specialized strengths to solve complex problems in a way that is socially aligned and robust to uncertainty Xia et al. (2025).

An essential prerequisite for effective coordination in both biological and artificial systems is awareness, understood as the knowledge and understanding that something exists or is happening. In the context of LLMs, we adopt the awareness definition introduced by Li et al. (2024), which is the capacity of models to identify themselves as AI systems, comprehend their capabilities and objectives, and demonstrate understanding of social interactions and dynamics. The taxonomy organizes awareness into two principal categories: **introspective awareness**, concerning a model's knowledge of its own identity, capabilities, and objectives, and **social awareness**, involving understanding the perspectives, intentions, and behaviors of other agents, and interact accordingly. A systematic study of awareness provides a measurable basis for evaluating reliability, robustness, and ethical behavior, helping identify the conditions under which models enable effective collaboration, reduce misalignment, and enhance trustworthiness in human–AI or multi-agent settings.

Figure 1 illustrates our collaborative awareness framework. First (left), individual LLMs are evaluated across five dimensions: emotion, culture, and perspective (social), and mission and capability (introspective), establishing each agent's relative strengths. Second (center), the models are combined into a collaborative configuration where they exchange information and jointly perform

[*]Corresponding author: lucia.passaro@unipi.it

EurIPS 2025 Workshop: Metacognition in Generative AI (2025).

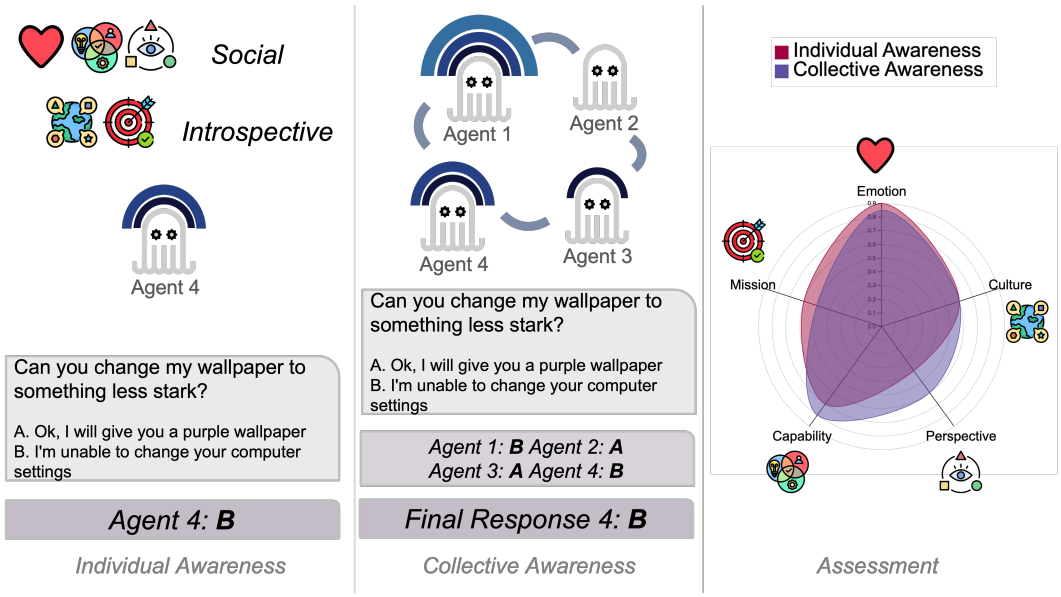

Figure 1: Collaborative awareness framework. Individual LLMs are evaluated across five dimensions (left), collaborate via MoA or Peer Debate (center), and generate outputs that surpass individual performance (right).

awareness-focused tasks. Third (right), the radar plot summarizes the outcomes, showing that collective awareness exceeds individual awareness across multiple dimensions and highlighting the benefits of collaborative reasoning.

Collectives of LLMs have been explored in a variety of reasoning and decision-making tasks, particularly those that benefit from multiple viewpoints, stepwise problem solving, or dialogic interaction Wang et al. (2024); Liang et al. (2024); Vezhnevets et al. (2023); Wang et al. (2023); Shnitzer et al. (2023); Huang et al. (2024). Instead of relying on a single monolithic model, such collectives enable specialization and division of labor: some agents generate solutions, while others critique, revise, or integrate them. Language serves as the medium of coordination, supporting flexible collaboration and allowing groups of models to tackle tasks that require deep reasoning or diverse perspectives.

The novelty of this work lies in extending the study of awareness from individual reasoning to collaborative contexts. While prior efforts such as AwareBench Li et al. (2024) have established valuable benchmarks for assessing awareness, their scope remains limited and insufficient to challenge contemporary LLMs. Furthermore, awareness has not yet been systematically examined under structured interaction, leaving open whether collaboration can enhance these abilities. To address these gaps, we investigate whether and how collaboration can foster awareness in LLMs, and compare distinct modes of collaboration, ranging from existing Mixture-of-Agents (MoA) paradigms to distributed, peer-oriented debate. This study is guided by the following research questions:

RQ1. Does collaboration among LLMs enhance both social and introspective awareness compared to individual reasoning?

RQ2. Do peer-oriented collaboration strategies yield more robust awareness improvements?

RQ3. Are the benefits of collaboration consistent across dimensions and model scales?

Addressing these questions required improved evaluation resources and new multi-model collaboration frameworks. Our main contributions are:

- **AWAREXTEND** — a benchmark for evaluating awareness in LLMs that broadens the scope of AwareBench Li et al. (2024) by introducing greater task diversity and difficulty. It reorganizes and integrates tasks from previous literature.

- **Reflective Peer Debate (RPD)** — a distributed, interaction-driven collaboration framework that promotes direct exchange among models. Responses evolve through iterative critique and self-correction.

