Epistemically-Constrained Belief Harmonization: Integrating Symbolic Structure with Probabilistic Consensus

Adam Kostka[1], Jarosław A. Chudziak[1]

[1]Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland

Abstract

Aggregating inconsistent judgments from multiple reasoners is a long-standing problem in artificial intelligence, with mature foundations in consensus theory, belief pooling, and epistemic logic. The topic has gained renewed significance with the rise of LLM ensembles, where the judgments of specialist LLMs, retrieval-augmented agents, and judge models often diverge even on factual matters. The core problem is that such divergent judgments usually reflect differences in the information sources accessed, not random noise. How can an aggregation rule recognize this distinction and reconcile beliefs without forcing agents that cannot represent the same fine-grained truth to converge? In this paper, we propose a symbolic-probabilistic framework that unites symbolic epistemic structure with probabilistic belief dynamics. Agents maintain probability distributions constrained by their epistemic partitions, while a critic agent computes a dynamic target belief and specialists shift only toward epistemic projections of that target. The framework supports a convergence guarantee and provides a diagnostic interpretation of the remaining disagreement. The contribution is a concise symbolic-statistical approach to feasible consensus in settings where agents cannot necessarily represent the same knowledge.

Keywords

epistemic partitions, belief harmonization, symbolic-statistical integration, multi-agent LLM systems, Jensen-Shannon divergence

1. Introduction

Modern AI pipelines increasingly rely on ensembles of reasoning components. A single decision may synthesize outputs from a retrieval-augmented Large Language Model (LLM), a tool-using agent, and a domain-specialized judge model [1, 2]. These reasoners solve the same problem, but differ in the knowledge they can access, the distinctions they can represent, and the examples on which they were trained. The combined output looks like a distribution over hypotheses, but the individual probabilities come from reasoners that are not interchangeable.

There are two mature approaches which handle parts of the problem. Statistical consensus and opinion pooling [3, 4] reconcile numerical states, assuming their homogeneity, while epistemic logic and belief revision [5, 6] define what an agent can represent and how it can revise its beliefs. The resulting question is this: how can an aggregation rule distinguish reducible disagreement from disagreement caused by representational limits, reconcile beliefs in the reducible case, and still leave room for disagreement in the latter case despite harmonization? Recent results in neuro-symbolic integration [7] and in the internal beliefs of multi-agent systems powered by LLMs [8] make progress towards this problem in exactly this sense.

This problem can be solved succinctly through Epistemically-Constrained Belief Harmonization. Agents have probabilistic representations over a common space of possible worlds, while each agent has a symbolic epistemic partition over the worlds it cannot distinguish. Critic agents compute a dynamic target belief from the overall state, and specialists harmonize towards a projection of this target belief within their own epistemic partitions, where the strength of updating is controlled by an adaptive

Joint Workshop on Statistics and Knowledge Integration for Logic, Learning, Ethical Decisions, and LLMs (SKILLED-LLMs 2026), co-located with FLoC 2026, Lisbon, Portugal

adam.kostka.stud@pw.edu.pl (A. Kostka); jaroslaw.chudziak@pw.edu.pl (J. A. Chudziak)

0009-0004-8174-2109 (A. Kostka); 0000-0003-4534-8652 (J. A. Chudziak)

trust weight that decreases with Jensen-Shannon divergence. The projection is the symbolic-statistical interface, and all updates satisfy the agent's representational constraints.

There are three key contributions. First, we provide a formal model in which symbolic epistemic structure imposes constraints on probabilistic belief dynamics in a heterogeneous multi-agent setting. Second, we prove that beliefs are guaranteed to converge to a bounded region around each agent's projected target, so agents seek harmony only to the extent that is representationally feasible. Third, we demonstrate our model with two formalized examples inspired by large language models. The first addresses consensus among specialists with different emphases in a review process, while the second diagnoses disagreements caused by differing knowledge cutoffs.

