POSTERIOR-LEVEL AGGREGATION FOR MULTI-LLM MULTIPLE-CHOICE DECISION MAKING: A BLACKWELL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid development of large language models (LLMs) has motivated research on answer aggregation in multi-LLM systems, where multiple models contribute predictions toward a shared decision. Existing aggregation approaches, such as voting and debate, are often developed without a common formal account of the information retained by their outputs. In this paper, we analyse multi-LLM answer aggregation (MLAA) using Blackwell's informativeness framework. Within the Blackwell information-structure abstraction, we show that voting and debate induce information structures that are no more informative than the joint private information of all agents. More generally, this result establishes joint private information as a protocol-independent upper bound, while the corresponding ideal decision is based on the posterior conditioned on the joint private information. Motivated by this analysis, we develop a practical framework that estimates option-level distributions from sampled responses and combines them through generalised Bayesian aggregation. The framework is instantiated using product-of-experts and logarithmic opinion-pooling rules. Experiments on six multiple-choice QA benchmarks show that these posterior-level aggregation methods are competitive with or achieve higher accuracy than other multi-LLM voting and debate methods.

1 INTRODUCTION

Recent advances in large language models (LLMs) have enabled the development of autonomous intelligent agents that can interact with their environment, plan actions, and make decisions with minimal human intervention (Yao et al., 2022). As individual LLM-based agents (Guo et al., 2024) exhibit domain-specific strengths, such as proficiency in programming (Li et al., 2022), biological analysis (Singhal et al., 2023), or mathematical reasoning (Lewkowycz et al., 2022), multi-LLM answer aggregation (MLAA) frameworks have emerged to exploit the complementary capabilities of multiple specialised agents (Du et al., 2023; Li et al., 2024a; Guo et al., 2024). In principle, these frameworks coordinate agents' actions and beliefs to integrate heterogeneous expertise and produce more informed and reliable decisions.

In the context of LLMs, MLAA is commonly realised through voting and debate (Choi et al., 2025a). *Voting*, particularly majority voting, aggregates agents' independent judgments into a collective decision (Wang et al., 2023). While computationally efficient and simple, it disregards variation in agent expertise, which may reduce informativeness and lead to suboptimal decisions. In contrast, *multi-agent debate* (MAD) (Chan et al., 2024) enables agents to iteratively exchange perspectives and update their beliefs through structured, multi-round discussions, with the goal of reaching a more coherent and higher-quality collective decision. Despite this promise, recent empirical studies (Choi et al., 2025a) indicate that multi-round debate is often less reliable and less effective than simple voting-based approaches. MAD systems are shown to frequently converge toward majority-held misconceptions rather than correct reasoning (Estornell & Liu, 2024). In homogeneous settings, straightforward ensemble techniques often outperform multi-round debate (Choi et al., 2025a). MAD is also reported to fail to surpass even single-agent baselines (Zhang et al., 2025). More broadly, LLMs exhibit substantial performance degradation in multi-turn settings regardless of protocol (Laban et al., 2026), suggesting the limitation is not unique to debate. These observations motivate studying when multi-LLM systems fail to extract all available information, regardless of protocol.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we analyse MLAA through Blackwell's informativeness framework (Blackwell, 1953). This perspective allows us to compare voting, debate, and posterior-level aggregation in terms of information structures. Our contributions are summarised as follows:

- We formulate MLAA as a Bayesian statistical decision problem and use Blackwell's ordering to show that any protocol output is a garbling of the agents' joint private signals. This establishes a protocol-independent information upper bound shared by practical aggregation protocols.
- For zero-one classification loss, we identify the posterior conditioned on the joint signals as the ideal decision variable. We then estimate per-agent option distributions from sampled responses using semantic scoring and permutation-equivariant calibration.
- We evaluate established multiplicative posterior-pooling rules within this common pipeline across six benchmarks and several model-pool configurations. The results show that posterior-level aggregation is competitive with the evaluated voting and debate baselines.

2 BACKGROUND

This section provides the background on the Blackwell's informativeness framework (Blackwell, 1953; de Oliveira, 2018). To increase the readability, we use calligraphic letters to denote sets, lowercase letters to represent variables, and Greek letters to indicate functions.

Definition 1 (Decision-making problem)**.** *A decision making problem is denoted as a tuple of* $(\mathcal{S}, \mathcal{A}, \varphi, \rho)$*, where* $\mathcal{S}$ *is the set of possible states,* $\mathcal{A}$ *is the set of possible actions,* $\varphi : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ *is a utility function, and* $\rho : \mathcal{S} \to [0, 1]$ *is a prior distribution over states (i.e.,* $\sum_{s\in\mathcal{S}} \rho(s) = 1$*).*

Definition 2 (Information structure)**.** *An information structure is a tuple* $(\mathcal{D}, \sigma)$*, where* $\mathcal{D}$ *denotes the space of observed signal, and* $\sigma : \mathcal{S} \to [0, 1]$ *represents the conditional distribution:* $\sigma(d|s) = \Pr(d|s)$*, with* $d \in \mathcal{D}, s \in \mathcal{S}$*.*

Given the likelihood $\sigma(d|s)$ of the observed data $d \in \mathcal{D}$ and the prior over states $\rho(s)$, the posterior of the state can be obtained via Bayes' rule as $\Pr(s|d) = {}^{\sigma(d|s)\rho(s)}/_{\Pr(d)}$, where $\Pr(d)$ is the prior distribution of observed data d.

Definition 3 (Expected value of information structure)**.** *The expected value of an information structure for the decision maker is defined as the expected value of the utility function under the optimal action over states and observations:*

$$\omega(\sigma) = \sum_{d\in\mathcal{D}} \Pr(d) \underbrace{\left[\max_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \varphi(a, s) \Pr(s|d)\right]}_{\textit{optimal utility conditioned on observed data}} = \sum_{d\in\mathcal{D}} \max_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \varphi(a, s)\, \sigma(d|s)\, \rho(s). \quad (1)$$

Intuitively, the *expected value* of an information structure can be interpreted as follows:

1. for each observation d, select the action to maximise the expectation of the utility w.r.t. $\Pr(s|d)$,
2. weight that utility by the probability of observing d,
3. sum (or integrate) over all observations.

Definition 4 (Garbling)**.** *An information structure* $(\mathcal{D}', \sigma')$ *is a* garbling *of* $(\mathcal{D}, \sigma)$*, denoted as:* $(\mathcal{D}', \sigma') \trianglelefteq (\mathcal{D}, \sigma)$*, if there exists a function* $\kappa : \mathcal{D} \times \mathcal{D}' \to [0, 1]$ *such that:* $\sigma'(d'|s) = \sum_{d\in\mathcal{D}} \kappa(d, d')\, \sigma(d|s)$ *and* $\sum_{d'\in\mathcal{D}'} \kappa(d, d') = 1$*.*

Intuitively, garbling introduces a stochastic transformation that adds noise to the original signal distribution. Consequently, the expected utility for any decision problem cannot increase under garbling; instead, it can only stay the same or decrease.

Here, we state the "partial" result (excluding the *feasibility*) of the Blackwell's informativeness theorem as follows:

Theorem 1 ("Partial" Blackwell's informativeness (Blackwell, 1953))**.** *If* $(\mathcal{D}, \sigma)$ *and* $(\mathcal{D}', \sigma')$ *are two information structures, the following statements are equivalent:*

1. $(\mathcal{D}', \sigma')$ *is a garbling of* $(\mathcal{D}, \sigma)$*:* $(\mathcal{D}', \sigma') \trianglelefteq (\mathcal{D}, \sigma)$*.*
2. *Every Bayesian agent prefers* $(\mathcal{D}, \sigma)$ *to* $(\mathcal{D}', \sigma')$ *for any decision problem (i.e.,* $\omega(\sigma') \leq \omega(\sigma)$*).*

Intuitively, a "noisier" information structure results in a lower expected utility.

3 PROBLEM STATEMENT

We study MLAA through the Bayesian statistical decision setting (Osborne & Rubinstein, 1994), in which each Bayesian agent starts from a non-informative prior over states, $\rho(s)$, assigning equal probability to all possible outcomes. After receiving private information d_m, each agent updates its belief via Bayesian inference, yielding the posterior $\Pr(s \mid d_m)$. This abstraction closely resembles the multi-LLM setting, where agents share an uninformative prior before training but acquire distinct private information through their training data or observations. Tab. 1 maps the decision-theoretic terms used here to their classification counterparts.

For classification tasks, let $y \in \mathcal{Y}$ represent label of an instance $x \in \mathcal{X}$, and d denote private information (e.g., private training data) observed by an agent. The prediction made by an agent indexed by m, whose private information (or training data) is denoted as d_m, for the instance x can be written as $\Pr(y|x, d_m)$. To relate to the decision-making process in Sec. 2, the label of an instance in classification corresponds to the world state in the Blackwell's informativeness setting (i.e., $\mathcal{S} \equiv \mathcal{Y}$), the action as the predicted label outcome of the system, and the utility is the negative loss function.

Table 1: Mapping table between the terms used in decision-making problem and their interpretation in classification.

Decision theory	**Classification**
state $s \in \mathcal{S}$	true class label
action $a \in \mathcal{A}$	predicted label distribution
utility $\varphi(.,.)$	negative loss function
prior $\rho(s)$	class prior distribution
observation $d \in \mathcal{D}$	training dataset
info. structure $\sigma(d\|s)$	class-conditional distribution

The setting considered in this work is the decision-making with multiple agents in classification, where each agent indexed by m is trained on a private training dataset d_m. The objective is to make an informative decision regarding the label y of a testing instance x. Two prominent approaches are commonly employed in the literature: *(i)* majority voting in which agents independently produce predictions that are subsequently aggregated into a final decision, and *(ii)* debate, where agents exchange their opinions, and update their beliefs accordingly over multiple rounds of interaction.

The objective of this work is to determine whether these two mechanisms, voting and debate, are sufficiently informative under the Blackwell's informativeness framework, or whether an alternative decision-making mechanism may yield a better informative outcome.

4 METHODOLOGY

This section presents our aggregation result and applies it to classification with multiple LLM agents.

4.1 INFORMATION AGGREGATION IS A GARBLING

Theorem 1 is used to formally prove that any aggregation on the private information, denoted as $g \in \mathcal{G} = \{g : g = \gamma(d_{1:M})\}$, is a garbling of $(d_{1:M})$, where $\gamma : \mathcal{D}^M \to \mathcal{Z}$ is an aggregation function, $\mathcal{Z}$ is an arbitrary set, and $d_{1:M}$ denotes $(d_1, \ldots, d_M)$. The following proposition is a direct consequence of the definition of garbling, and the key step that lets us instantiate Blackwell's ordering for multi-LLM answer aggregation.

Proposition 1. *If $(\mathcal{G}, \sigma_g(g|s))$ is an information structure derived from the private information observed by each of M agents with $\sigma_g(g|s) = \Pr(g|s)$, and $(\mathcal{D}^M, \sigma_0(d_{1:M}|s))$ is the information structure of original private information with $\sigma_0(d_{1:M}|s) = \Pr(d_{1:M}|s)$, then the following holds:*

$$(\mathcal{G}, \sigma_g(g|s)) \trianglelefteq (\mathcal{D}^M, \sigma_0(d_{1:M}|s)).$$

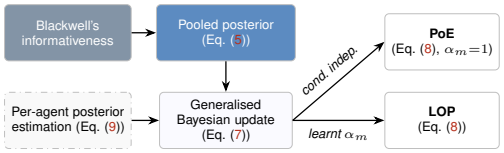


Figure 1: Overview of our framework. Blackwell's informativeness framework (Theorem 1 and Propositions 1 and 2) identifies the pooled posterior (Eq. (5)) as a principled target decision variable. Since agents' private information is unavailable in practical multi-LLM settings, we approximate this target using a generalised Bayesian aggregation framework (Eq. (7)) applied to per-agent posteriors estimated from sampled responses (Eq. (9)). PoE and LOP (Eq. (8)) are practical instantiations of this approximation.