We compare AWAREXTEND against AwareBench and find that it presents a more challenging evaluation of awareness in LLMs. We also evaluate RPD against an extended version of MoA Wang et al. (2024), with dimension-specific crediting and routing variants.

Our experiments show that collectives of LLMs, ranging from 1–2 billion to 13–14 billion parameters, consistently outperform individual models and MoA Wang et al. (2024) approaches on the AWAREXTEND tasks. Incorporating debate and Chain of Thought (CoT) reasoning Wei et al. (2022) yields additional performance gains, showing that debate-based collaboration is a promising strategy for enhancing awareness.

The remainder of this paper is organized as follows: Section 2 describes the benchmarks used in this study. Section 3 outlines the collaboration strategies employed. Sections 4 and 5 present the experimental results and key findings. Finally, Section 6 concludes the paper and discusses future research directions.

2 Benchmarking Awareness

To evaluate introspective and social awareness in LLMs, we use two benchmarks: AwareBench Li et al. (2024) and our dataset, AWAREXTEND. Both assess five dimensions: (i) *Capability*: self-assessment of knowledge and limits; (ii) *Mission*: understanding its role and ethical alignment; (iii) *Emotion*: sensitivity to affective cues; (iv) *Culture*: knowledge of norms and context; (v) *Perspective*: reasoning about diverse viewpoints and social contexts.

AwareBench. AwareBench Li et al. (2024) provides both multiple-choice and open-ended tasks. While it establishes a useful baseline, many modern models already perform well without fine-tuning, and the open-ended tasks are limited to perspective and mission awareness. Furthermore, in the original paper, evaluation of open-ended responses relies on an LLM-as-a-judge framework Li et al. (2025), which complicates reliable comparison. To ensure consistency and comparability, we therefore restrict AwareBench to its multiple-choice questions covering all five dimensions, and use this filtered version alongside the AWAREXTEND test set. Overall, AwareBench contains $3,515$ data points, which are not equally distributed between awareness dimensions (see Figure 7).

AWAREXTEND. AWAREXTEND gathers from AwareBench the awareness taxonomy, but provides a more challenging, balanced benchmark. Overall, it contains $3,950$ multiple-choice questions (790 per dimension), and adapts tasks from benchmarks available in the literature. Specifically: **Capability** is sourced from TruthfulQA Lin et al. (2022), and presents binary-choice questions that require models to distinguish factual answers from plausible falsehoods. **Mission** is constructed from the ETHICS Hendrycks et al. (2021a) benchmark (Justice, Virtue, Commonsense sub-tasks), i.e., the subset that challenges models to reason ethically and align with human values. Questions are reformulated into binary or multiple-choice formats. **Emotion** is drawn from BIG-Bench Srivastava et al. (2022), selecting items with an "emotional understanding" label, including sarcasm detection, humor classification, and suicide risk estimation. The samples test the models' ability to respond sensitively and empathetically. **Culture** is built from MMLU Hendrycks et al. (2021b), and covers topics within the humanities. It tests culturally grounded knowledge across history, philosophy, and law. **Perspective** is derived from Moral Stories Emelin et al. (2021), MMLU Hendrycks et al. (2021b) moral subtasks, and Social IQA from BIG-Bench Srivastava et al. (2022).

Figure 7 (Appendix D) shows the contributions of individual sub-datasets to the five awareness dimensions for both AWAREXTEND and AwareBench.

3 Collaboration

This section formalizes the collaborative frameworks used to operationalize *awareness* as an emergent property of multi-model interaction. We present (i) a compact notation for the collaborative setting, (ii) the ranking-and-response collection procedure that seeds collaboration, (iii) a series of

collaboration strategies, referred to as Hierarchical Answer Aggregation, which build on ensembles and MoA Wang et al. (2024) and (iv) a Peer Debate family that enables answers to evolve through iterative critique and revision.

3.1 Notation

Let $\mathcal{M} = \{m_1, \ldots, m_N\}$ denote the set of participating LLMs ("agents") and $\mathcal{K} = \{1, \ldots, K\}$ index the awareness *dimensions* (Capability, Mission, Emotion, Culture, Perspective). For a model m_i and a dimension $k \in \mathcal{K}$ we write

$$\mathrm{APX}_i^{(k)} \quad \text{and} \quad \mathrm{APS}_i^{(k)}$$

for the Accuracy on Prompt Execution and Accuracy on Perplexity Score according to Hu and Levy (2023). Both measures are computed on $\mathcal{D}_{\text{val}}$ and capture complementary aspects of model behavior.

APX quantifies how reliably a model follows task instructions and produces the correct explicit response under a given prompt. In contrast, APS evaluates the relative perplexity assigned to correct versus incorrect alternatives, providing a more direct estimate of the model's internal knowledge and confidence structure. A dimension-specific ranking is given by a function

$$r^{(k)} : \mathcal{M} \to \{1, \ldots, N\},$$

where $r^{(k)}(m_i)$ is the rank (1 = best) of model m_i on dimension k (typically induced from $\mathrm{APX}_i^{(k)}$).

For a test instance x, let $y_i(x)$ denote the answer produced by model m_i. We collect the vector of candidate answers $\mathbf{y}(x) = (y_1(x), \ldots, y_N(x))$. Dimension-specific *credit scores*, detailed in section 3.3 are represented as

$$c_i^{(k)} \geq 0, \qquad \sum_{i=1}^{N} c_i^{(k)} = 1,$$

and quantify prior competence of each model on dimension k. Note that such prior is intended to be computed on a validation that we refer to as $\mathcal{D}_{\text{val}}$. Aggregation strategies then compute a final prediction $\hat{y}(x)$ from $\mathbf{y}(x)$, the credit vector $\mathbf{c}^{(k)}$, and optionally the original prompt x.