2. The Formal Model

The model consists of three components. The symbolic component specifies an agent's capability to represent possible worlds through a partition. The statistical component represents an agent's beliefs as a probability distribution constrained by that partition. The dynamic component lets a critic set a target belief that each specialist tracks.

2.1. Worlds, Partitions, and Admissible Beliefs

Let $W = \{w_0, \ldots, w_{m-1}\}$ be a finite set of possible worlds and let us consider n specialists $S_1, \ldots, S_n$ and a critic C. Each agent a is linked to an equivalence relation $R_a \subseteq W \times W$ partitioning W into information cells, such that $(w, w') \in R_a$ means that agent a does not distinguish between worlds w and w'. This follows the standards of Kripke semantics [5], reflecting the heterogeneity of representational capabilities without specifying its origin. Each agent also holds a probability distribution $\mu_a : W \to [0, 1]$. These distributions must be compatible with the partition, which gives the key notion of epistemic constraint.

Definition 1 (Epistemically constrained belief)**.** *A distribution μ_a is epistemically constrained by R_a if for all $w, w' \in W$, $(w, w') \in R_a$ implies $\mu_a(w) = \mu_a(w')$.*

2.2. Projection as the Symbolic-Statistical Bridge

The projection operation maps any probability distribution to an admissible one by averaging probability values within each cell of the agent's partition. Denote by $[w]_a$ the cell of R_a containing w. The definition of the projection is $\mathcal{P}_a(\mu)(w) = |[w]_a|^{-1} \sum_{w' \in [w]_a} \mu(w')$. This preserves the total probability mass of each cell while distributing that cell's mass uniformly across the worlds inside it. Clearly, $\mathcal{P}_a(\mu)$ is epistemically constrained with respect to R_a according to Definition 1, and the projection operation does not increase the distance between probability distributions measured in the L_1 norm, a property to be used later.

The projection operator realizes the symbolic-statistical integration. When the dynamics use an externally specified target distribution, no requirement is imposed on the agent to represent distinctions it cannot perceive. This provides a structural restriction on the update rule, as opposed to a particular weighting scheme for the posterior. Statistical techniques averaging probability distributions across agents generally fail to recognize this boundary. Our setting formally identifies this separation using $\mathcal{P}_a$, while the numerical dynamics determine the subsequent behavior.

2.3. Critic Dynamics and Specialist Update

The dynamics layer operates in discrete iterations. At each iteration t, the critic computes the diagnostic distribution $\mu_C^{(t)}(w) = n^{-1} \sum_j \mu_{S_j}^{(t)}(w)$ from the specialists' current beliefs, and updates its dynamic ideal in three steps. The first is a mixing step, $\hat{\mu}_C^{(t+1)} = \beta \mu_C^{*(t)} + (1 - \beta)\mu_C^{(t)}$, for some $\beta \in [0, 1)$. The

second adds a small uniform perturbation with $\epsilon \in (0, 1)$, and the third projects the result onto the critic's partition,

$$\mu_C^{*(t+1)} = \mathcal{P}_C\Big((1-\epsilon)\hat{\mu}_C^{(t+1)} + \epsilon\nu_{\text{unif}}\Big). \tag{1}$$

The ideal is internal and not imposed from the outside. It serves as a reflective goal based on the present state of the aggregate system, which increases the stability of the model against erroneous external feedback.

Next, each specialist computes an adaptive trust weight, which is a function of the Jensen-Shannon distance [9, 10] between its subjective distribution and the critic's diagnostic, given by $\alpha_j^{(t)} = \exp(-\kappa \mathrm{JS}(\mu_{S_j}^{(t)} \| \mu_C^{(t)}))$ with $\kappa > 0$. Specialists whose beliefs resemble the aggregate receive larger update weights and move more strongly toward their projected target, while highly divergent specialists update more cautiously. Every specialist always retains a positive weight, so none can ever be frozen.

$$\mu_{S_j}^{(t+1)} = (1-\alpha_j^{(t)})\mu_{S_j}^{(t)} + \alpha_j^{(t)}\mathcal{P}_j\Big(\mu_C^{*(t)}\Big). \tag{2}$$

A convex combination of two distributions that are each epistemically constrained by R_{S_j} is itself constrained, so the update preserves the symbolic layer by construction. The specialist moves only toward targets it can represent.