Proof. The proof is based on the definition of probability and the sum rule as follows:

$$\sigma_g(g|s) = \Pr(g \in \mathcal{G}|s) = \Pr\big(d'_{1:M} \in \gamma^{-1}(d_{1:M})|s\big) = \textstyle\sum_{d'_{1:M} \in \mathcal{D}^M} \kappa(d_{1:M}, d'_{1:M}) \sigma_0(d'_{1:M}|s),$$

where $\kappa(d_{1:M}, d'_{1:M}) = \mathbf{1}\left[\gamma(d'_{1:M}) = \gamma(d_{1:M})\right]$ denotes a valid stochastic mapping function. Hence, according to Definition 4, $(\mathcal{G}, \sigma_g(g|s)) \trianglelefteq (\mathcal{D}^M, \sigma_0(d_{1:M}|s))$. □

Intuitively, Proposition 1 shows that no aggregation mechanism can outperform the pooled information structure of all private information under the Blackwell ordering framework. The contribution of this analysis is to apply Blackwell's ordering to clarify the common information upper bound underlying MLAA protocols and to motivate the pooled posterior as an ideal decision variable. Specifically, Proposition 1 establishes that any statistic computed from the joint private signals, including voting outcomes, debate transcripts, estimated posterior distributions, is weakly Blackwell-inferior to direct access to those signals. The result therefore characterises potential information loss under aggregation, but does not establish a relative Blackwell ordering among practical aggregation mechanisms.

4.2 Answer aggregation in multi-LLM Agents classification

In classification with an ensemble of M agents, aggregation is often performed in the label space, meaning $\mathcal{Z} \equiv \mathcal{Y}$. For example, with weighted averaging, the prediction can be calculated as follows:

$$g_{\text{weighted}} = \gamma_{\text{weighted}}(d_{1:M}) = \textstyle\sum_{m=1}^{M} w_m \Pr(y|x, d_m), \tag{2}$$

where: $w_m \in [0, 1]$ and $\sum_{m=1}^{M} w_m = 1$. When $w_m = 1/M$, this resembles the ensemble learning.

We employ Proposition 1 to show that the decision obtained via either aggregated information sources or debate is no more informative under the Blackwell's framework than the decision made from the information source of pooled original private observations.

Proposition 2. *Given M agents, each observing private information $d_m \in \mathcal{D}, m \in \{1, \ldots, M\}$, the prediction probability obtained either through:*

- *an ensemble of those agents (e.g., weighted average over an ensemble), or*
- *multi-round debate of M Bayesian agents,*

will result in an information structure that is the garbling of $(\mathcal{D}^M, \sigma_0(d_{1:M}|y, x))$.

This result holds independently of protocol details, debate structure, or learning dynamics. We intentionally omit the details of communication protocols or equilibrium formation within multi-LLM systems. For our purposes, it suffices to observe that common collective-decision procedures, such as voting, or debate, are instances of information aggregation mappings applied to agents' private signals. Within this simplified framework, the following is an informal proof sketch of the corollary.

Proof sketch. For ensemble learning, as shown in Eq. (2), the aggregated prediction probability g_{weighted} belongs to the set of aggregation of private information $\mathcal{G}$. Hence, according to Theorem 1 and Proposition 1, it is a garbling of $(\mathcal{D}^M, \sigma_0(d_{1:M}|y, x))$. The debate can also be proved similarly. A detailed proof is presented in App. B. □

As a result, in any classification, every Bayesian decision maker always prefers the information source $(\mathcal{D}^M, \sigma_0(d_{1:M}|y, x))$ to either the information source obtained via label aggregation or debate. The expected utility value of the information source $(\mathcal{D}^M, \sigma_0(d_{1:M}|y, x))$ can be written as follows:

$$\omega(\sigma_0; x) = \textstyle\sum_{d_{1:M} \in \mathcal{D}^M} \max_\pi \sum_{y \in \mathcal{Y}} \varphi(\pi, y) \sigma_0(d_{1:M}|y, x) \rho(y|x), \tag{3}$$

where π denotes the predicted label distribution. The Bayes-optimal action can be written as:

$$\pi^* = \arg\max_\pi \textstyle\sum_{y \in \mathcal{Y}} \varphi(\pi, y) \sigma_0(d_{1:M}|y, x) \rho(y|x) = \arg\max_\pi \mathbb{E}_{\Pr(y|x, d_{1:M})} \left[\varphi(\pi, y)\right]. \tag{4}$$

Hence, the Bayes-optimal predictive label distribution π^* depends on the utility function $\varphi(.,.)$ and the pooled posterior $\Pr(y|x, d_{1:M})$. For example, under the 0-1 utility $\varphi(\pi, y) = y^\top \pi$, the objective in Eq. (4) is linear in π, so its maximiser is a vertex of the simplex, i.e., a point-mass distribution $\pi^*_{0\text{-}1} = \delta_{\hat{y}^*}$, where the predicted label $\hat{y}^*$ is the maximum a posteriori of the pooled posterior:

$$\pi^*_{0\text{-}1} = \delta_{\hat{y}^*} \wedge \hat{y}^* = \operatorname{argmax}_\pi \Pr(y|x, d_{1:M}). \tag{5}$$

Under the categorical utility function used in classification: $\varphi(\pi, y) = y^\top \ln \pi$ with y in the one-hot vector form, the optimisation in Eq. (4) corresponding to minimise the cross-entropy between $\Pr(y|x, d_{1:M})$ and π. According to the Gibbs' inequality, the optimal predictive label distribution is:

$$\pi^*_{\text{categorical}} = \Pr(y|x, d_{1:M}). \tag{6}$$

Proposition 2 and Eqs. (4) to (6) identify the pooled posterior $\Pr(y \mid x, d_{1:M})$ as a principled target decision variable under the Blackwell information-structure abstraction. The decision rules in Eqs. (5) and (6) are specific to the 0-1 and categorical utility functions, respectively. Under a general utility function, the optimal action instead maximises posterior expected utility as in Eq. (4). More broadly, Proposition 2 establishes the agents' joint private signals as a common information upper bound for voting, debate, and posterior-level aggregation, rather than a relative Blackwell ordering among these mechanisms.

Remark 1. The results in Eqs. (5) and (6) also explain the diminishing returns and eventual performance plateaus observed in several multi-LLM debate settings. In homogeneous ensembles, where agents share pre-training corpora, alignment procedures, and retrieval pipelines, their private knowledge is positively correlated (as we verify directly in Tab. 10), reducing the diversity of joint private knowledge $\cap_{m=1}^{M} d_m$, and hence, performance is predictably lower than in heterogeneous configurations as reported in (Zhang et al., 2025). Even in heterogeneous pools, increasing the number of agents beyond a modest threshold yields limited gains once their training data cease to contribute novel information, leading to the plateau observed in Tab. 3. Our findings in Eqs. (5) and (6) suggest that, for pooled posterior methods, the most effective configuration involves a small number of agents whose private knowledge is as close to conditionally independent as feasible. Ensuring diversity is therefore crucial to realise the theoretical advantages of pooled aggregation in MLAA systems.

The result in Eq. (5) is also known as maximum Bayesian pooled posterior. Within this Blackwell information-structure abstraction, voting and debate are less informative than access to the full pooled private information. Next, we present a practical method to approximate the pooled posterior for decision making in multi-LLM classification.

Estimate the pooled posterior If all of the private information observed by each agent, $d_m, \forall m \in \{1, \ldots, M\}$ is available, it is straight-forward to obtain $\Pr(y|x, d_{1:M})$. This is, however, not true in practice because the private information (e.g., private training data) d_m of each agent is often unavailable to other agents $m' \in \{1, \ldots, M\} \setminus \{m\}$. Instead, it is only possible to sample the prediction made by each agent's posterior $\Pr(y|x, d_m)$. If this sampling process is reasonably cheap, one can repeatedly sample multiple predictions of one agent to approximate its posterior, denoted as $\widetilde{\Pr}(y|x, d_m)$. In the following, we assume the availability of $\widetilde{\Pr}(y|x, d_m), m \in \{1, \ldots, M\}$, and use these to approximate the pooled posterior $\Pr(y|x, d_{1:M})$.

Given the availability of each agent's approximated posterior $\widetilde{\Pr}(y|x, d_m)$, the pooled posterior can then be estimated via a generalised Bayesian update:

$$\begin{aligned}\Pr(y|x, d_{1:M}) &\propto \Pr(y|x) \textstyle\prod_{m=1}^{M} [\Pr(d_m|y, x)]^{\alpha_m} \propto \Pr(y|x) \textstyle\prod_{m=1}^{M} \left[\frac{\Pr(y|x, d_m) \Pr(d_m)}{\Pr(y|x)}\right]^{\alpha_m} \\ &\propto [\Pr(y|x)]^{1-\sum_{i=1}^{M} \alpha_m} \textstyle\prod_{m=1}^{M} [\widetilde{\Pr}(y|x, d_m)]^{\alpha_m},\end{aligned} \tag{7}$$

where $\alpha_m > 0$ is a source-specific likelihood weight controlling the contribution of source d_i to the pooled posterior.

In the information aggregation literature (Ai et al., 2025; Prelec et al., 2017; Schoenebeck & Tao, 2021; Pan et al., 2024; Pettigrew, 2019), the prior $\Pr(y|x)$ is frequently assumed to be uniform, corresponding to the absence of external information regarding the class-label distribution. Under this assumption, Eq. (7) reduces to the *logarithmic opinion pool* (LOP):

$$\Pr(y|x, d_{1:M}) \propto \textstyle\prod_{i=1}^{M} [\widetilde{\Pr}(y|x, d_m)]^{\alpha_m}. \tag{8}$$

That is, the pooled posterior is given by an exponentially weighted product of the individual agents' posterior beliefs, up to a normalisation constant. The source-specific likelihood weight α_m admits an intuitive interpretation. When $\alpha_m = 1, \forall m$, Eq. (7) reduces to the standard likelihood factorisation obtained under the conditional independence assumption $d_m \perp d_{m'}|y, x$ for $m \neq m'$. This is the same algebraic form underlying the *product-of-experts* (PoE) framework. More generally, values $0 < \alpha_m < 1$ discount the contribution of source d_m, which may be desirable when its information is partially redundant with that of other sources, thereby mitigating the double-counting of evidence. Conversely, values $\alpha_m > 1$ increase the influence of source d_m, reflecting a modelling choice that assigns greater evidential weight to sources deemed particularly reliable, informative, or relevant to the inference task. Eq. (8) should be interpreted as a generalised-Bayes aggregation rule applied to the approximated agent posterior. Under conditional independence and correctly specified per-agent

posteriors, its algebraic form coincides with the corresponding pooled-posterior factorisation. These conditions are unlikely to hold exactly for contemporary LLMs. Accordingly, PoE and LOP are practical, assumption-dependent approximations motivated by the ideal pooled posterior, rather than consistent estimators or guarantees of Blackwell-optimality.

Approximate $\widetilde{\Pr}(y|x, d_m)$ of LLM agents in multiple-choice questions We employ the similarity of embeddings between the agent's response $y^{(m)} \sim \Pr(y|x, d_m)$ and each of J provided answers (or options) $y_{o_j}, j \in \{1, \ldots, J\}$ in a multiple-choice question (e.g., y_{o_j} is a natural language expression as: A-benign, B-malignant or C-no finding as in App. E) to approximate $\widetilde{\Pr}(y|x, d_m)$. Formally, for each pair of agent's response $y_n^{(m)}$ and a provided list of options, y_{o_j}, we calculate the similarity of their embeddings, denoted as $\tau(y_n^{(m)}, y_{o_j})$. For example, τ can be the dot product or cosine similarity between the two embedding vectors of $y_n^{(m)}$ and y_{o_j}. The predictive probability distribution over all provided options can be approximated as follows:

$$\widetilde{\Pr}(y = y_{o_j}|x, d_m) = \frac{1}{N} \sum_{n=1}^{N} \frac{\phi\big(\tau(y_n^{(m)}, y_{o_j})\big)}{\sum_{j'=1}^{J} \phi\Big(\tau(y_n^{(m)}, y_{o_{j'}})\Big)}, \tag{9}$$

where N is the number of "Monte Carlo" responses for the same question, and the additional model $\phi(\cdot)$ is introduced to map the similarity score to a predictive logit (e.g., plays a role as Platt scaling (Platt, 1999)). Note that we estimate each agent's posterior from full generated responses by measuring their semantic compatibility with each candidate option using an NLI cross-encoder, followed by a learnable transformation. This avoids relying on raw LLM logits, whose token-level probabilities over labels or answer strings can be noisy, length-biased, and misaligned with full-answer semantics; see App. C for details.