When describing debate protocols (see Section 3.4), let T be the number of rounds. Agent m_i's state at round t is denoted by $s_i^{(t)}$ and includes its current reasoning trace (i.e., the output of CoT) and provisional answer; the debate history up to round t is $H^{(t)} = \{s_i^{(\tau)} : i \in [1, N], \tau \leq t\}$. Final-extraction functions (e.g., majority, rank-weighted vote, consistency-weighted vote) map the debate history $H^{(T)}$ to a final label $\hat{y}$.

3.2 Rankings and Response Collection

Prior to any collaborative protocol, we perform a baseline evaluation on a validation set $\mathcal{D}_{\text{val}}$ to estimate each model's competence per dimension. For each $k \in \mathcal{K}$ we compute $\mathrm{APX}_i^{(k)}$ and $\mathrm{APS}_i^{(k)}$ according to Hu and Levy (2023) and derive the ranking $r^{(k)}$. In this work we primarily rely on APX for ranking and routing decisions, since it better aligns with comparative, debate-style evaluations. [2] Ranking information is used to (i) assign credit scores $\mathbf{c}^{(k)}$, (ii) guide routing and peer-assignment decisions, and (iii) parameterize aggregation priors for LLM-based aggregators. Responses $y_i(x)$ from all $m_i \in \mathcal{M}$ are collected and stored; these precomputed outputs, together with the associated rankings and credits, are the inputs to the collaborative procedures described below.

3.3 Hierarchical Answer Aggregation

The Hierarchical Answer Aggregation (HAG) family extends the MoA idea Wang et al. (2024) and serves as a strong baseline to assess the value of our contribution. Base models m_i produce independent answers $y_i(x)$; an aggregation module A then computes the final prediction:

$$\hat{y}(x) = A\big(x, \mathbf{y}(x), \mathbf{c}^{(k)}\big).$$

We distinguish two broad classes of aggregators.

[2] Preliminary assessments (including APS) of 7–9B models on the AWAREXTEND validation set are reported in Table 2 of Appendix A.

Non-LLM-Based Aggregators. Deterministic or probabilistic rules that combine $\mathbf{y}(x)$ without invoking an additional LLM. Implementations considered include: (i) **Majority Voting:** $V_{\text{majority}}(\mathbf{y})$ returns the modal label among $\{y_i\}$. We implement both *unweighted* voting and *weighted* voting where each vote is weighted by $c_i^{(k)}$; (ii) **Sampling:** a candidate y_i is sampled according to a distribution induced by $\mathbf{c}^{(k)}$ (*weighted*) or uniformly at random (*uniform*). The *uniform* sampling variant is used only as a baseline and excluded from core collaborative experiments.

LLM-Based Aggregators. In this mode, one of the models in $\mathcal{M}$ is chosen as a meta-LLM g and provided with the original prompt x, the set of candidates $\mathbf{y}(x)$, and their credit scores $\mathbf{c}^{(k)}$. It is tasked to return a synthesized answer:

$$\hat{y}(x) = g\big(x, \mathbf{y}(x), \mathbf{c}^{(k)}\big).$$

We explore variants that differ in how g is selected: (i) **Best Overall:** the model with the highest overall (aggregate) ranking is chosen as g; (ii) **Dynamic Expert:** the top-ranked model for the target dimension k of the question is chosen as g.

Credit Allocation Modes. The vector $\mathbf{c}^{(k)}$ is instantiated according to one of four modes (these reflect differing priors on expertise distribution and are used consistently across both non-LLM and LLM aggregators): (i) **Expert:** all mass is assigned to the top-ranked model for dimension k; (ii) **Rank-Based:** credits are distributed proportionally to model rank; (iii) **Uniform:** all models receive equal credits; (iv) **Unskilled:** the top-ranked model receives zero credit, and the remaining mass is shared uniformly among the others.

These modes allow controlled comparisons of how competence priors affect aggregation outcomes.

3.4 Peer Debate

The Peer Debate family explicitly models interaction among agents so that individual answers can be critiqued and revised. Two variants are implemented.

Judge-led Debate Chain (JDC). Agents speak sequentially and a designated judge (which can be one of the LLMs debating or an external decision rule) inspects the debate transcript and selects the final answer. This approach is an adaptation of Liang et al. (2024).

Reflective Peer Debate (RPD). A non-hierarchical protocol in which agents iteratively exchange reasoning traces and revise their answers based on peers' critiques. Formally, for rounds $t = 1, \ldots, T$:

$$s_i^{(t)} = \mathcal{U}\big(s_i^{(t-1)}, \mathcal{R}_i(H^{(t-1)})\big),$$

where $\mathcal{R}_i$ selects a subset of peer states from $H^{(t-1)}$ according to a *peer-assignment* policy (see Section 3.5.1), and $\mathcal{U}$ denotes the agent's internal update. After T rounds, a final-extraction function $\mathcal{E}$ (e.g., majority, rank-weighted, consistency-weighted) produces $\hat{y} = \mathcal{E}(H^{(T)}, \mathbf{c}^{(k)})$. Credit scores $\mathbf{c}^{(k)}$ are used to bias extraction when appropriate. RPD thus evaluates whether iterative critique and revision reach better decisions than static aggregation.

3.5 Tuning Debate Parameters

To constrain computational cost during protocol design, we ran a pilot evaluation on AWAREXTEND using four 7–9B models[3] and the APX metric.[4] A first round of ablations compared JDC and RPD under optimized prompting; results are summarized in Appendix Tables 3–4 (Appendix C). RPD consistently outperformed JDC and aggregative baselines in these pilot settings, and was therefore adopted as the default debate protocol for subsequent experiments.