3. Symbolic-Statistical Integration for LLM Ensembles

With the formal layer defined, the model can be illustrated through the paradigm of current LLM ensembles. The aim is neither to formalize transformers nor to explain their inner workings, but to offer language and an inference rule applicable to a set of reasoning agents, based on emerging literature on LLM collaboration [1], judge-based LLM systems [2, 11], and LLM-based debates [12, 13].

Table 1 summarizes how each formal concept maps onto corresponding entities in current LLM pipelines. Possible worlds form the common hypothesis space that an LLM ensemble reasons about. A specialist's partition captures the structural distinctions it can represent, constrained by its prompt, tools, retrieved corpus, or knowledge cutoff. The distribution represents its confidence over the possible worlds. The critic corresponds to an LLM acting as a coordinator or judge [12, 2], which forms a dynamic target passed to specialists through the projection operator $\mathcal{P}_j$.

Table 1
Mapping between formal objects and LLM ensemble components.

Formal object	LLM ensemble interpretation
Possible worlds W	Shared hypothesis or answer space
Partition R_{S_j}	Representational limits of an LLM agent
Distribution μ_{S_j}	Normalized confidence over hypotheses
Critic C	Coordinator or judge model
Dynamic ideal $\mu_C^{*(t)}$	Current target synthesized by the critic
Projection $\mathcal{P}_j$	Restriction to the agent's representable distinctions
Trust weight $\alpha_j^{(t)}$	Divergence-sensitive reliance on specialist j

Consider three language model agents with different retrieval capacities that assess the feasibility of a proposed therapy. Agent A has access to contemporary clinical trial registers, whereas Agent B relies on its own training data collected two years ago. Agent C has access to a database of drug interactions but lacks outcome reports. Their disagreements are not simply stochastic. Agent B cannot discern results that occurred after its cutoff, which is exactly the kind of restriction imposed by a knowledge-cutoff partition [14]. Opinion pooling via linear or weighted averaging would collapse their views into a single distribution shared by all agents, while the projected update approach requires each agent to adopt only the part of the critic's target that its partition allows it to represent.

In practice, such a framework can serve as a meta-level diagnostic layer within an LLM ensemble. A designer first fixes a finite hypothesis space, for example candidate answers, diagnoses, or review outcomes. The designer then assigns partitions by asking which worlds each agent can distinguish given its prompt, retrieval corpus, tools, or knowledge cutoff. Scores from specialists are normalized into distributions over W by calibration or by a task-specific scoring rubric. The update rule then compares two quantities: divergence from the projected target, which measures convergence within the agent's capacity, and divergence from the raw target, which measures the remaining gap caused by missing distinctions.

This procedure does not require access to transformer internals. It requires only a finite set of hypotheses, interpretable capability constraints, and comparable confidence scores. If the projected divergence is high, more harmonization or better calibration is needed. If the projected divergence is low but the raw divergence remains high, the issue is structural rather than behavioral, and the appropriate intervention is to change the agent's tools, retrieval access, or task decomposition [15, 16].

4. Formalized Coordination Scenarios

The model is illustrated through two scenarios set in an LLM ensemble context. Each scenario specifies a finite hypothesis space, partition assignments, initial beliefs, and 30 iterations of the dynamics. Scenario A uses an explicit eight-world space for three review aspects. Scenario B uses a four-world technology-assessment space and shows a disagreement that cannot be removed by harmonization alone, because one agent is unable to represent the critic's raw target.