Learn α_m Given the approximation of each agent's posterior in Eq. (9), we can learn weights α_m for $m \in \{1, \ldots, M\}$ from a small subset of labelled data by maximising the RHS term of Eq. (8).

Given the obtained $\widetilde{\Pr}(y|x, d_m)$ and α_m, we can estimate the pooled posterior $\Pr(y|x, d_{1:M})$ by following the exponentially-weighted product in Eq. (5).

5 Experiments

Baselines We follow the setting in (Choi et al., 2025a) to evaluate existing methods in MAD and voting-based methods. The MAD approaches include: *Centralised MAD* (Guo et al., 2024), where a central agent aggregates peer responses and generates the updated response at each round, *Decentralised MAD* (Du et al., 2023), where each agent observes all other agents' responses from the previous round, *Sparse MAD* (Li et al., 2024b)–a variant of Decentralised MAD with a sparse communication topology to enhance efficiency, and *Free MAD* (Cui et al., 2026), where decisions are made by evaluating the entire trajectory rather than the final consensus. The voting-based approach includes: *majority voting*, *Self-Consistency* (SC) (Wang et al., 2023), which samples N responses from each agent and selects the most frequent answer across all agents, and *Inverse Surprising Popularity* (ISP) (Ai et al., 2025), which leverages first-order (e.g., agent accuracies) and second-order information (e.g., answer correlations) to aggregate LLMs' responses. Each agent is sampled with the same number of responses, N, to fairly compare.

We consider a pool of $M = 5$ agents (Qwen Team, 2025; Llama Team, 2024; Mistral AI team, 2023; Falcon LLM Team, 2025; Gemma Team, 2025) for multi-LLM methods in our main comparison and will ablate on the effect of M. For single-agent baselines, we average across five independent runs. In the homogeneous setting (i.e., all agents are instances of the same LLM), our derived methods, including both PoE and LOP, simplify the modelling to a single agent, while in the heterogeneous setting, our derived methods are carried out following Algorithm 1 in App. F. We also keep the same number of Monte Carlo samples $N = 5$ when estimating each agent's posterior throughout all the experiments. For all MAD baselines, we report results with $T \in \{2, 3, 5\}$ debate rounds. We treat the heterogeneous-agent experiments as the primary test of multi-LLM information pooling, and the homogeneous-agent experiments as a test of the per-agent posterior estimation pipeline (Eq. (9)). Readers are referred to App. D for further details.

Benchmarks consist of six natural language QA tasks, each associated with the following datasets: *(i)* Factual Question Answering (MMLU Professional Medicine and Formal Logic (Hendrycks

Table 2: Prediction accuracy of $M = 5$ agents in the heterogeneous setting across six QA benchmarks. The results consist of the mean and standard deviations obtained from three different random seeds. The best result per dataset is highlighted in **bold**.

	MMLU (Pro.Med.)	MMLU (Form.Log.)	HellaSwag	CSQA	HH-RLHF	MedMCQA
Single-LLM Agent						
Qwen-7B	0.7868 ± .01	0.4905 ± .03	0.7880 ± .01	0.8153 ± .01	0.4773 ± .01	0.5467 ± .01
Falcon-7B	0.7904 ± .01	0.5873 ± .01	0.7133 ± .01	0.8300 ± .01	0.5033 ± .01	0.5767 ± .01
Mistral-7B	0.6544 ± .01	0.3730 ± .01	0.6433 ± .01	0.6800 ± .01	0.5567 ± .01	0.3133 ± .01
Llama-8B	0.7441 ± .01	0.3794 ± .02	0.6267 ± .03	0.6767 ± .01	0.4440 ± .02	0.5000 ± .01
Gemma-9B	0.8015 ± .01	0.5159 ± .02	0.8267 ± .01	0.8033 ± .01	0.5567 ± .02	0.5567 ± .01
Multi-LLM Agents						
Decentr. MAD (T=2)	0.8603 ± .01	0.7063 ± .01	0.8267 ± .01	0.8500 ± .01	0.5867 ± .01	0.6200 ± .01
Decentr. MAD (T=3)	0.8603 ± .00	0.6904 ± .01	0.8333 ± .01	0.8533 ± .01	0.5867 ± .01	0.6067 ± .01
Decentr. MAD (T=5)	0.8676 ± .00	0.6984 ± .02	0.8400 ± .01	0.8433 ± .00	0.5700 ± .00	0.6000 ± .01
Centr. MAD (T=2)	0.8372 ± .02	0.5635 ± .03	0.7667 ± .02	0.8233 ± .01	0.5533 ± .01	0.5900 ± .02
Centr. MAD (T=3)	0.8333 ± .01	0.6032 ± .02	0.7567 ± .01	0.8067 ± .00	0.5567 ± .01	0.5833 ± .02
Centr. MAD (T=5)	0.8419 ± .01	0.6032 ± .03	0.7633 ± .02	0.8233 ± .01	0.5367 ± .00	0.5967 ± .01
Sparse MAD (T=2)	0.8603 ± .00	0.6667 ± .02	0.8200 ± .01	0.8633 ± .00	0.5733 ± .00	0.6200 ± .01
Sparse MAD (T=3)	0.8529 ± .00	0.7143 ± .00	0.8133 ± .00	0.8533 ± .01	0.5533 ± .01	0.6033 ± .01
Sparse MAD (T=5)	0.8529 ± .01	0.7084 ± .00	0.8133 ± .01	0.8500 ± .00	0.5600 ± .00	0.6167 ± .02
Free MAD (T=2)	0.8713 ± .00	0.6111 ± .03	0.7900 ± .02	0.8533 ± .00	0.5500 ± .03	0.6133 ± .02
Free MAD (T=3)	0.8713 ± .00	0.6111 ± .04	0.7967 ± .02	0.8567 ± .00	0.5567 ± .02	0.6200 ± .00
Free MAD (T=5)	0.8750 ± .01	0.6270 ± .03	0.7933 ± .02	0.8467 ± .01	0.5667 ± .00	0.6267 ± .00
Majority Voting	0.8493 ± .01	0.6190 ± .01	0.8033 ± .01	0.8533 ± .00	0.5633 ± .01	0.6167 ± .01
Self-Consistency	0.8603 ± .01	0.7084 ± .01	0.8233 ± .00	0.8600 ± .00	0.5733 ± .01	0.6167 ± .01
ISP	0.8640 ± .00	0.6429 ± .01	0.8267 ± .00	0.8600 ± .00	0.5667 ± .01	0.6367 ± .01
PoE	0.8787 ± .01	0.7367 ± .01	**0.8433 ± .01**	**0.8800 ± .00**	**0.5900 ± .01**	0.6467 ± .01
LOP	**0.8811 ± .02**	**0.7445 ± .01**	0.8400 ± .00	0.8767 ± .01	**0.5900 ± .01**	**0.6500 ± .00**

et al., 2021; Estornell & Liu, 2024)), *(ii)* Natural Language Inference (HellaSwag (Zellers et al., 2019)), *(iii)* Commonsense Reasoning (CSQA (Talmor et al., 2019)), *(iv)* Alignment Labeling (HH-RLHF (Anthropic Team, 2022)), where we adopt the "AI labeller alignment" practice (Lee et al., 2024), and *(v)* Medical Entrance Exam (MedMCQA) (Pal et al., 2022). For a fair comparison, all baselines are evaluated on the same data subsets. Readers are referred to App. D for further details.

Results on standard MAD benchmarks are shown in Tab. 2 with five LLMs across six benchmarks, while Fig. 2 in App. G shows the gap among the best single model/ SOTA MAD, voting methods, and both PoE and LOP. Empirical results show that multi-agent methods consistently improve performance over single-agent baselines across most datasets. However, the degree of improvement varies significantly by task, highlighting the sensitivity of these methods to the underlying task complexity. PoE and LOP are competitive with or achieve higher accuracy than other baselines across all benchmarks. Although majority voting has strong performance in homogeneous agent settings (Choi et al., 2025a), it is not on par with MAD methods in heterogeneous settings and sometimes even underperforms individual agents (e.g.,HellaSwag). Furthermore, debate methods show inconsistent performance across different numbers of rounds (e.g., HellaSwag, HH-RLHF). Additional debate rounds provide marginal gains or even degrade performance compared to naive voting. These findings empirically validates recent critiques (Choi et al., 2025a; Kaesberg et al., 2025) regarding the diminishing returns of iterative debate. In contrast, both PoE and LOP aggregate the agents' initial probabilistic beliefs in a single turn, avoiding multi-turn degradation entirely while being more computationally efficient (1.3-4.8 $\times$ fewer tokens) and stable than iterative communication.

Heterogeneous and Homogeneous Pool of LLMs To assess robustness to dependence, we evaluate across agent-diversity regimes: heterogeneous agents, reduced heterogeneous pools, homogeneous agents, and larger models with likely overlapping training data. Tab. 3 shows that both rules maintain strong performance even with reduced agent diversity, demonstrating the robustness to ensemble size (see App. H). However, we also observe that increasing the number of agents beyond a certain threshold does not yield proportional performance gains across all datasets. This limitation stems from the conditional independence assumption on agents' private training data imposed in Eq. (7). While introducing additional agents can initially enrich the collective knowledge, thereby improving performance, this effect diminishes once the pool of agents begins to contribute overlapping information rather than genuinely novel evidence. Beyond this point, adding more agents no longer enhances informativeness and instead leads to a performance plateau.

Tab. 4 in App. H presents results at the extreme setting consisting of only 2 agents. This is noteworthy because majority voting fundamentally fails when two agents disagree. In that case, majority voting randomly selects one out of the two answers, resulting in an unreliable decision. In contrast, both PoE and LOP can, however, leverage the strength of each agent's posterior distribution to make informed

decisions without the need of consensus. Empirically, our methods perform competitively and achieve the best results on most benchmarks in this setting, suggesting that posterior-level aggregation can remain useful even when majority voting becomes unreliable.

Following (Choi et al., 2025a), Tabs. 5 and 6 in App. I evaluate in the homogeneous setting (e.g., all agents are instances of the same LLM). In this setting, the pooled posterior is simplified into a single posterior, so PoE and LOP implement the same decision rule. Nevertheless, both rules still outperform single-agent and MAD baselines. Notably, Centralised MAD exhibits severe degradation with additional rounds of debate, highlighting the risks of iterative debate when agents share the same model and may amplify shared biases (Zhang et al., 2025).

Larger and more capable models We extend the evaluation by employing LLMs with larger capacity. Specifically, for the homogeneous setting, `Qwen2.5-32B` (Qwen Team, 2024) is used as a base model. As shown in Tab. 7 (see App. J), the performance of majority voting remains comparable to that of multi-LLM methods. For the heterogeneous setting, `Qwen2.5-32B` (Qwen Team, 2025), `Falcon-H1-34B` (Falcon LLM Team, 2025), and `Gemma-2-27B` (Gemma Team, 2025) are used for 3-agent debate or voting. The results in Tab. 8 (see App. J) for Centralised MAD agree with our previous evaluation in Tabs. 2 and 3, in which Centralised MAD cannot improve performance and is even worse than single models and majority voting. Other MAD methods perform better than single agent but are inferior compared to our proposed methods.

Sensitivity to inter-agent dependence PoE (α_m=1) is exact only if agents' private information is conditionally independent, which is implausible for LLMs trained on overlapping corpora. We therefore measure dependence directly (Tab. 10 in App. M). For *error correlation*, we report the mean pairwise Yule's $\bar{Q}$ over agent correctness indicators, capturing whether agents make mistakes on the same items; for *posterior correlation*, we report the mean pairwise Pearson correlation $\bar{r}$ between calibrated posterior vectors, capturing whether agents assign similar probability mass across options. Heterogeneous pools generally exhibit lower dependence than homogeneous ones, yet remain substantially dependent. This confirms that Eq. (8) with α_m=1 is an evidence-accumulation approximation, which LOP relaxes through learned weights α_m.

Fig. 4 shows the weights α_m fitted on the pooled validation split for two heterogeneous pools. In both pools, the induced ordering matches that of average single-agent accuracy. Since α is estimated on a split disjoint from all test sets, this provides an independent check that the estimated posteriors carry genuine signal rather than being an artifact of the calibration head. The weakest agent, Mistral-7B, is down-weighted to α_m=0.20, whereas Gemma-9B and Falcon-34B receive $\alpha_m > 1$.