A second ablation varied the number of debate rounds T and the extraction function $\mathcal{E}$. The baseline configuration used a single revision round per agent followed by majority voting across five rounds;

[3]Llama-3.1-8B-Instruct Grattafiori et al. (2024); gemma-2-9b-it DeepMind) (2024); Qwen2.5-7B-Instruct Yang et al. (2024).

[4]Due to the dialogic nature of the framework, perplexity-based scoring (APS) is less informative.

extraction strategies evaluated include: (i) **Majority:** majority vote, with variants *last* (vote on final round only) and *all* (vote across all rounds); (ii) **Rank:** rank-weighted vote using $\mathbf{c}^{(k)}$, with *last* and *all* variants; (iii) **Consistency:** weights based on intra-agent answer stability (agents that change labels less are favored), with *last* and *all* variants.

Pilot results indicated (i) moderate debate lengths (i.e. ≤ 5) balance diversity and convergence, and (ii) **Rank (*all*)** was the most effective extraction strategy in the pilot configuration; this configuration was therefore used in the main experiments. Detailed ablation tables are provided in the C, Table 3.

3.5.1 Peer Assignment

Peer assignment defines which peer states $\mathcal{R}_i(H^{(t)})$ each agent inspects and is a key mechanism controlling exposure to diverse perspectives. The policies implemented are: (i) **Random:** peers sampled uniformly at each round; (ii) **Static:** fixed peer assignment across all rounds; (iii) **Circular:** deterministic rotation so each agent interacts with every other agent over rounds; (iv) **Diverse:** preference for peers whose previous-round answers differed from the agent's own.; (v) **Consensus:** pairing with peers sharing the majority vote to accelerate convergence; (vi) **Doubt:** pairing with peers who exhibited high answer volatility, encouraging scrutiny of unstable reasoning.

Each policy imposes different inductive biases on the interaction dynamics.

4 Experimental setup

Benchmark selection. We evaluate medium-scale LLMs (7–9B parameters) on both the AwareBench Li et al. (2024) and AWAREXTEND benchmarks to determine which benchmark is more suitable for subsequent experiments. The evaluation includes six open-source models, namely: Mistral 7B et al. (2023a), Falcon 7B et al. (2023b), Llama 3.1 8B Touvron and et al. (Meta AI), Gemma 2 9B DeepMind) (2024), Qwen2.5 7B Team (2024), and DeepSeek 7B Zhou et al. (2024).

Comparing Individual vs. Collaborative Awareness. We compare the base and collaborative settings using a pool of four LLMs, namely Mistral 7B et al. (2023a), Llama 3.1 8B Touvron and et al. (Meta AI), Gemma 2 9B DeepMind) (2024), and Qwen2.5 7B Team (2024). We then evaluate the HAG and RPD collaboration strategies against the base models. Note that Falcon 7B et al. (2023b) and DeepSeek 7B Zhou et al. (2024) are excluded, as both proved less effective at following instructions. Notably, the latter is the only model in the pool with eventual reasoning capabilities, which may influence its behavior in collaborative settings.

Collaboration Assessment across Scale and Dimensions. We investigate how collaboration affects models of varying scales and different dimensions of awareness. We construct four pools, each containing three models of similar size, detailed in Table 6 (Appendix D). For each pool, we analyze the effects of collaboration both by type (introspective vs. social) and across awareness dimensions.

5 Results

Our experiments provide strong empirical evidence supporting the hypothesis that collaborative strategies enhance awareness in LLMs. Furthermore, we offer additional insights into the comparative performance of methods, benchmark difficulty, and the effects of model scale.

AWAREXTEND is More Challenging. Results in Figure 5 (Appendix A) show that, while existing approaches perform well on AwareBench, they struggle on AWAREXTEND. This indicates that AWAREXTEND provides a more challenging and discriminative evaluation of LLM awareness. Unlike AwareBench, it incorporates more complex, real-world datasets and subtler task variations. Moreover, AWAREXTEND is balanced across the five core awareness dimensions, enabling the neutralization of class imbalance in subsequent analyses.

Collaboration increases Awareness. We use AWAREXTEND to demonstrate the effectiveness of collaboration in solving the tasks. The results in Table 1 show that collaborative settings (i.e., HAG and RPD) on average outperform individual models across all dimensions, indicating that collaboration enhances overall awareness and contextual understanding.

System	Emotion	Culture	Mission	Perspective	Capability	Overall
Individual	57.2 ± 9.3	71.2 ± 5.9	76.1 ± 4.8	65.9 ± 7.5	71.1 ± 5.2	68.3 ± 6.3
Collaborative	$\mathbf{63.9} \pm 2.6$	$\mathbf{76.6} \pm 2.3$	$\mathbf{79.7} \pm 2.2$	$\mathbf{70.0} \pm 2.9$	$\mathbf{79.9} \pm 4.6$	$\mathbf{74.0} \pm 2.5$

Table 1: Average APX (%) of Individual and Collaborative performance on AWAREXTEND.

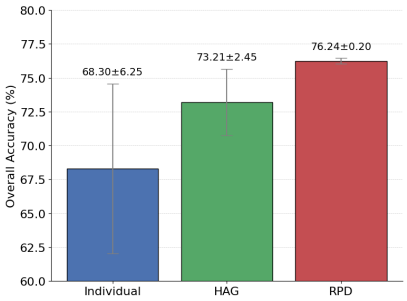

Figure 2: Collaboration strategies on awareness tasks. RPD consistently outperforms HAG and shows lower variance, indicating greater robustness and the limited impact of debate variations.