4.1. Scenario A: Multi-Aspect Review Harmonization

In this scenario, we model a peer review process with three LLM specialists, each examining the proposal from a different aspect [17]. The hypothesis space is the full Boolean product of novelty (N), methodology (M), and clarity (C), giving eight worlds $w_{111}, w_{110}, w_{101}, w_{011}, w_{100}, w_{010}, w_{001}, w_{000}$. For example, w_{111} denotes a proposal strong on all three aspects, while w_{000} denotes one weak on all three. Specialist S_1 distinguishes high from low novelty, S_2 distinguishes high from low methodology, and S_3 distinguishes high from low clarity. The critic uses the identity partition over all eight worlds.

Initially, each specialist assigns probability 0.8 uniformly to the worlds in its high-quality block and 0.2 uniformly to the complementary block. The critic's initial ideal assigns probability 0.4 to w_{111} and distributes the remaining mass uniformly. After 30 iterations with $\beta = 0.7$, $\kappa = 1.2$, and $\epsilon = 0.01$, all three specialists reach projected Jensen-Shannon divergence 3.36×10^{-9} from their projected targets. Their raw divergence from the critic's ideal is 1.33×10^{-7}. Thus, the system reaches agreement up to the distinctions each specialist can represent while retaining a full eight-world hypothesis space rather than collapsing the three aspects into a smaller ad hoc coding.

4.2. Scenario B: Diagnosing Knowledge-Cutoff Asymmetry

Scenario B illustrates the utility of the framework's diagnostic tools when one agent is subject to a structural epistemic constraint. The hypothesis space contains four possible worlds corresponding to an evaluation of an innovative technology. World w_0 corresponds to confirmation of viability by recent data, w_1 corresponds to lack of certainty, w_2 indicates that past theoretical insights had already pointed to problems with the technology, and w_3 stands for negative side effects revealed by recent data. Agent S_1 models a language system with an outdated knowledge cutoff, and therefore cannot discriminate worlds w_0, w_1, and w_3, giving the partition $\{\{w_0, w_1, w_3\}, \{w_2\}\}$. Agent S_2 and the critic have the full partition. Agent S_3 plays the role of a cautious intermediary with partition $\{\{w_0, w_1\}, \{w_2, w_3\}\}$. This scenario uses $\beta = 0.85$, $\kappa = 1.8$, and $\epsilon = 0.001$. The critic's initial ideal is $(0.65, 0.15, 0.10, 0.10)$, reflecting the up-to-date assessment.

Figure 1 shows how the probability assigned to w_0 evolves for each of the agents. The revealing case is Agent S_1. Its belief in w_0 increases and converges, but this agent fails to track the preference for

w_0 exhibited by the critic as well as Agent S_2 does, since its partition enforces equiprobability across w_0, w_1, and w_3. At iteration 30, the Jensen-Shannon divergence between the belief of Agent S_1 and its projected ideal is 3.1×10^{-6}, meaning near-perfect convergence with the ideal representable by Agent S_1. The divergence between Agent S_1 and the critic's raw ideal is approximately 2.9×10^{-3}, three orders of magnitude larger. This should not be seen as a failure to cooperate but as a measure of disagreement induced by the representational constraints of Agent S_1. The numerical value tells system designers that the right course of action is to improve the representational capacity of Agent S_1 rather than pursue further agreement [15, 16].

