

Weighted Proportional Conflict Redistribution for Heterogeneous LLM Ensembles

ANAS DAHABIAH
Faculty of Information Technology
Al-Sham Private University
Damascus
SYRIA

Abstract: - Smaller language models often exhibit high confidence in erroneous predictions, undermining traditional ensemble methods. In heterogeneous ensembles, simple averaging can grant disproportionate influence to overconfident but incorrect outputs. This paper introduces a reliability-weighted extension of the Proportional Conflict Redistribution rule (PCR6), where model-specific weights are derived from validation accuracy. We provide a formal treatment of the conflict measure, prove that the weighted rule avoids Zadeh's paradox in the high-conflict regime, and establish a quasi-associativity bound of order $O(K_{\max}^2)$. We further justify linear weighting from both Bayesian (conjugate Dirichlet) and minimax perspectives, and compare the approach with logarithmic pooling and Dirichlet calibration. Empirical evaluation on a multi-model MMLU setting shows that the primary practical benefit of weighted PCR6 in our experiments lies in improved calibration (substantially lower Expected Calibration Error) rather than large accuracy gains over simple averaging. Accuracy improvements appear most plausible when ensemble members differ markedly in reliability; when models are comparably strong, the method remains competitive while producing better-calibrated probabilities. The fusion step itself is computationally lightweight (tens of microseconds per query for small ensembles).

Key-Words: - ensemble learning, evidence theory, conflict resolution, language models, PCR6, Zadeh's paradox, quasi-associativity, reliability weighting, logarithmic pooling, Dirichlet calibration

Received: March 12, 2026. Revised: April 11, 2026. Accepted: May 22, 2026. Published: October 13, 2026.

1 Introduction

Combining language models of different sizes - from frontier systems such as GPT-4 [1] to small open models—balances cost and performance, but creates a problem: smaller models sometimes express high confidence in wrong answers. Simple averaging gives such errors too much weight, raising the question of whether accounting for model quality can do better.

We examine this through evidence theory: Dempster-Shafer combination has known issues with conflicting sources [4], [12], and while Proportional Conflict Redistribution addresses some of these it still treats all sources equally [5]. Our contribution weights the redistribution by empirical validation accuracy.

Theoretically, we prove that the weighted approach avoids classical paradoxes while maintaining desirable mathematical properties, justify the weighting scheme from Bayesian and minimax viewpoints, and compare it with logarithmic pooling and Dirichlet calibration. Empirically, we find that the most consistent

advantage in our experiments is improved probability calibration; accuracy gains over simple averaging are modest and appear to depend on the degree of reliability heterogeneity among the ensemble members.

The rest proceeds as follows. Section 2 covers background. Section 3 presents weighted PCR6 with formal definitions, complete proofs, an algorithm, Bayesian/minimax justification, and comparisons. Section 4 describes experiments. Section 5 analyzes results. Section 6 concludes.

2 Background

2.1 Recent Ensemble Approaches

Ensemble methods for language models have evolved considerably. Wang et al. (2024) [15] introduced Mixture-of-Agents with 8–12 pp gains at 5–10× latency cost. Jiang et al. (2023) [9] proposed LLM-Blender with a learned ranker. Chen et al. (2023) [2] developed FrugalGPT for cost optimization. Recent work [13], [14] continues to

highlight calibration challenges in modern LLMs. A separate line of work quantifies uncertainty within a single model rather than across an ensemble, e.g. semantic entropy over sampled generations (Kuhn et al., 2023 [10]) or black-box confidence elicitation (Xiong et al., 2024 [16]); such single-model uncertainty estimates are complementary to, rather than a substitute for, the cross-model fusion problem addressed here.

2.2 Dempster-Shafer and PCR

Dempster-Shafer theory represents uncertainty by assigning mass to sets of hypotheses [4], [12]. Dempster's rule has a well-known problem: normalizing away conflict leads to paradoxes when sources disagree strongly (Zadeh, 1979) [17]. PCR6 (Dezert & Smarandache, 2006) [5] redistributes conflict proportionally to squared commitments, but assumes equal source reliability—inappropriate when model quality varies substantially.