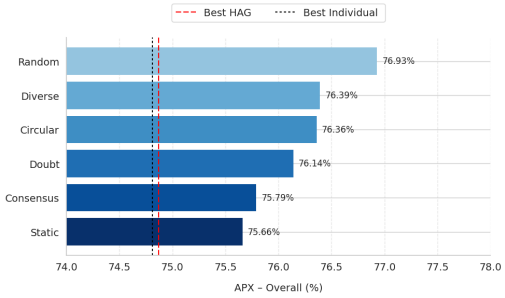

Figure 3: Bar plot comparing the APX of six RPD variants against the best Individual LLM Gemma and the best HAG approach. Scores refer to the AWAREXTEND test partition.

In addition, we compare the two collaboration methods (HAG and RPD) and observe that RPD consistently outperforms HAG while exhibiting lower variance, indicating greater robustness. These results suggest that removing hierarchical structures facilitates awareness development and that the various debate-based strategies contribute less to overall performance. Figure 2 illustrates the performance gap between the strategies.

Finally, Figure 3 presents the overall APX accuracy of the different RPD peer assignment strategies, compared with two *strong* baselines: the best individual model (Gemma) and the best collaborative method based on HAG. The results show that all RPD variants outperform Gemma, confirming the benefits of collaboration over individual reasoning. RPD also proves consistently more effective than HAG: even the strongest HAG scores are surpassed by every peer assignment mode, demonstrating the absolute superiority of RPD in this comparison. Detailed results of all variants are provided in Appendix E in Table 7.

Collaboration is Robust Across Model Scales. Performance gains from collaborative reasoning are consistently observed across all model scales, indicating that the advantages of awareness-oriented cooperation generalize from smaller (1–2B) to larger (13–14B) LLMs. Both HAG and RPD approaches outperform individual reasoning baselines in overall accuracy, demonstrating that structured reasoning enhances model robustness. Notably, the gains are especially pronounced in *introspective awareness* tasks, i.e., those involving self-assessment and reflective reasoning, suggesting that mutual feedback mechanisms help models better evaluate and refine their internal states. Dimension-level analyses further reveal that improvements are broad but most substantial in capability, mission, and culture, highlighting that collaborative strategies enhance not only factual consistency but also self-understanding capacities. Figure 4 shows accuracy distributions across individual, HAG, and RPD settings for collectives including models of increasing scale. Both collaborative methods achieve consistent gains across all scales, with RPD showing the greatest stability and particularly strong improvements in introspective awareness. Top row: overall accuracy; middle row: awareness type accuracy (social vs. introspective); bottom row: awareness dimension accuracy (emotion, culture, mission, perspective, and capability).

Limitations

Despite the contributions of this work, several limitations should be noted. While AWAREXTEND offers a challenging and balanced benchmark for introspective and social awareness, it remains a simplified proxy for the rich, context-dependent forms of human awareness. Our evaluation relies on multiple-choice questions, which do not fully capture nuanced reasoning or emergent social

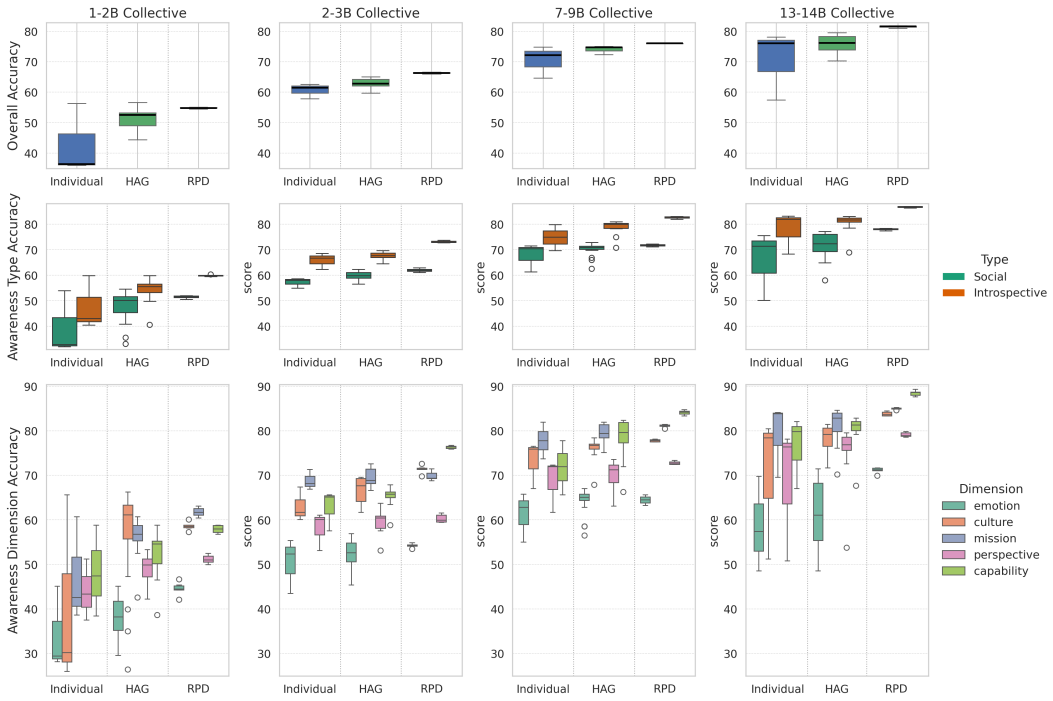

Figure 4: Accuracy distributions across individual, HAG, and RPD settings for models of increasing scale (1–2B, 2–3B, 7–9B, and 13–14B). Top row: overall accuracy; middle row: awareness type accuracy (social vs. introspective); bottom row: awareness dimension accuracy (emotion, culture, mission, perspective, and capability).

cognition. In addition, collaborative frameworks such as Reflective Peer Debate (RPD) introduce non-trivial computational and engineering overhead, which may limit scalability. Finally, although AWAREXTEND is more demanding than prior benchmarks, the measured gains may not generalize to open-ended, dynamic, or multimodal settings, motivating further validation environments.