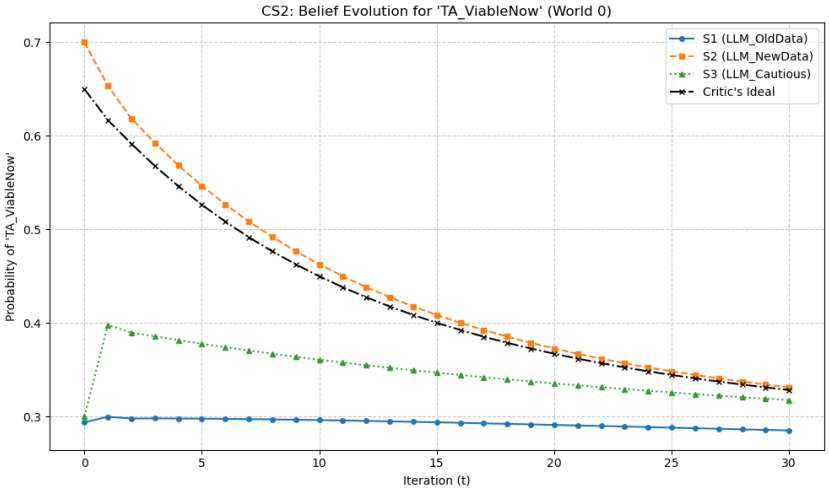

Figure 1: Scenario B: probability of w_0 ("currently viable") over 30 iterations. Specialist S_1 (outdated cutoff) is constrained by its partition and cannot track the critic's preference to the same extent as S_2 (up-to-date). The residual gap between S_1 and the critic's ideal is structural, not a failure of convergence.

4.3. Comparison with Simple Baselines

The experiments compare the proposed dynamics with static linear pooling, complete-graph DeGroot averaging, and no harmonization. In Scenario A, the proposed update reaches maximum projected JS 3.36×10^{-9}, while pooling and DeGroot remain at 1.64×10^{-2} and no harmonization remains at 5.06×10^{-2}. In Scenario B, the corresponding values are 6.55×10^{-6}, 1.29×10^{-2}, and 7.52×10^{-2}. Pooling and DeGroot also produce shared distributions that are not generally admissible for each specialist's partition. A 48-point grid over β, κ, and ϵ for Scenario B kept the maximum projected divergence below 8.89×10^{-6}.

Together, the two scenarios highlight the dual nature of the model. In Scenario A, it is shown that heterogeneous but symmetric representational constraints lead to agreement over an explicit eight-world space, whereas Scenario B illustrates how asymmetric structures create persistent disagreement that the model can measure and explain.

5. Convergence Guarantee

The scenarios of Section 4 exemplify a fundamental feature of the dynamics in which each specialist eventually follows its own projected target trajectory. In this section, the phenomenon is stated as a convergence theorem. The result has the usual Lyapunov-style form of target tracking, but with an agent-specific target and with heterogeneous partitions retained.

We define the tracking error of specialist j at time t by

$$D_j^{(t)} \;=\; \mathrm{JS}\Big(\mu_{S_j}^{(t)} \,\|\, \mathcal{P}_j\Big(\mu_C^{*(t)}\Big)\Big), \qquad D^{(t)} \;=\; \max_j D_j^{(t)}. \tag{3}$$

The theorem establishes a result regarding the Jensen-Shannon divergence between each specialist and the associated projected target, maximized over all specialists, asymptotically. It states that the maximum over specialists of this divergence stays, in the long run, within a bounded neighborhood determined by the parameters of the dynamics and the size of the hypothesis space.

Theorem 1 (Convergence to a neighborhood of the projected ideal)**.** *Consider the dynamics of Section 2 with finite W, $\beta \in [0, 1)$, $\epsilon \in (0, 1)$, and $\kappa > 0$. There exists a constant $B \geq 0$ depending on these parameters and on $|W|$ such that*

$$\limsup_{t \to \infty} D^{(t)} \;\leq\; B.$$

Moreover, if $\mu_C^{*(t)}$ converges so that its step-wise drift tends to zero, then $D^{(t)} \to 0$ and convergence is pointwise.

The proof applies a Lyapunov function-like technique with $D^{(t)}$ as the potential, alongside an upper bound on the drift of the dynamic ideal and a lower bound on trust weights. The key arguments and expression of B are provided in Appendix A.

The theorem has an intuitive meaning behind the dynamics. The model does not enforce unrealistic synchronization between agents. Instead, each specialist aligns itself with the component of the target which it can capture, with the tracking error to the projected target bounded by the drift of the ideal. The gap from the raw target may persist when it reflects distinctions outside the specialist's partition. In practice, this offers a rigorous notion of long-run equilibrium against which system engineers can benchmark real-world trajectories.