3 Method

3.1 Setup and Notation

Consider multiple-choice questions with answer set $\Theta = \{A, B, C, D\}$. Each model i produces a probability distribution p_i that is converted to a basic belief assignment:

$$m_i(\{X\}) = (1 - \varepsilon) p_i(X) \text{ for } X \in \Theta, \text{ and } m_i(\emptyset) = \varepsilon, \quad (1)$$

where $\varepsilon = 0.05$. This parameter represents residual uncertainty (hallucination / OOD risk) and prevents degeneracy when a model assigns near-zero mass to the correct answer. A direct sensitivity sweep on the test set ($\varepsilon \in \{0.01, 0.03, 0.05, 0.08, 0.10, 0.15, 0.20\}$) shows fused accuracy varies by at most 0.1 percentage points (68.6%–68.7%) and ECE varies by at most 0.004 (0.079–0.084) across this range, confirming the reported results are not an artifact of the specific choice $\varepsilon = 0.05$.

3.2 Formal Definition of the Conflict Measure

Definition 1 (Pairwise Conflict). For two BBAs m_i, m_j on Θ :

$$K_{ij} = 1 - \sum_{X \in \Theta} m_i(\{X\}) \cdot m_j(\{X\}). \quad (2)$$

Definition 2 (Global Conflict Level). For N sources:

$$K = (2 / (N(N-1))) \sum_{\{1 \leq i < j \leq N\}} K_{ij}. \quad (3)$$

$K_{\max} = \max_{\{i < j\}} K_{ij}$ is used in the quasi-associativity bound. As shown below, $K_{ij} \in [0,1]$ and is symmetric; $K_{ij} = 0$ exactly when both sources are certain and in full agreement (point

masses on the same hypothesis), and $K_{ij} = 1$ exactly when their singleton supports are disjoint—precisely the regime in which Dempster's rule normalizes by $(1-K) = 0$ and is undefined, while PCR6 remains well-defined by construction.

Proof of basic properties. Write $m_i(X)$ for $m_i(\{X\})$. (i) Boundedness: since $\sum_X m_i(X) \leq 1$ and $\sum_X m_j(X) \leq 1$ (the remainder being ignorance mass on $\emptyset$), the Cauchy–Schwarz inequality gives $\sum_X m_i(X)m_j(X) \leq \sqrt{(\sum_X m_i(X)^2) \cdot (\sum_X m_j(X)^2)} \leq 1$, so $K_{ij} = 1 - \sum_X m_i(X)m_j(X) \in [0,1]$. (ii) Symmetry is immediate from commutativity of the product inside the sum. (iii) $K_{ij} = 0$ requires $\sum_X m_i(X)m_j(X) = 1$; by the equality condition of Cauchy–Schwarz this forces both $\sum_X m_i(X)^2 = 1$ and m_i, m_j proportional, which (since each $m_i(X) \leq 1$) holds only when both BBAs place unit mass on the same singleton hypothesis—zero ignorance and full agreement. (iv) $K_{ij} = 1$ requires $\sum_X m_i(X)m_j(X) = 0$, i.e. the singleton supports of m_i and m_j are disjoint; this is exactly the case where Dempster's rule divides by $(1-K_{ij}) = 0$ and is undefined, whereas the PCR6 redistribution of Section 3.3 remains well-defined because it never divides by $(1-K)$. ■

3.3 Weighted PCR6

For sources i, j with weights $w_i, w_j > 0$ ($\sum w_k = 1$), the conflicting product $m_i(\{A\}) \cdot m_j(\{X\})$ ($X \neq A$) is redistributed as:

$$\Delta_i(A) = [w_i \cdot m_i(\{A\})^2 \cdot m_j(\{X\})] / [w_i \cdot m_i(\{A\}) + w_j \cdot m_j(\{X\})], \quad (4)$$

and symmetrically for $\Delta_j(X)$. The fused BBA adds the conjunctive consensus mass to all redistributed contributions. When $w_i \gg w_j$ the more reliable source recovers most of the conflicting mass.

Explicitly, summing every pairwise transfer together with the conjunctive consensus gives the closed-form fused mass on singleton A :

$$m_{\text{fused}}(\{A\}) = m_c(\{A\}) + \sum_{\{i=1\}^{\wedge} \{N\}} \sum_{\{j \neq i\}} \sum_{\{X \neq A\}} [w_i m_i(\{A\})^2 m_j(\{X\})] / [w_i m_i(\{A\}) + w_j m_j(\{X\})], \quad (5)$$

where $m_c(\{A\}) = \prod_{\{k=1\}^{\wedge} \{N\}} m_k(\{A\})$ is the conjunctive consensus mass. The outer sums range over every source i and every other source j , and the inner sum over every hypothesis X conflicting with A ; this is exactly what Algorithm 1 computes.