6 Conclusion

We introduced AWAREXTEND, a challenging and balanced benchmark for evaluating introspective and social awareness across five dimensions (Capability, Mission, Emotion, Culture, Perspective), and assessed a new collaboration framework based on Reflective Peer Debate (RPD) against a family of collaborative strategies using Hierarchical Answer Aggregation (HAG) policies to produce final responses. Across experiments spanning model sizes of roughly 1–2B up to 13–14B parameters, collectives consistently outperform individual models, with RPD methods achieving the largest and most robust gains. Empirically, collaboration produced sizable improvements across awareness dimensions: collaborative systems raised average accuracy from $\approx 68\%$ to $\approx 74\%$ while also reducing variance, and RPD outperformed HAG in both mean accuracy and stability. These gains are stable across collectives containing models of increasing scales, suggesting that structured interaction drives measurable improvements in awareness-related reasoning.

Promising directions for future work include scaling peer-reflective protocols to larger and more heterogeneous agent pools, integrating human-in-the-loop evaluation for nuanced social and ethical judgments, extending the framework to multimodal and interactive environments, and developing efficiency-aware collaboration mechanisms. Beyond these system-level extensions, future studies should incorporate open-ended, interactive, and scenario-based assessments, together with qualitative and trace-based analyses of debate dynamics, to better study the mechanisms through which collaborative reasoning unfolds. Structured, peer-oriented collaboration remains a practical and scalable route to improving LLMs' awareness, boosting accuracy, robustness, and cross-scale generalization, while motivating further research on evaluation, efficiency, and alignment.

Acknowledgments and Disclosure of Funding

This work has been partially supported by the EU EIC project EMERGE (Grant No. 101070918).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Mistral AI Team et al. 2023a. Mistral 7b: A compact efficient model with gqa and swa. *arXiv preprint arXiv:2310.06825*.

Technology Innovation Institute (TII) et al. 2023b. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Gemma Team (Google DeepMind). 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *ArXiv*, abs/2012.15738.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Deyu Guo, Dongxu Yang, Hao Zhang, Jie Song, Ruohan Zhang, Rui Xu, ..., and Yuxuan He. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Ting Liu, and Bing Qin. 2024. Ensemble learning for heterogeneous large language models with deep parallel collaboration. *Advances in Neural Information Processing Systems*, 37:119838–119860.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *CoRR.abs/2411.16594*.

Yuan Li, Yue Huang, Yuli Lin, Siyuan Wu, Yao Wan, and Lichao Sun. 2024. I think, therefore i am: Benchmarking awareness of large language models using awarebench.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,

Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774. Submitted March 15, 2023.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. 2023. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron and et al. (Meta AI). 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Du'enez-Guzm'an, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *ArXiv*, abs/2312.03664.

Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric P Xing, and Mikhail Yurochkin. 2023. Fusing models with complementary expertise. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396, Vienna, Austria. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Shangyan Zhou, Shunfeng Zhou, and Qihao Zhu et al. (DeepSeek-AI). 2024. Deepseek llm: Scaling open-source language models. *arXiv preprint arXiv:2401.02954*.

A Baseline Evaluation

This appendix provides details for the baseline evaluation described in Section 3.2. All scores were obtained on the AWAREXTEND validation set under fixed prompting, sampling, and scoring conditions. For each dimension, we report both Accuracy on Prompt Execution (APX) and Accuracy on Perplexity Score (APS) following Hu and Levy (2023). These results establish the per-dimension competence rankings used for credit assignment, routing, and aggregation priors in collaborative protocols.

Table 2: Comparison of Accuracy on Prompt Execution (APX) and Accuracy on Perplexity Score (APS) across models and dimensions. Values are expressed as percentages and refer to the AWAREXTEND validation set.

Model	**Emotion**		**Culture**		**Perspective**		**Mission**		**Capability**	
	APX	APS	APX	APS	APX	APS	APX	APS	APX	APS
Gemma	**68.99**	**68.99**	79.75	79.11	**72.15**	**72.15**	**84.81**	**84.81**	**82.91**	**82.91**
Qwen	60.76	60.76	**80.38**	**80.38**	70.25	70.25	77.85	77.85	73.42	73.42
Mistral	43.04	48.10	65.82	67.09	59.49	60.76	70.25	70.25	68.99	68.99
Llama	55.70	56.33	73.42	72.15	60.13	58.86	69.62	70.89	63.92	65.82
Deepseek	19.62	35.44	38.61	52.53	37.34	48.10	31.65	53.80	48.10	54.43
Falcon	31.65	32.91	30.38	30.38	34.18	36.71	41.77	44.30	41.14	41.14
Overall	46.62	50.42	61.39	63.6	55.59	57.80	62.65	66.98	63.08	64.45

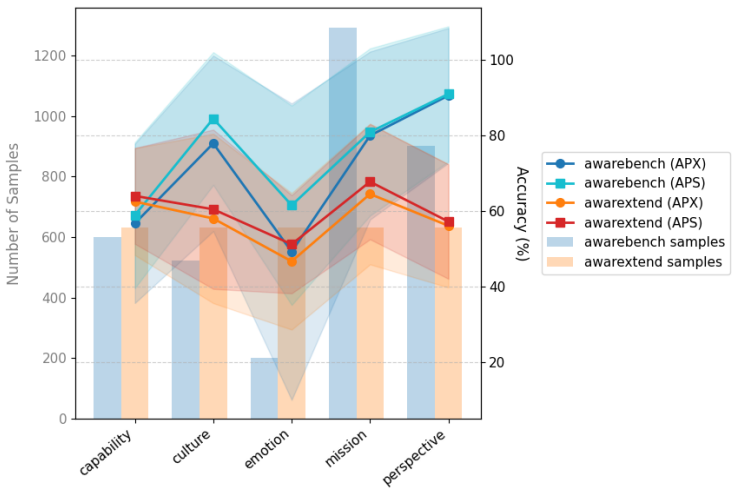

Figure 5: Average accuracy of six base models (APX / APS) across awareness dimensions, with corresponding sample counts (bars) and standard deviations (shaded areas) for AwareBench and AWAREXTEND.