6. Related Work

Our proposal sits at the intersection of statistical consensus, epistemic logic, and modern LLM coordination. Consensus algorithms are used to formally describe the process of reaching agreement between agents [3, 4, 18], and linear opinion pools provide probabilistic generalizations of such schemes. In all such cases, there is usually some kind of shared representational capacity among the agents involved, so the disagreements are seen as noise to average out. Our proposed projection operator differs from this tradition by allocating individual projections of the shared target to each agent, while the endogenous critic distinguishes our work from target-oriented consensus.

Formal reasoning and belief revision provide standard semantics for agents' epistemic capacities [5, 19], and specify how an agent's epistemic state can be updated upon receiving new information [6, 20]. Dynamic epistemic logic describes how agents' epistemic structures respond to external events [21]. Aumann's agreement theorem [22] makes a strong structural-symmetry assumption. All of these ideas influence our symbolic layer but do not specify how probability mass should evolve across agents over time. Our contribution lies in combining the projection operator with the reflective critic, which translates qualitative indistinguishability into a quantitative restriction on the belief dynamics.

A growing body of work integrates the outputs of multiple LLMs [1]. Techniques such as ChatEval [12], multi-agent debate systems [13], and LLM-as-a-judge workflows [2, 11] are extremely effective in practice but often lack a theoretical explanation for their workings. Ongoing efforts to synergize logical reasoning with knowledge management in multi-agent LLM systems [23] and to evaluate beliefs in such systems [8] underscore the need for formal tools that describe the internal epistemic structure of heterogeneous reasoners. Our model introduces formal machinery to coordinate such reasoners and is compatible with neuro-symbolic integration [7], which uses symbolic constraints to regulate statistical operations. Our framework does not replace or diminish the use of tool-use and retrieval augmentation techniques [15, 16], but can be combined with them as a diagnostic layer.

7. Limitations, Outlook, and Conclusion

The proposed framework is formal and diagnostic in nature. Translating concrete LLM outputs into calibrated probability distributions relies on calibration techniques outside the scope of this paper, and extracting realistic partitions from an agent's prompts, tools, or retrieval abilities remains a nontrivial engineering problem. The scaling factor B depends on the size of the hypothesis space, so the framework is most suitable for coordination within structured, finite hypothesis spaces, where scalability may require approximation or factored partitions.

The main result of this paper is that symbolic and statistical techniques can be combined at the level of update rules, not only representations. The benefit is not just an additional aggregation technique but a formal diagnosis that separates disagreement to be reconciled from disagreement that signals unavailable distinctions. Potential follow-up directions include estimating partitions from observations of an agent, applying the framework to real-time LLM ensembles as a diagnostic layer, and generalizing the critic to more complex aggregation rules.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 4.6 in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style, and Formatting assistance. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, H. D. Nguyen, Multi-agent collaboration mechanisms: A survey of LLMs, arXiv preprint arXiv:2501.06322 (2025). doi:10.48550/arXiv.2501.06322.

[2] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, B. Zhang, L. Ni, W. Gao, Y. Wang, J. Guo, A survey on LLM-as-a-judge, The Innovation 7 (2026) 101253. doi:10.1016/j.xinn.2025.101253.

[3] M. H. DeGroot, Reaching a consensus, Journal of the American Statistical Association 69 (1974) 118–121. doi:10.1080/01621459.1974.10480137.

[4] R. Olfati-Saber, J. A. Fax, R. M. Murray, Consensus and cooperation in networked multi-agent systems, Proceedings of the IEEE 95 (2007) 215–233. doi:10.1109/JPROC.2006.887293.

[5] R. Fagin, J. Y. Halpern, Y. Moses, M. Y. Vardi, Reasoning About Knowledge, MIT Press, Cambridge, MA, 1995.

[6] A. Darwiche, J. Pearl, On the logic of iterated belief revision, Artificial Intelligence 89 (1997) 1–29. doi:10.1016/S0004-3702(96)00038-0.

[7] B. C. Colelough, W. Regli, Neuro-symbolic AI in 2024: A systematic review, arXiv preprint arXiv:2501.05435 (2025). doi:10.48550/arXiv.2501.05435.