3.4 Algorithm

Algorithm 1 (Weighted PCR6 Fusion). Input: BBAs $m_1, \dots, m_N$ on Θ ; weights $w_1, \dots, w_N$ with $\sum w_i = 1$. (1) Initialize fused mass to zero. (2) Compute conjunctive consensus $m_c(\{\emptyset\}) = \prod_i m_i(\{\emptyset\})$ for

each singleton. (3) For every ordered pair (i,j) and distinct hypotheses (A,X) , add the weighted PCR6 transfer $\Delta_i(A)$ to the fused mass of A . (4) Add the consensus mass and optionally normalize. Complexity is $O(N^2 M^2)$ naively and effectively $O(N^2 M)$ for small M ($M=4$ in our experiments). Measured overhead for $N=3$ is on the order of tens of microseconds per query.

Theorem 1 (Computational Complexity).

Algorithm 1 runs in $O(N^2 M^2)$ time, which is $O(N^2 M)$ whenever the number of hypotheses M is held fixed ($M = O(1)$), as in every experiment reported here ($M = 4$).

Proof. Step (2) requires one multiplication per (source, hypothesis) pair, costing $O(NM)$. Step (3) enumerates all $N(N-1) = O(N^2)$ ordered pairs of distinct sources (i,j) ; for each pair it enumerates all $M(M-1) = O(M^2)$ ordered pairs of distinct hypotheses (A,X) and performs $O(1)$ arithmetic per pair to compute and accumulate the transfer Δ . This gives $O(N^2 M^2)$ total work for Step (3), which dominates Step (2) and the $O(M)$ normalization in Step (4). Hence the algorithm runs in $O(N^2 M^2)$ time overall, i.e. $O(N^2 M)$ when M is treated as a constant not scaling with problem size, consistent with the fixed four-way multiple-choice setting used throughout this paper. ■

3.5 Theoretical Properties

3.5.1 Weight Selection: Bayesian and Minimax Perspectives

Bayesian perspective. Let a_i be the unknown true accuracy of model i . A conjugate Dirichlet (or Beta) prior on the accuracy vector yields a posterior mean that, under a weakly informative prior and moderate validation size, is nearly proportional to the observed validation accuracies. Linear weights $w_i \propto \hat{a}_i$ therefore coincide with the Bayes estimator under standard conjugate analysis.

Minimax perspective. Under bounded ℓ_∞ perturbations of the empirical accuracy vector (modeling finite-sample noise or mild distribution shift), linear weighting is Lipschitz and does not amplify errors near the extremes, unlike Softmax or power transformations. It also never drives any weight exactly to zero when $\hat{a}_i > 0$, preserving ensemble diversity.

3.5.2 Comparison with Logarithmic Pooling and Dirichlet Calibration

Logarithmic pooling forms $p_{\log}(\theta) \propto \prod_i p_i(\theta)^{\{w_i\}}$. It is externally Bayesian but exhibits a

veto property: a single near-zero probability collapses the product. Weighted PCR6 redistributes rather than multiplies, avoiding this brittleness.

Dirichlet calibration [18] learns a linear map on log-probabilities and improves marginal calibration. It does not, however, perform explicit conflict redistribution. In our experiments it is a strong baseline for accuracy and remains complementary to conflict-aware fusion.

Proposition 1 (Zadeh's Paradox Avoidance).

Assumptions. Let sources i and j report singleton masses $\alpha = m_i(\{A\}) \in (0,1)$ and $\beta = m_j(\{B\}) \in (0,1)$ on disjoint hypotheses $A \neq B$, with reliability weights $w_i, w_j > 0$. Consider the high-conflict regime $\alpha, \beta \rightarrow 1-\epsilon$, i.e. both sources approach full, confident commitment to their respective, conflicting hypotheses, so $K_{ij} \rightarrow 1$ as $\alpha, \beta \rightarrow 1$.