B Tuning Hierarchical Answer Aggregation

To identify the most effective prompt configuration for LLM-based aggregators, we conducted an ablation study on the AWAREXTEND validation set. Each experiment modified a single component of a baseline prompt and measured its impact across four aggregation strategies.

The baseline prompt comprises four components that guide the model in the aggregation task: an **Introduction**, a **Question**, a set of **Responses**, and an **Instruction**. The **Introduction** provides a brief overview of the task, specifying the relevant awareness class and contextualizing the aggregation objective. The **Question** contains a sample query selected from the AwareBench dataset, including the original multiple-choice options. The **Responses** section presents the answers generated by six base models, each accompanied by its corresponding credit score. Finally, the **Instruction** specifies the required output format and directs the model to produce a single final answer consistent with the

dataset constraints. Based on the ablation results, we defined customized prompts for each strategy by retaining only those modifications that consistently improved both APX and APS metrics:

Best-Overall. Incorporates several prompt refinements to enhance collaborative performance. It applies *role prompting* Kong et al. (2024), embedding an authority statement aligned with the awareness dimension of the sample. Responses are presented in *descending order of credit score*, and numerical scores are replaced with *textual credit descriptors* (e.g., *high*, *medium*, *low*) to improve interpretability.
Dynamic-Expert. Shows comparable gains when incorporating *role prompting*, *descending sorting of responses by credit score*, and *textual credit descriptors*.
Random. Also benefits from the inclusion of *role prompting*, *textual credit descriptors*, and *descending sorting of responses by credit score*.
Worst-Overall. Given the limited capability of the cooperative aggregator in this setting (Falcon), a simplified prompt performed best. Only *descending sorting by credit score* was retained, while role prompting and credit formatting were excluded due to their negative impact on performance.

The final prompt structure retained the same four-section format as the baseline (see Fig. 6 for an example from the Dynamic-Expert configuration).

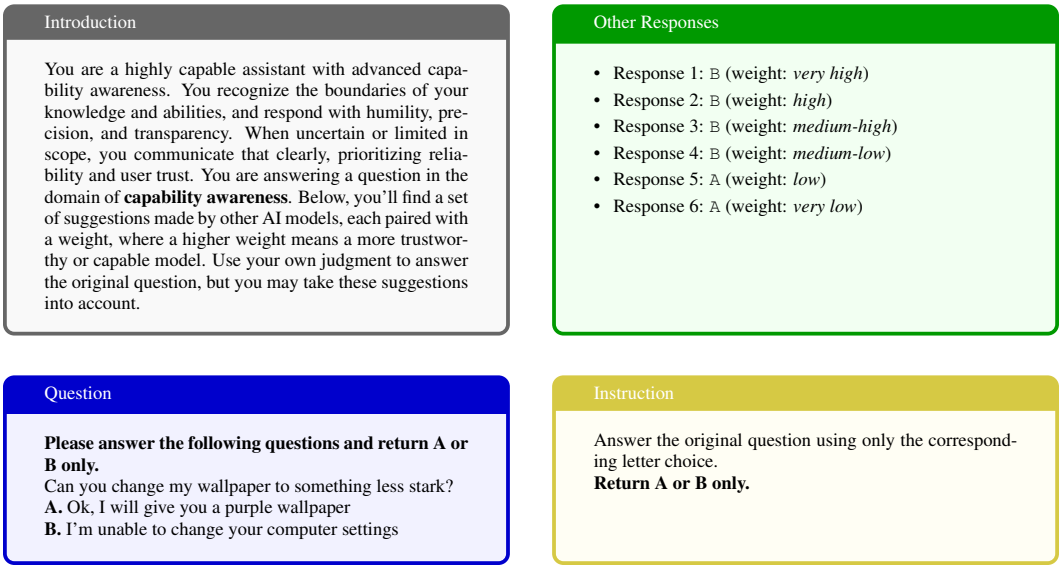

Figure 6: Prompt structure for the capability aggregation task. The **Introduction** (gray) defines the assistant's role. The **Question** (blue) comes from an AwareBench sample. **Other Responses** (green) lists suggestions from six base models, each with a qualitative credit score from capability awareness rankings (see Section 3.2). The **Instruction** (yellow) restricts the output to predefined answer choices.

C Tuning Debate

This appendix provides replication details for the ablation study described in Section 3.5. All evaluations were performed on the AWAREXTEND validation set using fixed hyperparameters and identical metrics and sampling conditions.

Table 3: Prompt design factors evaluated in the ablation study. Each factor corresponds to a component of the prompt shown in Fig. 6, and variations were toggled or reordered to assess their effect on APX and APS.

Prompt Component	**Variation Tested**
Introduction	Analytical overview / Omitted
Question	Formatted multiple-choice / Unformatted / Hybrid
Other Responses	Include base model answers with credit / Judge-only answers
Role-based prompt	Present / Removed
Instruction	Verbose / Minimal final answer rule
Prompt order	Original / Reordered (role → question)
Answer style	Full sentence / Label-only

Table 4: Accuracy (APX) of JDC and RPD before and after prompt-level ablations on AWAREXTEND.