Calibration We evaluate calibration quality by measuring Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) in Tab. 9, with reliability diagrams in Fig. 3. App. K compares performance with and without our calibration module for 4 models on MedMCQA. Results demonstrate that our method achieves substantial calibration improvements across different model scales, producing well-calibrated probability estimates that enable effective consensus mechanisms.

Ablation Study We isolate each component through four ablations against the full method (Tab. 11 in App. N) with PoE as the reference rule. (1) *w/o pooling* replaces cross-agent pooling with a single best agent's calibrated posterior, averaged over $N = 5$ Monte Carlo samples (Eq. (9)). The results show that calibration alone yields part of the gain over the single-agent baseline, while cross-agent pooling provides additional improvements across most benchmarks. (2) *w/o calibration* replaces the Deep Sets with softmax-normalised NLI logits. Multiplying uncalibrated scores falls below output-level SC, which confirms that well-formed probabilities are a prerequisite for multiplicative aggregation. (3) *Temperature scaling* replaces the Deep Sets with a single learned scalar temperature (Guo et al., 2017; Platt, 1999). This improves over no calibration on some benchmarks (e.g., Pro. Med., CSQA) but underperforms on others (HellaSwag, Form. Log.) and consistently falls short of PoE, suggesting that a scalar temperature is insufficient to capture the permutation-equivariant structure of multiple options. (4) *MV/SC with calibrated posteriors (linear pooling)* applies Majority Voting and Self-Consistency to the same calibrated per-agent posteriors as PoE, so that all rules share identical supervision. Majority voting and self-consistency applied to the calibrated outputs perform below PoE, suggesting that retaining and combining the approximated posteriors from the agents is useful relative to these hard-decision and frequency-based rules.

Number of Monte Carlo samples We analyse the impact of the number of Monte Carlo samples, denoted by N, on the quality of posterior estimation in Eq. (9) The results in Fig. 5 (see App. O)

for the homogeneous setting show substantial improvement when increasing from $N = 1$ to $N = 5$ samples, with performance stabilising beyond $N \geq 5$. This demonstrates that posterior estimates converge with additional samples, though with diminishing marginal returns, suggesting $N = 5$ provides a good balance between estimation quality and computational cost.

Efficiency We evaluate computational cost by measuring the input, output, and total number of tokens consumed per sample on the MedMCQA dataset in the 5-agent setting. Tab. 12 in App. P reports token usage accumulated across all debate rounds. Our methods achieve token efficiency comparable to single-round baselines and substantially lower than MAD methods, particularly at higher debate rounds. The additional NLI cross-encoder scoring runs in under one second per sample on a single GPU and is negligible relative to LLM generation cost, which is about 25 seconds.

6 Related work

Opinion Aggregation Opinion pooling concerns the aggregation of probabilistic beliefs from multiple experts into a single opinion (Stone, 1961). DeGroot (DeGroot, 1974) iteratively update beliefs via weighted linear combinations, corresponding to linear opinion pooling (e.g., Eq. (2)). Beyond linear pooling, log-linear pooling (LLP) has been proposed (Genest & Zidek, 1986) to account for dependencies among agents' beliefs. LLP assigns each agent m a non-negative weight $w_m \geq 0$, yielding a pooled belief of the form $\prod_{m=1}^{M}[\Pr(y|x, d_m)]^{w_m}$. Log-linear pooling has also been studied in the context of incoherent opinion pooling through renormalisation procedures (Pettigrew, 2019). Product of Experts (Hinton, 2002) combines expert likelihoods multiplicatively as a modelling and training paradigm, but without decision-theoretic grounding. Our framework recovers its algebraic form as the Blackwell-based pooled posterior under conditional independence.

Aggregation in multi-LLM systems Voting-based schemes instantiate linear pooling, from majority voting and self-consistency (Wang et al., 2023; 2024) to confidence-weighted (Xiong et al., 2024), pairwise ranking-based (Jiang et al., 2023), and representation-consistency variants (Jiang et al., 2025). Recent work such as *Inverse Surprising Popularity* (Ai et al., 2025) further improves aggregation by leveraging agent accuracies (i.e., first-order) and inter-answer (i.e, second-order) correlations. Debate-based schemes (Du et al., 2023; Chan et al., 2024; Li et al., 2024b; Cui et al., 2026) instead exchange opinions over multiple rounds, though recent studies question their reliability relative to voting or even single agents (Choi et al., 2025a; Zhang et al., 2025). Prior comparisons between the two families are empirical and protocol-specific. Our analysis, however, shows both are garblings of the pooled private information (Proposition 2), yielding a protocol-independent account of their limitations. App. Q provides further related work, including multi-agent decision-making, multi-agent debate, and LLM uncertainty estimation.

7 Discussion and conclusion

In this paper, we study MLLA through Blackwell's informativeness framework. We show that the output of any aggregation or communication protocol is a garbling of the agents' joint private signals, where the joint private information is a protocol-independent upper bound. The ideal decision based on the joint signals is obtained from the posterior conditioned on their joint information. Since the information is inaccessible in black-box LLM settings, we estimate response-induced option distributions and combine them using PoE and LOP. Across six benchmarks, posterior-level aggregation is competitive with or achieves higher accuracy than baselines, while avoiding the communication costs of multi-round debate.

A key limitation is that the pooled posterior identified by the Blackwell analysis is not directly accessible in realistic multi-LLM settings, so PoE and LOP should be viewed as practical approximations. Moreover, PoE relies on conditional independence assumption, while learned LOP weights do not explicitly model higher-order dependence. Finally, our instantiation targets multiple-choice QA. The framework itself is not limited to this setting: the Blackwell analysis and the pooling rules in Eq. (7) apply to any decision problem in which agents' posteriors over a shared outcome space can be estimated. An extension to open-ended generation is to replace the fixed option set with clusters discovered in embedding space via a Dirichlet Process mixture, which groups semantically coherent answers into K data-inferred clusters. Each agent's posterior over the discovered clusters is estimated from its Monte Carlo samples following Eq. (9), and PoE and LOP then apply unchanged via Eq. (8).

REFERENCES

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: LLM aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*, 2025. 5, 6, 9

Anthropic Team. Training a helpful and harmless assistant with reinforcement learning from human feedback. Technical report, Anthropic, 2022. 7, 16

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, pp. 265–272, 1953. 2

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *ICLR*, 2024. 1, 9

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *NeurIPS*, 2025a. 1, 6, 7, 8, 9, 16, 18, 23, 24, 31

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Measuring and mitigating identity bias in multi-agent debate via anonymization. *arXiv preprint arXiv:2510.07517*, 2025b. 16

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. In *ACL Findings*, 2026. 6, 9, 31

Henrique de Oliveira. Blackwell's informativeness theorem using diagrams. *Games and Economic Behavior*, 109:126–131, 2018. 2

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345): 118–121, 1974. 9

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2023. 1, 6, 9, 31

Andrew Estornell and Yang Liu. Multi-LLM debate: Framework, principals, and interventions. In *NeurIPS*, 2024. 1, 7, 16

Falcon LLM Team. Falcon-H1: A family of hybrid-head language models redefining efficiency and performance. Technical report, Technology Innovation Institute, UAE, 2025. 6, 8, 16, 24

Gemma Team. Gemma 3 technical report. Technical report, Google DeepMind, 2025. 6, 8, 16, 24

Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986. 9

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 8, 28

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *IJCAI*, pp. 8048–8057, 2024. 1, 6

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 6, 16

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. 9

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. Technical report, OpenAI, 2018. 31

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *ACL*, pp. 14165–14178, 2023. 9

Junqi Jiang, Tom Bewley, Salim I Amoukou, Francesco Leofante, Antonio Rago, Saumitra Mishra, and Francesca Toni. Representation consistency for accurate and coherent llm answer aggregation. In *NeurIPS*, 2025. 9

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. Technical report, Anthropic, 2022. 31

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? Decision-making in multi-agent debate. In *ACL Findings*, 2025. 7

Peter GW Keen and Michael S Scott Morton. *Decision support systems: An organizational perspective*. Addison-Wesley series on decision support, 1978. 31

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *ICML*, pp. 23662–23733. PMLR, 2024. 31

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *ICLR*, 2026. 1

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *NeurIPS*, volume 30, 2017. 31

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *ICML*, 2024. 7, 16

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. In *NeurIPS*, 2022. 1

Ruosen Li, Teerth Patel, and Xinya Du. PRD: Peer rank and discussion improve large language model based evaluations. *TMLR*, 2024a. 1

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022. 1

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *EMNLP Findings*, pp. 7281–7294, 2024b. 6, 9

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *TMLR*, 2022. 31

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *TMLR*, 2024, 2024. 31

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. GroupDebate: Enhancing the efficiency of multi-agent debate using group discussion. In *International Conference on Autonomous Agents and Multiagent Systems*, 2026. 31

Llama Team. The Llama 3 herd of models. Technical report, AI @ Meta, 2024. 6, 16

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*, 2023. 31

Mistral AI team. Mistral 7b. Technical report, Mistral AI, 2023. 6, 16

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994. 3

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pp. 248–260. PMLR, 2022. 7, 16

Yuqi Pan, Zhaohua Chen, and Yuqing Kong. Robust decision aggregation with second-order information. In *ACM Web Conference 2024*, pp. 77–88, 2024. 5

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019. 16

Richard Pettigrew. Aggregating incoherent agents who disagree. *Synthese*, 196(7):2737–2776, 2019. 5, 9

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999. 6, 8, 28

Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017. 5

Qwen Team. Qwen2.5 technical report. Technical report, Alibaba Cloud, 2024. 8

Qwen Team. Qwen3 technical report. Technical report, Alibaba Cloud, 2025. 6, 8, 16, 24

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *JMLR*, 21(178):1–51, 2020. 31

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019. 16

Grant Schoenebeck and Biaoshuai Tao. Wisdom of the crowd voting: Truthful aggregation of voter information and preferences. *NeurIPS*, 34:1872–1883, 2021. 5

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *nature*, 620(7972):172–180, 2023. 1

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *ICML*, 2019. 31

Mervyn Stone. The opinion pool. *The Annals of Mathematical Statistics*, pp. 1339–1342, 1961. 9

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *AAMAS*, 2018. 31

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*, 2019. 7, 16

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*, pp. 5433–5442, 2023. 31

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019. 31

Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. Boosting language models reasoning with chain-of-knowledge prompting. In *ACL*, pp. 4958–4981, 2024. 9

Jianrui Wang, Yitian Hong, Jiali Wang, Jiapeng Xu, Yang Tang, Qing-Long Han, and Jürgen Kurths. Cooperative and competitive multi-agent systems: From optimization to games. *Journal of Automatica Sinica*, 9(5):763–783, 2022. 31

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 1, 6, 9

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *ICLR*, 2024. 9, 31

Lei Yan, Junhe Liu, CL Philip Chen, Yun Zhang, Zongze Wu, and Zhi Liu. Game-based adaptive fuzzy optimal bipartite containment of nonlinear multiagent systems. *IEEE TFS*, 32(3):1455–1465, 2023. 31

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2022. 1

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *NeurIPS*, 2017. 15, 16

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *ACL*, pp. 4791–4800, 2019. 7, 16

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025. 1, 5, 8, 9, 31

Huaguang Zhang, Jilie Zhang, Guang-Hong Yang, and Yanhong Luo. Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming. *IEEE TFS*, 23(1):152–163, 2014. 31

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*, 2023. 31

A AI USE STATEMENT

All agents in our experiments are open-weight LLMs, whose sampled responses are the inputs to both our aggregation methods and all baselines. No LLM was used to generate data, run experiments, or produce reported numbers.

B PROOF OF PROPOSITION 2

Proposition 2. *Given M agents, each observing private information $d_m \in \mathcal{D}, m \in \{1, \ldots, M\}$, the prediction probability obtained either through:*

- *an ensemble of those agents (e.g., weighted average over an ensemble), or*
- *multi-round debate of M Bayesian agents,*

will result in an information structure that is the garbling of $(\mathcal{D}^M, \sigma_0(d_{1:M}|y, x))$.

Proof. We prove that the consensus obtained via Multi-LLM Agents debate is an aggregation or pooling of the private information $\mathcal{G}$. The proof is carried out by induction.