Statement. (a) Under Dempster's rule of combination, the ratio of combined masses $m_D(A)/m_D(B) = [\alpha(1-\beta)] / [\beta(1-\alpha)]$ is independent of the reliability weights w_i, w_j , and becomes ill-conditioned as $\alpha, \beta \rightarrow 1$ (Zadeh's instability). (b) Under weighted PCR6, the ratio of the two dominant conflict-redistribution transfers satisfies, exactly and for all $\alpha, \beta \in (0,1)$:

$$\Delta_i(A) / \Delta_j(B) = (w_i / w_j) \cdot (\alpha / \beta), \quad (6)$$

which remains finite and is governed by the reliability ratio w_i/w_j even as $\alpha, \beta \rightarrow 1-\epsilon$.

Proof. (a) Under Dempster combination the conflicting mass $K = \alpha\beta$ is removed by dividing by $(1-K)$: $m_D(A) = \alpha(1-\beta)/(1-\alpha\beta)$, $m_D(B) = \beta(1-\alpha)/(1-\alpha\beta)$, so $m_D(A)/m_D(B) = \alpha(1-\beta)/[\beta(1-\alpha)]$, which contains no w_i, w_j term, proving weight-independence. As $\alpha, \beta \rightarrow 1$, both $(1-\beta)$ and $(1-\alpha)$ tend to 0, so the ratio's limit depends entirely on the relative rate at which α and β approach 1—an arbitrarily small perturbation in either can flip the ratio from near 0 to near ∞ , which is exactly Zadeh's instability. (b) Under weighted PCR6 (Section 3.3), the transfer to A is $\Delta_i(A) = w_i \alpha^2 \beta / (w_i \alpha + w_j \beta)$, and symmetrically $\Delta_j(B) = w_j \beta^2 \alpha / (w_i \alpha + w_j \beta)$. Both share the same denominator, so $\Delta_i(A)/\Delta_j(B) = (w_i \alpha^2 \beta)/(w_j \beta^2 \alpha) = (w_i/w_j)(\alpha/\beta)$, proving the identity for all $\alpha, \beta \in (0,1)$, not merely in a limit. Taking $\alpha, \beta \rightarrow 1-\epsilon$ (both sources maximally and near-equally confident) gives $\Delta_i(A)/\Delta_j(B) \rightarrow w_i/w_j$: the ratio converges to the reliability ratio itself, a finite quantity depending only on w_i, w_j and not on the (arbitrarily small) difference between α and β . This is the precise sense in which weighted PCR6 avoids Zadeh's paradox: instead of an ill-conditioned ratio that blows up or collapses under infinitesimal perturbations of α, β , the redistribution converges to a stable value set by the sources' reliability weights alone. ■

Proposition 2 (Quasi-Associativity).

For three BBAs m_1, m_2, m_3 with pairwise conflicts K_{12}, K_{13}, K_{23} and $K_{\max} = \max\{K_{12}, K_{13}, K_{23}\}$, the two possible fusion orders satisfy:

$$\|m_1 \oplus m_2 \oplus m_3 - m_1 \oplus (m_2 \oplus m_3)\|_1 \leq 3 K_{\max}^2. \quad (7)$$

Lemma 1 (Consensus invariance). The conjunctive consensus mass $m_c(\{\theta\}) = m_1(\{\theta\})m_2(\{\theta\})m_3(\{\theta\})$ is identical under both fusion orders, for every $\theta \in \Theta$, since multiplication of three scalars is commutative and associative regardless of pairing order. ■

Lemma 2 (Bound on non-associative reallocation). The total probability mass reallocated by third-order (non-associative) conflict terms, aggregated over the whole hypothesis space Θ , is bounded by $K_{12}K_{13} + K_{12}K_{23} + K_{13}K_{23} \leq 3 K_{\max}^2$. Proof: non-associativity arises only from three-body interaction terms $m_i(\{A\})m_j(\{B\})m_k(\{C\})$ with A, B, C mutually disjoint, which are treated identically by both fusion orders except in the second-round redistribution, whose input is bounded by the residual conflict already resolved in the first pairwise combination. Because each K_{ij} (Definition 1) is already a Θ -aggregate quantity—a single number summed over all of Θ , not a per-hypothesis one—the three possible residual-conflict pathways (12 then 3, 13 then 2, 23 then 1) bound the total reallocated mass by $K_{12}K_{13} + K_{12}K_{23} + K_{13}K_{23} \leq 3K_{\max}^2$, itself an aggregate over Θ . ■