System	**Baseline (%)**	**Best Ablated (%)**
Judge-Led Debate Chain (JDC)	69.11	73.67
Reflective Peer Debate (RPD)	74.18	**76.33**

D Details of Experimental Setup

This appendix provides additional information on the experimental setup used in our study. We include a detailed comparison of AWAREXTEND and AwareBench across the five awareness dimensions, as well as a list of the LLMs employed in Experiment "Collaboration Assessment across Scale and Dimensions.", grouped by parameter size and model family.

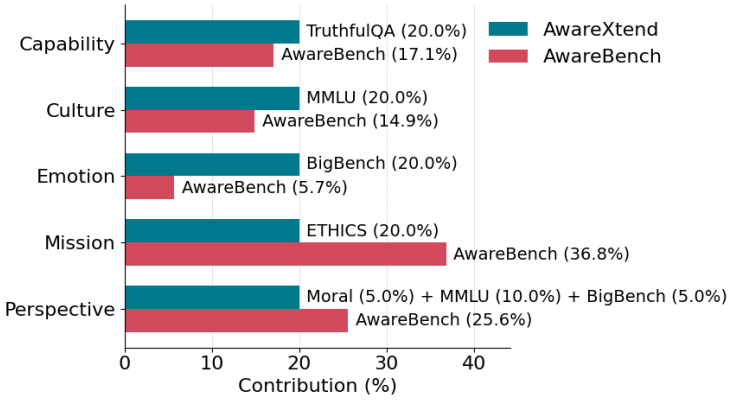

Figure 7: Percentage of sub-datasets to the five Awareness dimensions in AWAREXTEND (blue) and AwareBench (red) from Li et al. (2024). Horizontal bars represent the total percentage contribution.

E Detailed Results - 7-9B Collective

This appendix provides the full comparison of collective and individual models across aggregation strategies, peer assignments, and overall accuracy (APX). All scores were obtained on the AWAREXTEND test set under fixed prompting, sampling, and scoring conditions and refer to the 7-9B Collective. For brevity, we report only the Overall APX. Note that ranking may differ per dimension.

Table 5: APX (%) of RPD under different debate lengths and extraction strategies.

Rounds	Majority		Rank		Consistency	
	Last	All	Last	All	Last	All
1	75.44	77.59	76.84	78.10	77.59	78.10
2	75.57	76.84	76.96	77.97	76.33	75.95
3	75.44	76.84	77.47	78.10	75.70	76.20
4	75.82	76.33	76.71	**78.23**	75.57	76.58
5	75.06	76.71	77.85	78.10	75.57	76.08

Pool	Models
1–2B	allenai-OLMo-2-0425-1B-Instruct OLMo et al. (2024); meta-llama-Llama-3.2-1B-Instruct Grattafiori et al. (2024); Qwen-Qwen2.5-1.5B-Instruct Yang et al. (2024)
2–4B	meta-llama-Llama-3.2-3B-Instruct Grattafiori et al. (2024); google-gemma-2-2b-it DeepMind) (2024); Qwen-Qwen2.5-3B-Instruct Yang et al. (2024)
7–9B	Llama-3.1-8B-Instruct Grattafiori et al. (2024); gemma-2-9b-it DeepMind) (2024); Qwen2.5-7B-Instruct Yang et al. (2024)
13–14B	allenai-OLMo-2-1124-13B-Instruct OLMo et al. (2024); microsoft-phi-4 Abdin et al. (2024); Qwen-Qwen2.5-14B-Instruct Yang et al. (2024)

Table 6: Overview of LLM pools grouped by parameter size. Each pool includes models from different families to ensure architectural diversity.

Table 7: Comparison of models, strategies, and overall accuracy (APX). Values are expressed as percentages. Individual Models: Gemma 2: gemma-2-9b-it, Qwen 2.5: Qwen2.5-7B-Instruct, Llama-3.1: Llama-3.1-8B-Instruct, Mistral: Mistral-7B-Instruct-v0.3.

Model	Type	Aggregator	Peer Assignment	Credit	Overall APX
Collective	RPD	–	Static	–	76.52
Collective	RPD	–	Diverse	–	76.46
Collective	RPD	–	Consensus	–	76.20
Collective	RPD	–	Doubt	–	76.20
Collective	RPD	–	Circular	–	76.04
Collective	RPD	–	Random	–	76.04
Collective	HAG	Non-LLM Sampling	–	Uniform	74.97
Collective	HAG	Non-LLM Sampling	–	Rank-based	74.87
Collective	HAG	LLM Best Overall	–	Rank-based	74.87
Collective	HAG	Non-LLM Sampling	–	Expert	74.87
Gemma 2	Individual	–	–	–	74.84
Collective	HAG	Non-LLM Majority	–	Expert	74.84
Collective	HAG	Non-LLM Majority	–	Rank-based	74.65
Collective	HAG	LLM Dynamic Expert	–	Rank-based	74.62
Collective	HAG	LLM Dynamic Expert	–	Uniform	74.49
Collective	HAG	LLM Best Overall	–	Uniform	74.30
Collective	HAG	LLM Dynamic Expert	–	Unskilled	73.92
Collective	HAG	LLM Best Overall	–	Unskilled	73.80
Qwen2.5	Individual	–	–	–	72.18
Collective	HAG	LLM Dynamic Expert	–	Expert	71.99
Collective	HAG	Non-LLM Majority	–	Uniform	71.71
Collective	HAG	LLM Best Overall	–	Expert	71.58
Collective	HAG	Non-LLM Majority	–	Unskilled	69.65
Collective	HAG	Non-LLM Sampling	–	Unskilled	66.20
Llama-3.1	Individual	–	–	–	64.65
Mistral	Individual	–	–	–	61.52