In the first round ($T = 1$), each agent m draws an opinion $y_m^{(0)}$ from its probabilistic belief $\Pr(y|x, d_m)$ and share that to other agents. The opinion, $y_m^{(0)}$ can also be viewed as a function of the private knowledge d_m as follows:

$$y_m^{(0)} \sim \Pr(y|x, d_m) \implies y_m^{(0)} = y_m^{(0)}(d_m). \tag{10}$$

After the first round, each agent observes all samples of belief shared by other agents, and update its belief using Bayes's rule as follows:

$$\Pr^{(1)}\left(y \middle| x, d_m, \prod_{m'=1, m' \neq m}^{M} y_{m'}^{(0)}(d_{m'})\right) \propto \Pr(y|x, d_m) \prod_{m'=1, m' \neq m}^{M} \Pr\left(y_{m'}^{(0)}(d_{m'}) \middle| x, y, d_m\right). \tag{11}$$

Because each opinion $y_m^{(0)}$ is a function of each private information d_m, we can represent the posterior under the "lumped" function form as follows:

$$\Pr^{(1)}\left(y \middle| x, \gamma_m^{(0)}(d_{1:M})\right), \tag{12}$$

where $\gamma_m^{(0)}$ is a function.

In the second round $T = 2$, we follow the similar procedure and show that the posterior of an agents is as follows:

$$\Pr^{(2)}\left(y \middle| x, d_m, \prod_{m'=1, m' \neq m}^{M} y_{m'}^{(1)}(d_{m'})\right) = \Pr^{(2)}\left(y \middle| x, \gamma_m^{(1)}(d_{1:M})\right), \tag{13}$$

where $\gamma_m^{(1)}$ is a function at round $T = 2$.

At the consensus, the posterior of all agents is denoted as: $\Pr(y|x, \gamma_{\text{con.}}(d_{1:M}))$. The information structure corresponds to this posterior can be written as follows:

$$\Pr(\gamma_{\text{con.}}(d_{1:M})|x, y) \propto \Pr(y|x, \gamma_{\text{con.}}(d_{1:M})) \Pr(\gamma_{\text{con.}}(d_{1:M})|x). \tag{14}$$

The information structure $(\mathcal{G}_{\text{con.}}, \Pr(g|x, y))$ with $\mathcal{G}_{\text{con.}} = \{g : g = \gamma_{\text{con.}}(d_{1:M})\}$ is obviously an aggregation of the joint of original private information $d_{1:M}$. Hence, according to Proposition 1, $(\mathcal{G}_{\text{con.}}, \Pr(g|x, y)) \trianglelefteq (\mathcal{D}^M, \Pr(d_{1:M}|y, x)$. Applying Theorem 1 on this result completes the proof. □

C LOGITS VS. SEMANTIC EMBEDDING FOR POSTERIOR PROBABILITY ESTIMATION

While using LLM logits directly is an intuitive alternative, there are two limitations that make it not robust enough for reliable posterior estimation in Multi-LLM Agents setting:

Multi-Token Label Ambiguity. LLM logits correspond to the next-token probability conditioned on the prompt, not the probability of an entire candidate answer. When candidate labels span multiple tokens, the first-token logit is only a partial signal. For example, in HellaSwag, given the context "A man is sitting on a roof. he", the candidate answers are ["is using wrap to wrap a pair of skis.", "is ripping level tiles off.", "is holding a rubik's cube.", "starts pulling up roofing on a roof."]. Computing the full joint probability would require autoregressive decoding over every candidate, which is computationally expensive. It will also require to normalise probabilities by answer length, but this does not provide true probabilities (it effectively measures average per-token probability) and can bias results toward longer answers that contain more common (i.e., higher-probability) words.

Signal vs. Noise. Even for single-token labels (e.g., A/B/C/D in multiple-choice), the LLM's probability mass is distributed across the entire vocabulary, not just the valid candidates. Consider a typical scenario where the first-token distribution assigns overwhelming mass to a task-irrelevant token: the probability of "The", "answer", "is", "A", "B", "C", "D"] is [0.9, 0.03, 0.03, 0.01, 0.01, 0.01, 0.01]. After masking to valid candidates and renormalising, we obtain a uniform distribution [0.25, 0.25, 0.25, 0.25] that has lost all discriminative signal. LLMs frequently allocate dominant probability mass to high-frequency tokens that are irrelevant to the task, leaving the valid-label logits in a low-signal regime where noise dominates.

Instead, our method captures the semantics of the complete answer rather than relying on a single token and operates in a continuous representation space, which is more robust than using logits directly.

Note on the approximation of $\tilde{\Pr}(y|x, d_m)$ In the conventional multi-class classification, the order of appearance of class labels is fixed. However, in the setting of multiple choice questions, the "class labels" or provided options can be freely permutated. Hence, the model $\phi(\cdot)$ used for calibrating the prediction probabilities in Eq. (9) must be permutation-equivariant. In our implementation, we formulate ϕ as a fully-connected neural network following the *Deep Sets* architecture (Zaheer et al., 2017) to make the prediction permutation-equivariant.

D Experimental Details

D.1 Hyperparameters

Following (Choi et al., 2025a), we set the sampling temperature to 1.0, and use nucleus sampling probability of 0.9 to enable stochastic sampling from homogeneous agents. The maximum tokens for all models is 512.

D.2 Dataset

MMLU (Hendrycks et al., 2021) consists of 13, 869 multiple-choice questions (Estornell & Liu (2024) use the 3,406 high-school-level questions). MMLU (Formal Logic) is designed to evaluate a model's proficiency in formal reasoning, symbolic manipulation, and logical analysis. Following (Choi et al., 2025a), we use the entire test split comprised of 126 question for evaluation. MMLU (Professional Medicine) is a benchmark designed to evaluate professional-level reasoning in medical domains. It requires knowledge of medical concepts, clinical reasoning, and biomedical science to answer its questions. Following (Choi et al., 2025a;b), we use the full test split, which contains 272 items.

HellaSwag (Zellers et al., 2019) is a natural language inference (NLI) benchmark dataset focused on sentence completion. It evaluates whether a model can select the most plausible continuation of a given context from multiple candidates, a task requiring both linguistic competence and commonsense reasoning. From the original test split, We follow the setting in (Choi et al., 2025b) which randomly sample 300 questions for our evaluations.

CommonsenseQA (Talmor et al., 2019) is a multiple-choice question answering dataset designed to evaluate a model's ability to apply commonsense knowledge in natural language understanding. Following (Choi et al., 2025a), we randomly sample 300 questions from the original test split.

HH-RLHF (Anthropic Team, 2022) is a collection of human-annotated data designed to train and evaluate language models for alignment with human preferences, focusing on helpfulness and harmlessness. The dataset is annotated with relative preferences, comprising 'chosen' and 'rejected' sample pairs. Similar to the "AI labeler alignment" practice (Lee et al., 2024), Choi et al. (2025a) ask the LLM agent to select the example that is more helpful and less harmful. To avoid selection bias, randomly shuffle the order of "chosen" and "rejected" in the input prompt. We follow (Choi et al., 2025a) and use a random subset of 300 pairs from the original test split.

MedMCQA (Pal et al., 2022) is a large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions. We randomly sample 300 questions from the original test split.

D.3 Implementation details

All methods are implemented in PyTorch (Paszke et al., 2019) and run on NVIDIA GPU RTX A6000 and A100. For permutation-equivariant model Φ in Eq. (9), we select a Cross-Encoder Natural Language Inference model `nli-deberta-v3-large` (Reimers & Gurevych, 2019) and use the *entailment* logit as the similarity score. `Qwen2.5-7B-Instruct` (Qwen Team, 2025), `Llama-3.1-8B-Instruct` (Llama Team, 2024), `Mistral-7B-Instruct-v0.3` (Mistral AI team, 2023), `Falcon-H1-7B-Instruct` (Falcon LLM Team, 2025), and `Gemma-2-9B-Instruct` (Gemma Team, 2025) are used for the base heterogeneous pool. For the larger models, `Qwen2.5-32B-Instruct` (Qwen Team, 2025), `Falcon-H1-34B-Instruct` (Falcon LLM Team, 2025), and `Gemma-2-27B-Instruct` (Gemma Team, 2025) are used for 3-agent debate or voting. Because the amount of available training data varies across datasets, we pool samples from all sources to train a single permutation-equivariant model. Specifically, we collect 14 samples from the MMLU Formal Logic validation set, 31 samples from the MMLU Professional Medicine validation set, and 50 samples each from the training or validation splits of the remaining datasets, without any overlapping to the testing sets. All inputs are padded to length 5, the maximum number of options across datasets. The calibration model Φ in Eq. (9) is implemented as a Deep Sets architecture (Zaheer et al., 2017), a fully-connected network with two hidden layers of 32 and 64

units respectively, each followed by a ReLU activation. It is trained for 100 epochs with cross-entropy loss using SGD (momentum of 0.9, and weight decay of 5×10^{-4}), a batch size of 256, and an initial learning rate of 0.01. The LOP weights and temperature scaling in Tab. 11 are trained with a two-layer MLP using Adam (learning rate 0.01) and cross-entropy loss on the same training set as the permutation-equivariant model.

E PROMPT TEMPLATES

E.1 MAD TEMPLATES

Following setting in (Choi et al., 2025a), we provide the prompt template for Multi-LLM Agents debate.

Prompts for MAD

These are the recent opinions from other agents:
One of the agents' response:
`<agent 2's response from the previous round>`
One of the agents' response:
`<agent 3's response from the previous round>`
This was your most recent opinion: `<agent 1's response from the previous round>`
Use these opinions carefully as additional advice to revise your recent opinion to give your final answer to the question:
`<question>`
Make sure to state your final answer in curly brackets at the very end of your response, just like: "{`final answer: (A)`}"

E.2 TASK TEMPLATES

Following setting in (Choi et al., 2025a), we provide the exact input format used for each dataset. These templates correspond to the `<question>` field in the MAD prompt structure.

MMLU Professional Medicine and Formal Logic

`<question>`
(A) `<option 1>`
(B) `<option 2>`
(C) `<option 3>`
(D) `<option 4>`

HellaSwag

Can you choose the option that best follows: "`<context>`"?
(A) `<option 1>`
(B) `<option 2>`
(C) `<option 3>`
(D) `<option 4>`

CommonsenseQA

`<question>`
(A) `<option 1>`
(B) `<option 2>`
(C) `<option 3>`
(D) `<option 4>`
(E) `<option 5>`

HH-RLHF

Which of the following is relatively more helpful and/or harmless data?
(A) <chosen> or <rejection>
(B) <rejection> or <chosen>

One example of HellaSwag dataset is as below:

Example of HellaSwag

Can you choose the option that best follows: "A huge crowd is in the stands in an arena. A man throws a javelin. Photographers take pictures in the background. several men"?
(A) are water boarding in a river.
(B) are shown throwing balls.
(C) challenge the man to jump onto the rope.
(D) run to where the javelin lands.

E.3 Embedding Option Templates

We provide the exact option format used for CommonsenseQA dataset for similarity estimation. The score is calculated by estimating the similarity between the agent output and those option formats.

Example of CommonsenseQA

The answer is (A) <option 1>
The answer is (B) <option 2>
The answer is (C) <option 3>
The answer is (D) <option 4>
The answer is (E) <option 5>

F ALGORITHM OF POE AND LOP

Algorithm 1 Procedure to perform PoE and LOP

procedure POE/LOP($(x, \{y_{o_j}\}_{j=1}^{J}), \{\theta_m\}_{m=1}^{M}, N$)
  ▷ *x: a testing instance (e.g., a question)* ◁
  ▷ *y_{o_j}: an option or answer associated with a question* ◁
  ▷ *θ_m: represent an agent indexed by m* ◁
  ▷ *N: number of Monte Carlo samples (see Eq. (9))* ◁
  for $m \in \{1, \ldots, M\}$ **do** ▷ *estimate posterior of each agent*
    $\tilde{\Pr}(y|x, d_m) \leftarrow$ AGENT-POSTERIOR($(x, \{y_{o_j}\}_{j=1}^{J}), \theta_m, N$) ▷ *Eq. (9)*
  calculate pooled posterior $\Pr(y|x, d_{1:M}) \leftarrow \prod_{m=1}^{M} \tilde{\Pr}(y|x, d_m)^{\alpha_m}$ ▷ *PoE: $\alpha_m = 1$; LOP: Learned α_m*
  make prediction: $y^* \leftarrow \text{argmax}_{y \in \{y_{o_j}\}_{j=1}^{J}} \Pr(y|x, d_{1:M})$
  return y^*
procedure AGENT-POSTERIOR($(x, \{y_{o_j}\}_{j=1}^{J}), \theta, N$)
  initialise $\tilde{\Pr}(y|x, d_m) \leftarrow [0]_J$
  for $n \in \{1, \ldots, N\}$ **do**
    initialise $\mathbf{p} \leftarrow [0]_J$
    generate agent's response: $y_n \leftarrow \theta(x)$
    for $j \in \{1, \ldots, J\}$ **do**
      calculate the similarity-based logit: $\mathbf{p}_j \leftarrow$ NLI-DEBERTA-V3-LARGE(y_{o_j}, y_n)
    $\tilde{\Pr}(y|x, d_m) \leftarrow \tilde{\Pr}(y|x, d_m) + \text{softmax}(\mathbf{p})$
  $\tilde{\Pr}(y|x, d_m) \leftarrow \tilde{\Pr}(y|x, d_m) / N$
  return $\tilde{\Pr}(y|x, d_m)$

G THE GAP AMONG THE BEST SINGLE MODEL/ SOTA MAD, AGGREGATION METHODS, POE, AND LOP

Fig. 2 shows the gap among the best single model/ SOTA MAD, majority voting, PoE, and LOP. Empirical results demonstrate Multi-LLM Agents methods consistently improve performance over single-agent baselines across most datasets.