[8] A. Kostka, J. A. Chudziak, Evaluating theory of mind and internal beliefs in LLM-based multi-agent systems, in: Computational Collective Intelligence. ICCCI 2025, volume 16138 of *Lecture Notes in Computer Science*, Springer, 2026, pp. 18–32. doi:10.1007/978-3-032-09318-9_2.

[9] J. Lin, Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory 37 (1991) 145–151. doi:10.1109/18.61115.

[10] T. M. Cover, J. A. Thomas, Elements of Information Theory, 2 ed., Wiley-Interscience, 2006.

[11] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, in: Advances in Neural Information Processing Systems, volume 36, 2023. doi:10.52202/075280-2020.

[12] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, ChatEval: Towards better LLM-based evaluators through multi-agent debate, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=FQepisCUWu.
[13] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of the 41st International Conference on Machine Learning, 2024.
[14] B. Fatemi, M. Kazemi, A. Tsitsulin, K. Malkan, J. Yim, J. Palowitch, S. Seo, J. Halcrow, B. Perozzi, Test of time: A benchmark for evaluating LLMs on temporal reasoning, arXiv preprint arXiv:2406.09170 (2024). doi:10.48550/arXiv.2406.09170.
[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 9459–9474.
[16] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, in: Advances in Neural Information Processing Systems, volume 36, 2023. doi:10.52202/075280-2997.
[17] P. Sukpanichnant, A. Rapberger, F. Toni, PeerArg: Argumentative peer review with LLMs, in: Proceedings of the First International Workshop on Next-Generation Language Models for Knowledge Representation and Reasoning (NeLaMKRR), 2024. doi:10.48550/arXiv.2409.16813.
[18] R. Hegselmann, U. Krause, Opinion dynamics and bounded confidence models, analysis and simulation, Journal of Artificial Societies and Social Simulation 5 (2002) 2. URL: https://www.jasss.org/5/3/2.html.
[19] J. Y. Halpern, Reasoning About Uncertainty, MIT Press, Cambridge, MA, 2003.
[20] S. Konieczny, R. Pino Pérez, Merging information under constraints: A logical framework, Journal of Logic and Computation 12 (2002) 773–808. doi:10.1093/logcom/12.5.773.
[21] H. van Ditmarsch, W. van der Hoek, B. Kooi, Dynamic Epistemic Logic, Springer, 2007.
[22] R. J. Aumann, Agreeing to disagree, The Annals of Statistics 4 (1976) 1236–1239. doi:10.1214/aos/1176343654.
[23] A. Kostka, J. A. Chudziak, Synergizing logical reasoning, knowledge management and collaboration in multi-agent LLM system, in: Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, Tokyo University of Foreign Studies, Tokyo, Japan, 2024, pp. 203–212. URL: https://aclanthology.org/2024.paclic-1.19/.

A. Proof Sketch for Theorem 1

The proof of the theorem relies on three key aspects. One, the dynamic ideal satisfies a bound on its per-step drift. Two, the projection mapping is non-expansive in the L_1 norm, which guarantees that the projected ideal maintains the bound. Three, applying a Lyapunov-type contraction to the difference $D^{(t)}$ leads to the desired $\limsup$ bound.

A.1. Bounded Ideal Drift

Let $L_t = \|\mu_C^{*(t+1)} - \mu_C^{*(t)}\|_1$ denote the step-wise L_1 drift of the dynamic ideal. By the definitions in Section 2.3 and Equation (1), and by L_1 non-expansiveness of $\mathcal{P}_C$, the drift satisfies the recurrence

$$L_t \;\le\; (1-\epsilon)\beta\, L_{t-1} \;+\; 2(1-\epsilon)(1-\beta).$$

Since $(1-\epsilon)\beta < 1$, this linear recurrence is contractive and L_t is uniformly bounded by some constant $L \ge 0$ depending on β and ϵ.