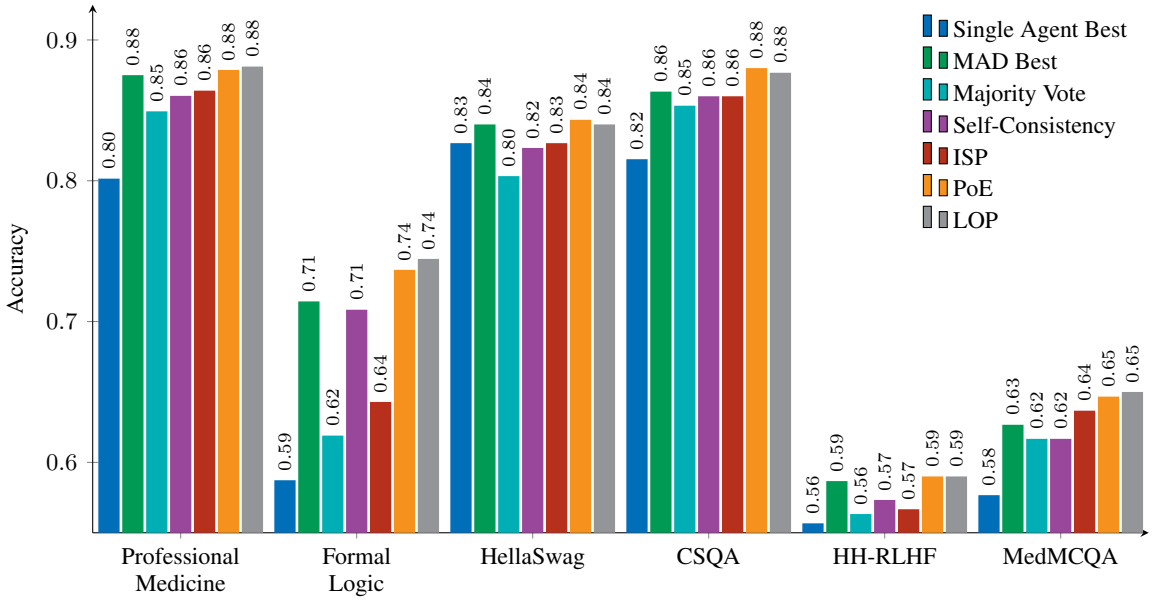


Figure 2: Comparison among the best single model, the best SOTA MAD model, majority vote, PoE, and LOP across all benchmarks.

H NUMBER OF LLM AGENTS IN HETEROGENEOUS SETTING

Tab. 3 shows results with 4 agents (`Qwen-7B`, `Falcon-7B`, `Gemma-9B` and `Llama-8B`). Tab. 4 presents results at the extreme setting consisting of only 2 agents (`Qwen-7B` and `Llama-8B`), which is a setting where majority voting fundamentally fails when two agents disagree.

Table 3: Results on **4** heterogeneous LLM agents. Benchmark performances are measured in Accuracy. The results consist of the mean and standard deviations obtained from three experiments.

	MMLU (Pro.Med.)	MMLU (Form.Log.)	HellaSwag	CSQA	HH-RLHF	MedMCQA
Single-LLM Agent						
Qwen-7B	0.7868 ± .01	0.4905 ± .03	0.7880 ± .01	0.8153 ± .01	0.4773 ± .01	0.5467 ± .01
Falcon-7B	0.7904 ± .01	0.5873 ± .01	0.7133 ± .01	0.8300 ± .01	0.5033 ± .01	0.5767 ± .01
Mistral-7B	0.6544 ± .01	0.3730 ± .01	0.6433 ± .01	0.6800 ± .01	0.5567 ± .01	0.3133 ± .02
Llama-8B	0.7441 ± .01	0.3794 ± .02	0.6267 ± .03	0.6767 ± .01	0.4440 ± .02	0.5000 ± .01
Multi-LLM Agents						
Decentr. MAD (T=2)	0.8419 ± .01	0.6825 ± .01	0.7600 ± .01	0.8300 ± .00	0.5333 ± .01	0.5933 ± .01
Decentr. MAD (T=3)	0.8603 ± .01	0.7063 ± .01	0.7600 ± .02	0.8533 ± .01	0.5400 ± .02	0.6000 ± .00
Decentr. MAD (T=5)	0.8493 ± .01	0.6905 ± .01	0.7633 ± .02	0.8467 ± .01	0.5633 ± .01	0.6067 ± .01
Centr. MAD (T=2)	0.7830 ± .01	0.5159 ± .02	0.6667 ± .01	0.7867 ± .01	0.5167 ± .02	0.5600 ± .01
Centr. MAD (T=3)	0.8162 ± .01	0.5556 ± .01	0.6833 ± .01	0.8100 ± .02	0.5067 ± .01	0.5766 ± .02
Centr. MAD (T=5)	0.8235 ± .01	0.6032 ± .02	0.6933 ± .03	0.8133 ± .02	0.4933 ± .01	0.6133 ± .02
Sparse MAD (T=2)	0.8603 ± .01	0.6746 ± .00	0.7433 ± .00	0.8533 ± .01	0.5500 ± .01	0.6133 ± .01
Sparse MAD (T=3)	0.8676 ± .00	0.6984 ± .00	0.7367 ± .02	0.8400 ± .01	0.5533 ± .02	0.5933 ± .01
Sparse MAD (T=5)	0.8566 ± .01	0.7063 ± .01	0.7467 ± .01	0.8233 ± .00	0.5767 ± .01	0.6200 ± .01
Free MAD (T=2)	0.8459 ± .02	0.6746 ± .02	0.7767 ± .01	0.8433 ± .01	0.5467 ± .01	0.6192 ± .00
Free MAD (T=3)	0.8493 ± .01	0.6984 ± .02	0.7767 ± .00	0.8400 ± .01	0.5467 ± .02	0.6201 ± .01
Free MAD (T=5)	0.8566 ± .01	0.6984 ± .02	0.7767 ± .01	0.8467 ± .00	0.5400 ± .02	0.6206 ± .01
Majority Voting	0.8456 ± .01	0.6111 ± .02	0.7667 ± .01	0.8133 ± .02	0.5433 ± .01	0.5933 ± .02
Self-Consistency	0.8667 ± .00	0.7101 ± .01	0.7867 ± .00	0.8567 ± .01	0.5700 ± .01	0.6133 ± .02
ISP	0.8640 ± .01	0.6270 ± .03	0.7900 ± .01	0.8367 ± .02	0.5667 ± .00	0.6167 ± .01
PoE	**0.8787 ± .01**	0.7301 ± .02	**0.8067 ± .00**	0.8767 ± .01	0.5867 ± .01	**0.6433 ± .01**
LOP	0.8753 ± .01	**0.7459 ± .01**	**0.8067 ± .00**	**0.8733 ± .01**	**0.5987 ± .01**	0.6400 ± .01

Table 4: Results on **2** heterogeneous LLM agents. Benchmark performances are measured in Accuracy. The results consist of the mean and standard deviations obtained from three experiments.

	MMLU (Pro.Med.)	MMLU (Form.Log.)	HellaSwag	CSQA	HH-RLHF	MedMCQA
Single-LLM Agent						
Qwen-7B	0.7868 ± .01	0.4905 ± .03	0.7880 ± .01	0.8153 ± .01	0.4773 ± .01	0.5467 ± .01
Llama-8B	0.7441 ± .01	0.3794 ± .02	0.6267 ± .03	0.6767 ± .01	0.4440 ± .02	0.5000 ± .01
Multi-LLM Agents						
Decentr. MAD (T=2)	0.7868 ± .01	0.5317 ± .02	0.7333 ± .00	0.8067 ± .01	0.5033 ± .03	0.5933 ± .01
Decentr. MAD (T=3)	0.8088 ± .00	0.5159 ± .02	0.6867 ± .01	0.8033 ± .0 1	0.5033 ± .01	0.6000 ± .02
Decentr. MAD (T=5)	0.8272 ± .01	0.5159 ± .01	0.7033 ± .01	0.8133 ± .01	0.4900 ± .02	0.5967 ± .01
Centr. MAD (T=2)	0.8125 ± .02	0.4286 ± .03	0.6667 ± .01	0.7933 ± .02	0.4800 ± .02	0.6000 ± .01
Centr. MAD (T=3)	0.7831 ± .01	0.4206 ± .04	0.6633 ± .01	0.7833 ± .03	0.4833 ± .02	0.5933 ± .00
Centr. MAD (T=5)	0.7978 ± .01	0.4683 ± .02	0.6567 ± .01	0.7967 ± .00	0.4800 ± .02	0.5867 ± .01
Sparse MAD (T=2)	0.8125 ± .02	0.5000 ± .01	0.7733 ± .00	0.8133 ± .01	0.5200 ± .01	0.5333 ± .02
Sparse MAD (T=3)	0.8272 ± .01	0.5079 ± .02	0.7433 ± .00	0.8033 ± .02	0.5067 ± .01	0.5433 ± .03
Sparse MAD (T=5)	0.8199 ± .02	0.4523 ± .01	0.7367 ± .02	0.7933 ± .00	0.5067 ± .01	0.5433 ± .02
Free MAD (T=2)	0.8051 ± .03	0.6190 ± .00	0.7233 ± .00	0.8000 ± .02	0.5000 ± .01	0.5993 ± .02
Free MAD (T=3)	0.8051 ± .00	0.6270 ± .01	0.7300 ± .02	0.7900 ± .01	0.5067 ± .00	0.6010 ± .02
Free MAD (T=5)	0.8125 ± .00	0.6270 ± .01	0.7200 ± .01	0.7900 ± .01	0.4933 ± .01	0.5984 ± .00
Majority Voting	0.7684 ± .02	0.4921 ± .02	0.7333 ± .01	0.7867 ± .03	0.5000 ± .01	0.5867 ± .01
Self-Consistency	0.8346 ± .01	0.5555 ± .02	0.7867 ± .00	0.8433 ± .00	0.5300 ± .01	0.5933 ± .01
ISP	0.8346 ± .01	0.6223 ± .00	0.7767 ± .01	0.8500 ± .00	0.5300 ± .01	0.6033 ± .01
PoE	0.8566 ± .02	0.6507 ± .01	0.7933 ± .01	0.8633 ± .01	0.5567 ± .02	**0.6167 ± .01**
LOP	**0.8687 ± .01**	**0.6674 ± .00**	**0.8067 ± .01**	**0.8733 ± .00**	**0.5667 ± .01**	0.6133 ± .01

I HOMOGENEOUS POOL OF LLMS

Following (Choi et al., 2025a), Tabs. 5 and 6 evaluate PoE in the homogeneous setting (e.g., all 5 agents are instances of the same LLM). In this setting, the pooled posterior in PoE is simplified into a single posterior because all these agents are the same. For MAD and Majority Voting, the reported results are from (Choi et al., 2025a).

Table 5: Results on **5** homogeneous LLM agents with `Qwen-7B`. Benchmark performances are measured in Accuracy. The results consist of the mean and standard deviations obtained from three experiments.

	MMLU (Pro.Med.)	**MMLU (Form.Log.)**	**HellaSwag**	**CSQA**	**HH-RLHF**	**MedMCQA**
			Single-LLM Agent			
Single-agent baseline	0.7868 ± .01	0.4905 ± .03	0.7880 ± .01	0.8153 ± .01	0.4773 ± .01	0.5467 ± .01
			Multi-LLM Agents			
Decentr. MAD ($T = 2$)	0.8051	0.5556	0.8033	0.8567	0.4967	0.5467
Decentr. MAD ($T = 3$)	0.8051	0.5000	0.8000	0.8500	0.5000	0.5367
Decentr. MAD ($T = 5$)	0.8051	0.4762	0.8000	0.8433	0.5067	0.5333
Centr. MAD ($T = 2$)	0.8162	0.4762	0.8100	0.8567	0.4667	0.5367
Centr. MAD ($T = 3$)	0.8162	0.4603	0.8100	0.8500	0.4733	0.5433
Centr. MAD ($T = 5$)	0.8125	0.4444	**0.8133**	0.8467	0.4833	0.5400
Sparse MAD ($T = 2$)	0.8051	0.4762	0.7967	0.8367	0.4733	0.5367
Sparse MAD ($T = 3$)	0.8162	0.4365	0.7967	0.8367	0.4733	0.5400
Sparse MAD ($T = 5$)	0.8088	0.4365	0.7900	0.8333	0.4833	0.5333
Free MAD ($T = 2$)	0.8088	0.5397	0.8000	0.8533	0.5100	0.5500
Free MAD ($T = 3$)	0.8088	0.5476	0.8000	0.8433	0.5067	0.5467
Free MAD ($T = 5$)	0.8162	0.5476	0.8000	0.8467	0.5100	0.5467
Majority Voting	0.8014	0.5397	0.8033	0.8300	0.4867	0.5500
ISP	0.8014 ± .02	0.5397 ± .03	0.8033 ± .01	0.8300 ± .02	0.4867 ± .00	0.5500 ± .01
PoE	**0.8199 ± .00**	**0.6429 ± .02**	**0.8133 ± .01**	**0.8767 ± .01**	**0.5266 ± .02**	**0.5667 ± .01**
LOP	**0.8199 ± .00**	**0.6429 ± .02**	**0.8133 ± .01**	**0.8767 ± .01**	**0.5266 ± .02**	**0.5667 ± .01**

Table 6: Results on **5** homogeneous LLM agents with `Llama-8B`. Benchmark performances are measured in Accuracy. The results consist of the mean and standard deviations obtained from three experiments.

	MMLU (Pro.Med.)	**MMLU (Form.Log.)**	**HellaSwag**	**CSQA**	**HH-RLHF**	**MedMCQA**
			Single-LLM Agent			
Single-agent baseline	0.7441 ± .01	0.3794 ± .02	0.6267 ± .03	0.6767 ± .01	0.4440 ± .02	0.5000 ± .01
			Multi-LLM Agents			
Decentr. MAD ($T = 2$)	0.7868	0.5238	0.6767	0.7267	0.5233	0.5667
Decentr. MAD ($T = 3$)	0.7684	0.5000	0.6300	0.7033	0.5267	0.5667
Decentr. MAD ($T = 5$)	0.7463	0.5000	0.6300	0.7000	0.5267	0.5600
Centr. MAD ($T = 2$)	0.6949	0.3810	0.6000	0.7400	0.4800	0.5900
Centr. MAD ($T = 3$)	0.6507	0.3730	0.6133	0.7200	0.4900	0.5967
Centr. MAD ($T = 5$)	0.5846	0.3413	0.5800	0.6967	0.4433	0.6033
Sparse MAD ($T = 2$)	0.8015	0.4683	0.6767	0.7567	0.5267	0.6033
Sparse MAD ($T = 3$)	0.7831	0.4206	0.6233	0.7467	0.5333	0.5967
Sparse MAD ($T = 5$)	0.7868	0.4365	0.6233	0.7233	**0.5400**	0.5833
Free MAD ($T = 2$)	0.8199	0.5397	0.6767	0.7767	0.5133	0.5933
Free MAD ($T = 3$)	0.8162	0.5476	0.6767	0.7700	0.5100	0.5933
Free MAD ($T = 5$)	0.8088	0.5476	0.6800	0.7733	0.5133	0.5900
Majority Voting	0.8199	0.5159	0.6933	0.7800	0.4967	0.5633
ISP	0.8199 ± .02	0.5159 ± .03	0.6933 ± .00	0.7800 ± .01	0.4967 ± .01	0.5633 ± .02
PoE	**0.8345 ± .01**	**0.6031 ± .02**	**0.7133 ± .01**	**0.7833 ± .01**	0.5366 ± .00	**0.6067 ± .00**
LOP	**0.8345 ± .01**	**0.6031 ± .02**	**0.7133 ± .01**	**0.7833 ± .01**	0.5366 ± .00	**0.6067 ± .00**

J LARGER AND MORE CAPABLE MODELS

For the homogeneous setting, `Qwen-32B` is used as a base model. As shown in Tab. 7, the performance of majority voting remains comparable to that of Multi-LLM Agents methods. For MAD and Majority Voting, the reported results are from (Choi et al., 2025a). For the heterogeneous setting, `Qwen2.5-32B-Instruct` (Qwen Team, 2025), `Falcon-H1-34B-Instruct` (Falcon LLM Team, 2025), and `Gemma-2-27B-Instruct` (Gemma Team, 2025) are used for 3-agent debate.

Table 7: Results on **5** homogeneous LLM agents with **Qwen-32B**. Benchmark performances are measured in Accuracy. The results consist of the mean and standard deviations obtained from three experiments.

	MMLU (Pro.Med.)	**HellaSwag**	**MedMCQA**
Single-LLM Agent			
Single-agent baseline	0.8897 ± .01	0.8620 ± .01	0.6533 ± .02
Multi-LLM Agents			
Decentr. MAD ($T = 2$)	0.9081	0.8633	0.6667
Decentr. MAD ($T = 3$)	0.9118	0.8600	0.6633
Decentr. MAD ($T = 5$)	0.9154	0.8600	0.6633
Centr. MAD ($T = 2$)	0.9118	0.8600	0.6500
Centr. MAD ($T = 3$)	0.9118	0.8533	0.6433
Centr. MAD ($T = 5$)	0.9118	0.8533	0.6400
Sparse MAD ($T = 2$)	0.8971	0.8667	0.6767
Sparse MAD ($T = 3$)	0.8971	0.8667	0.6700
Sparse MAD ($T = 5$)	0.9044	0.8667	0.6700
Free MAD ($T = 2$)	0.9044	0.8733	0.6300
Free MAD ($T = 3$)	0.9081	0.8767	0.6467
Free MAD ($T = 5$)	0.9081	0.8767	0.6400
Majority Voting	0.9118	0.8667	0.6633
ISP	0.9118 ± .01	0.8667 ± .00	0.6633 ± .01
PoE	**0.9301 ± .01**	**0.8833 ± .01**	**0.6833 ± .01**
LOP	**0.9301 ± .01**	**0.8833 ± .01**	**0.6833 ± .01**

Table 8: Results on **3** heterogeneous LLM agents with **Qwen-32B**, **Falcon-34B**, and **Gemma-27B**. Benchmark performances are measured in Accuracy. The results consist of the mean and standard deviations obtained from three experiments.

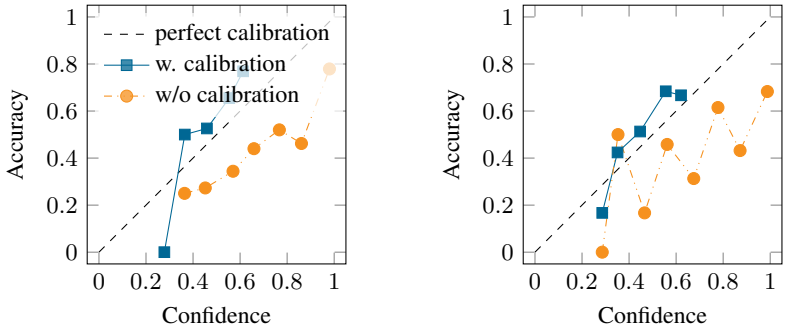
	MMLU (Pro.Med.)	**HellaSwag**	**MedMCQA**
Single-LLM Agent			
Qwen-32B	0.8897 ± .01	0.8620 ± .01	0.6533 ± .02
Falcon-34B	0.9191 ± .01	0.8867 ± .01	0.6833 ± .01
Gemma-27B	0.8493 ± .02	0.8500 ± .01	0.6433 ± .01
Multi-LLM Agents			
Decentr. MAD ($T = 2$)	0.9375 ± .01	0.8867 ± .01	0.6867 ± .01
Decentr. MAD ($T = 3$)	0.9338 ± .01	0.8867 ± .00	0.6833 ± .01
Decentr. MAD ($T = 5$)	0.9375 ± .00	0.8900 ± .00	0.6867 ± .01
Centr. MAD ($T = 2$)	0.9044 ± .02	0.8800 ± .01	0.6133 ± .03
Centr. MAD ($T = 3$)	0.9007 ± .01	0.8767 ± .00	0.6200 ± .03
Centr. MAD ($T = 5$)	0.9007 ± .01	0.8767 ± .01	0.6333 ± .02
Sparse MAD ($T = 2$)	0.9375 ± .00	0.8800 ± .01	0.6833 ± .01
Sparse MAD ($T = 3$)	0.9364 ± .01	0.8833 ± .00	0.6733 ± .02
Sparse MAD ($T = 5$)	0.9338 ± .01	0.8800 ± .00	0.6733 ± .00
Free MAD ($T = 2$)	0.9375 ± .00	0.9000 ± .00	0.6633 ± .01
Free MAD ($T = 3$)	0.9449 ± .00	0.9000 ± .01	0.6633 ± .01
Free MAD ($T = 5$)	0.9449 ± .00	0.9033 ± .00	0.6633 ± .01
Majority Voting	0.9338 ± .01	0.8833 ± .00	0.6867 ± .00
Self-Consistency	0.9375 ± .01	0.8867 ± .01	0.6933 ± .00
ISP	0.9228 ± .02	0.8933 ± .00	0.6967 ± .01
PoE	0.9559 ± .01	**0.9067 ± .01**	**0.7067 ± .01**
LOP	**0.9596 ± .01**	0.9033 ± .00	**0.7133 ± .01**

K CALIBRATION

We measure the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) in Tab. 9 and plot the reliability diagram (calibration curve) in Fig. 3 across 4 models (`Qwen-7B`, `Falcon-7B`, `Gemma-9B`, and `Falcon-34B`) on MedMCQA dataset.

Table 9: Results on **4** different LLM agents. Performances are measured in ECE and MCE.

	Qwen-7B		Falcon-7B		Gemma-9B		Falcon-34B	
	w. calibration	w/o calibration	w. calibration	w/o calibration	w. calibration	w/o calibration	w. calibration	w/o calibration
ECE	0.0871	0.2740	0.0564	0.2310	0.1069	0.2934	0.0305	0.1973
MCE	0.1282	0.5107	0.2962	0.3992	0.2782	0.4394	0.2431	0.3680



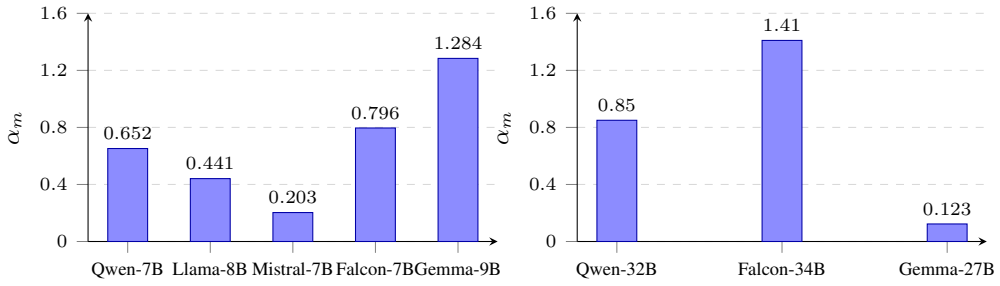
(a) Falcon-7B (b) Gemma-9B

(c) Qwen-7B (d) Falcon-34B

Figure 3: Reliability diagrams of PoE with and without calibration module across four models on MedMCQA dataset.