By the same non-expansiveness argument, $\|\mathcal{P}_j(\mu_C^{*(t+1)}) - \mathcal{P}_j(\mu_C^{*(t)})\|_1 \le L$ for every specialist j. The projected ideal therefore drifts no faster than the raw ideal.

A.2. Lower Bound on Trust Weights

The Jensen-Shannon divergence has an upper limit of $\ln 2$ between any two probability distributions [9]. The incorporation of the above inequality along with the trust value introduced in Section 2.3 results in

$$\alpha_j^{(t)} \;\geq\; \exp(-\kappa \ln 2) \;=\; 2^{-\kappa} \;=:\; \delta \;>\; 0.$$

In this expression, the minimum value of δ depends only on the predefined constant κ, implying that no specialist will be frozen, regardless of how far it may be from the diagnosis.

The positivity of the trust value is a necessary condition to prove the Lyapunov stability of the algorithm. Otherwise, it would be possible that a specialist is forever stuck far away from its expected target point, thus contradicting our claim. The exponential dependence on κ ensures that δ is positive for any finite value of κ.

A.3. Lyapunov Contraction

Let $Y_j^{(t)} = \mathcal{P}_j(\mu_C^{*(t)})$ and $V_j(t) = \mathrm{JS}(\mu_{S_j}^{(t)} \| Y_j^{(t)})$. By the convexity of Jensen-Shannon divergence in its first argument and the specialist update in Equation (2),

$$V_j(t+1) \;\leq\; (1-\alpha_j^{(t)})\mathrm{JS}(\mu_{S_j}^{(t)} \| Y_j^{(t+1)}) \;+\; \alpha_j^{(t)} \mathrm{JS}(Y_j^{(t)} \| Y_j^{(t+1)}).$$

Using the positivity of the projected ideal (ensured by $\epsilon > 0$), one can bound $\mathrm{JS}(\mu_{S_j}^{(t)} \| Y_j^{(t+1)}) \leq V_j(t) + C_1 L$ for some constant C_1 depending on ϵ and $|W|$, and $\mathrm{JS}(Y_j^{(t)} \| Y_j^{(t+1)}) \leq C_2 L^2$ for some constant C_2. These are standard consequences of information-theoretic inequalities under a positive-support assumption.

Combining the two bounds, and using $\alpha_j^{(t)} \geq \delta$, we obtain a recurrence of the form

$$V_j(t+1) \;\leq\; (1-\delta) V_j(t) \;+\; E,$$

with $E = (1-\delta)C_1 L + \delta C_2 L^2$. Taking the maximum over specialists yields $D^{(t+1)} \leq (1-\delta) D^{(t)} + E$, and the standard analysis of contractive linear recurrences gives $\limsup_{t\to\infty} D^{(t)} \leq B$ with $B = E/\delta$. When the per-step drift L_t itself tends to zero (that is, when the dynamic ideal converges), the same argument applied with time-varying L_t in place of L yields $E_t \to 0$, so $D^{(t)} \to 0$ and convergence becomes pointwise.

A.4. Simulation Parameters

The simulations use 30 iterations and the natural-logarithm Jensen-Shannon divergence. Scenario A uses $\beta = 0.7$, $\kappa = 1.2$, and $\epsilon = 0.01$. Its eight worlds are the Boolean combinations of novelty, methodology, and clarity. Each specialist S_j places probability 0.8 uniformly on the worlds in its high-quality block and 0.2 uniformly on the complementary block, while the critic's initial ideal assigns 0.4 to w_{111} and distributes the remaining mass uniformly. Scenario B uses $\beta = 0.85$, $\kappa = 1.8$, and $\epsilon = 0.001$. Its initial specialist distributions are $(0.25, 0.25, 0.10, 0.25)$ for S_1, $(0.70, 0.10, 0.10, 0.10)$ for S_2, and $(0.30, 0.30, 0.15, 0.25)$ for S_3, followed by projection to the corresponding partitions. The critic's initial ideal is $(0.65, 0.15, 0.10, 0.10)$.