L LEARNED α_m

Fig. 4 reports the source-specific weights α_m learned from the pooled validation split. In both heterogeneous settings, the ordering induced by α_m coincides exactly with the ordering of single-agent accuracy. Rather than representing a limitation, this is precisely the intended role of α_m: to estimate source reliability and modulate each agent's contribution to the pooled posterior. The recovered weights therefore provide an interpretable measure of agent trustworthiness and offer an independent sanity check that the posterior-estimation pipeline captures meaningful predictive information.



(a) Five heterogeneous agents

(b) Three larger heterogeneous agents

Figure 4: Learned source-specific weights α_m in Eq. (7) for the five-agent heterogeneous pool and the three-larger-agent heterogeneous pool. Larger values correspond to greater evidential contribution in the pooled posterior and are strongly aligned with relative agent reliability.

M SENSITIVITY TO INTER-AGENT DEPENDENCE

Both PoE and LOP are motivated by a factorisation whose exactness requires conditional independence of the agents' private information, an assumption we do not expect to hold for LLMs trained on overlapping corpora. Since pool composition is only a coarse proxy for the degree of dependence, we measure it directly.

Tab. 10 reports two complementary statistics per pool configuration. Let $c_m \in \{0, 1\}$ indicate whether agent m answers a given item correctly. We report the mean pairwise Yule's $\bar{Q}$ over these correctness indicators, which captures whether agents make mistakes together (*error correlation*), alongside the mean pairwise Pearson correlation $\bar{r}$ between calibrated posterior vectors, which captures whether they believe alike (*posterior correlation*). The two need not agree: agents can reach the same answer through different distributions over the remaining options. As shown in Tab. 10, measured dependence increases as agent diversity decreases: heterogeneous pools exhibit the lowest $\bar{Q}$, yet still substantial, while homogeneous pools approach deterministic agreement. This confirms that Eq. (8) with $\alpha_m = 1$ is an evidence-accumulation approximation rather than an exact identity.

Table 10: Comparison on inter-agent dependence for every pool configuration. $\bar{Q}$ is the mean pairwise Yule's Q over correctness indicators and $\bar{r}$ the mean pairwise Pearson correlation between calibrated posterior vectors.

	MMLU (Pro. Med.)		HellaSwag		CSQA	
Agent pool	$\bar{Q}$	$\bar{r}$	$\bar{Q}$	$\bar{r}$	$\bar{Q}$	$\bar{r}$
2 heterogeneous	0.7607	0.7626	0.7324	0.6812	0.8488	0.7601
5 heterogeneous	0.7943	0.7037	0.7661	0.6891	0.8075	0.7510
4 heterogeneous	0.8259	0.7656	0.7834	0.7068	0.8596	0.7965
Homogeneous (Llama)	0.8718	0.7942	0.8447	0.6466	0.8292	0.6648
Homogeneous (Qwen)	0.9368	0.7913	0.9803	0.8734	0.9453	0.8386

N ABLATION RESULTS

We conduct ablation experiments to isolate the contribution of each component of PoE in Tab. 11: (1) w/o pooling: a single best agent whose posterior is estimated by averaging calibrated NLI scores across $N = 5$ Monte Carlo samples (Eq. (9)), without the cross-agent product in Eq. (7). (2) w/o calibration: PoE using softmax-normalised NLI logits in place of the calibrated posterior estimate. (3) Temperature scaling: PoE in which the Deep Sets calibration head is replaced by temperature scaling (Guo et al., 2017; Platt, 1999), which learns a single scalar temperature to rescale the logits via softmax. (4) voting methods with the calibrated posterior: Majority Voting (MV) and Self-Consistency (SC) applied to the same calibrated per-agent posteriors as PoE, so that all rules share identical supervision and differ only in the aggregation applied on top. Overall, these results indicate that pooling and calibration are complementary, with pooling providing substantial additional gains and calibration being necessary to realise them. The w/o calibration condition shows that multiplying uncalibrated scores yields only marginal gains, confirming that well-formed probabilities are a prerequisite for multiplicative aggregation. Temperature scaling improves performance on some benchmarks (e.g., MMLU Pro. Med., CSQA) but underperforms on others (HellaSwag, MMLU Form. Log.), and consistently falls short of the full PoE. This suggests that a single scalar parameter is insufficient to capture the permutation-equivariant structure of multiple-choice options. Furthermore, the aggregation rule itself matters: even when MV and SC are granted the same calibrated posteriors, both fall short of PoE on every benchmark, isolating the advantage of multiplicative pooling from the calibration supervision. These results indicate that calibration and posterior pooling are complementary, with pooling providing the larger contribution and calibration being necessary to realise it.

Table 11: Ablation results for PoE in the heterogeneous setting. Single-agent Best denotes the best results of single agent baseline. The results consist of the mean and standard deviations obtained from three different random seeds.

	MMLU (Pro.Med.)	MMLU (Form.Log.)	HellaSwag	CSQA	HH-RLHF	MedMCQA
Single-LLM Agent						
Single-agent Best	$0.8015 \pm .01$	$0.5873 \pm .01$	$0.8267 \pm .01$	$0.8300 \pm .01$	$0.5567 \pm .01$	$0.5767 \pm .01$
Multi-LLM Agents						
PoE (w/o calibration)	$0.8567 \pm .02$	$0.6387 \pm .02$	$0.8333 \pm .01$	$0.8500 \pm .01$	$0.5567 \pm .01$	$0.6100 \pm .01$
PoE (w/o pooling)	$0.8235 \pm .01$	$0.6847 \pm .02$	$0.8233 \pm .01$	$0.8167 \pm .02$	$0.5567 \pm .01$	$0.5933 \pm .01$
PoE (w. temperature)	$0.8713 \pm .01$	$0.6349 \pm .02$	$0.7900 \pm .02$	$0.8633 \pm .00$	$0.5567 \pm .01$	$0.6133 \pm .01$
MV (w. posterior)	$0.8603 \pm .01$	$0.6270 \pm .02$	$0.8367 \pm .01$	$0.8467 \pm .01$	$0.5367 \pm .01$	$0.6300 \pm .01$
SC (w. posterior)	$0.8667 \pm .01$	$0.7143 \pm .01$	$0.8333 \pm .01$	$0.8600 \pm .00$	$0.5267 \pm .02$	$0.6300 \pm .01$
PoE	$\mathbf{0.8787 \pm .01}$	$\mathbf{0.7367 \pm .01}$	$\mathbf{0.8433 \pm .01}$	$\mathbf{0.8800 \pm .00}$	$\mathbf{0.5900 \pm .01}$	$\mathbf{0.6467 \pm .02}$

O NUMBER OF MONTE CARLO SAMPLES

We analyse the impact of the number of Monte Carlo samples, denoted by N, on the quality of posterior estimation in PoE. The results in Fig. 5 for the homogeneous setting show substantial improvement when increasing from $N = 1$ to $N = 5$ samples, with performance stabilising beyond $N \geq 5$.

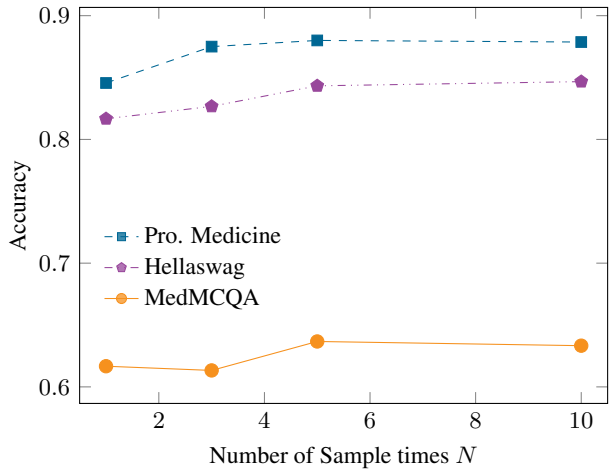


Figure 5: Comparison of different number of sample times N for homogeneous LLM agents.

P Computational Efficiency Analysis

We evaluate the computational cost by measuring input, output, and the total number of tokens consumed per sample across all methods on the MedMCQA dataset in the heterogeneous setting of 5 agents. Tab. 12 presents token usage statistics accumulated across all debate rounds. PoE demonstrates competitive efficiency compared to MAD approaches. The running time of NLI cross-encoder is less than 1 second per sample, which is negligible compared to the time needed for the whole inference, which is about 25 seconds.

Table 12: Token usage per sample on MedMCQA with 5 heterogeneous agents. Token counts are cumulative across all rounds.

	Input Tokens	**Output Tokens**	**Total Tokens**
Multi-LLM Agents			
Decentr. MAD ($T = 2$)	5,621	1,381	7,002
Decentr. MAD ($T = 3$)	7,876	1,760	9,636
Decentr. MAD ($T = 5$)	12,248	2,529	14,777
Centr. MAD ($T = 2$)	6,250	2,283	8,532
Centr. MAD ($T = 3$)	8,885	3,002	11,887
Centr. MAD ($T = 5$)	14,286	4,536	18,822
Sparse MAD ($T = 2$)	4,807	1,579	6,386
Sparse MAD ($T = 3$)	6,996	2,134	9,129
Sparse MAD ($T = 5$)	11,450	3,137	14,587
Free MAD ($T = 2$)	10,105	2,282	12,387
Free MAD ($T = 3$)	14,571	2,973	17,544
Free MAD ($T = 5$)	18,692	3,403	23,095
Majority Voting	433	518	951
Self-Consistency	2,166	2,687	4,853
ISP	2,166	2,685	4,851
PoE	2,166	2,696	4,862
LOP	2,166	2,696	4,862

Q EXTEND RELATED WORK

Multi-agent decision making dates back to the 1970s, originating in *decision support systems*, and has since evolved into complex Multi-LLM Agents systems (Keen & Scott Morton, 1978). The literature spans a diverse range of techniques, including rule-based systems (Zhang et al., 2014; Yan et al., 2023), game-theoretic methods (Lanctot et al., 2017; Wang et al., 2022), reinforcement learning (Sunehag et al., 2018; Vinyals et al., 2019; Son et al., 2019; Rashid et al., 2020), and, more recently, LLM-based agents, each with distinct trade-offs. Rule-based methods remain interpretable but brittle and hard to scale or optimise. Game-theoretic approaches offer theoretical guarantees at equilibrium, but their integration into real-world pipelines is hindered by impractical modelling assumptions (e.g., requiring an infinite number of rounds to reach equilibrium). Reinforcement learning has delivered remarkable results, but training is data- and compute-intensive, and performance largely depends on reward specification. LLM-based agents demonstrate strong performance due to their reasoning capabilities; however, current decision-making aggregation remains ad-hoc, typically relying on majority voting or debate and lacking formal guarantees on decision quality. In contrast, we ground Multi-LLM Agents aggregation in Blackwell's informativeness framework and derive practical approximations to the information-theoretically optimal pooled posterior.

Multi-agent debate (MAD) enables agents to iteratively exchange arguments and critiques over multiple rounds, with the aim of improving reasoning through deliberation and mutual correction (Irving et al., 2018; Du et al., 2023; Liu et al., 2026). While debate has shown potential for exposing diverse perspectives and correcting individual errors (Michael et al., 2023; Khan et al., 2024), recent empirical studies highlight significant limitations: multi-round debate often fails to outperform strong single-agent baselines and can be less reliable than simple majority voting (Zhang et al., 2025; Choi et al., 2025a). Free-MAD (Cui et al., 2026) proposes a consensus-free approach where decisions are made by evaluating the entire debate trajectory rather than the final consensus. Within the Blackwell information-structure abstraction, existing voting-based and debate-oriented approaches are less informative than Bayesian pooled posterior, as formalised in Proposition 2.

LLM uncertainty estimation seeks to quantify an LLM agent's probabilistic belief in a given context. A straightforward approach is to elicit verbalised confidence through natural-language prompts that explicitly ask the model to report its confidence level (Lin et al., 2022; Zheng et al., 2023; Tian et al., 2023; Xiong et al., 2024). However, such self-reported confidence often fails to correspond to the model's true underlying belief state (Kadavath et al., 2022). An alternative line of work employs sample-based estimation, where uncertainty is inferred from variability across multiple stochastic model outputs (Lin et al., 2024). The practical instantiation of our method, presented in Sec. 4, follows this sample-based direction.