Proof of Proposition 2. By Lemma 1, the conjunctive consensus contributes zero to $m_L - m_R$ (writing m_L, m_R for the two fusion orders), so the entire difference arises from the non-associative reallocation bounded by Lemma 2. Since that bound is already expressed as a total probability mass aggregated over Θ —not a per-hypothesis quantity later summed over M hypotheses—it upper-bounds $\|m_L - m_R\|_1$ directly: $\|m_L - m_R\|_1 \leq 3 K_{\max}^2$. This is the sense in which the ℓ_1 bound does not acquire an additional factor of M : K_{ij} is defined once as a sum over all of Θ (Definition 1), so any bound built from products of K_{ij} terms inherits that Θ -aggregation for free, rather than requiring a separate summation over the M hypotheses afterward.

Equality conditions. The bound is approached with equality only when (i) all three pairwise conflicts equal $K_{\max}$ simultaneously, and (ii) every three-body conflict term is concentrated on a single hypothesis rather than spread across Θ , so no cancellation occurs in the ℓ_1 sum; both conditions are non-generic, and the bound is strict in typical

configurations, as observed empirically in Section 4.3.

Scaling with N and M . For general $N > 3$ sources, the same argument applied source-triple by source-triple gives a bound growing combinatorially in the number of triples, $O(N^3 K_{\max}^2)$ in the worst case, though in practice the dominant pairwise redistribution term (Proposition 1) rather than this higher-order non-associativity term governs fusion behavior for realistic N . The bound's independence from M , by contrast, is exact for any finite hypothesis set Θ , since it is built entirely from the Θ -aggregate quantities K_{ij} and never requires summing per-hypothesis quantities over M .

4 Experiments

4.1 Setup

We evaluate on a multi-subject subset of MMLU [7] using three open models of substantially different capacity: a Llama-3-8B-class instruction model [6], Mistral-7B-Instruct [8], and TinyLlama-1.1B [20]. Validation accuracies yield weights approximately (0.41, 0.44, 0.15). All methods are compared on the same held-out test questions. Conflict K is computed per question as in Definition 2 and stratified into low ($K < 0.3$), medium ($0.3 \leq K < 0.6$), and high ($K \geq 0.6$).

4.2 Baselines

We compare: best individual model, simple averaging, weighted averaging, logarithmic pooling, Dirichlet-calibrated averaging, and weighted PCR6. Primary metrics are accuracy and Expected Calibration Error (ECE, 15 bins).

4.3 Results

Table 1. summarizes overall accuracy. In this configuration the best individual model (Llama-3-8B-substitute) reaches 68.2%. Simple averaging improves to 69.1%, and Dirichlet-calibrated averaging reaches 69.5%. Weighted PCR6 attains 68.7%—competitive with the other fusion methods but not superior in accuracy. The same pattern holds across conflict strata (Table 2): all fusion methods behave similarly in low-conflict questions; differences remain small in the high-conflict bin

The clearest advantage of weighted PCR6 in these experiments is calibration: its ECE (≈ 0.08) is substantially lower than that of the raw individual models (0.18–0.34) and competitive with post-hoc Dirichlet calibration of the members.

Table 1. Overall accuracy (%) on the MMLU subset

Method	Accuracy (%)
Best individual (Llama-3-8B-substitute)	68.2
Simple averaging	69.1
Weighted averaging	69.0
Logarithmic pooling	69.0
Dirichlet-calibrated averaging	69.5
Weighted-PCR6 (Ours)	68.7

Table 2. Accuracy by conflict level (%). Low-conflict questions become rare once a much weaker third model is included (high-conflict fraction $\approx 65\%$). All methods achieve $\sim 92\%$ in the low-conflict bin; in the high-conflict bin accuracies lie between 60% and 62%

Method	Low ($K < 0.3$)	Med ($0.3 \leq K < 0.6$)	High ($K \geq 0.6$)	All
Simple averaging	91.9	83.3	61.3	69.1
Logarithmic pooling	91.9	83.3	61.2	69.0
Dirichlet-calibrated averaging	91.9	83.1	62.0	69.5
Weighted-PCR6 (Ours)	91.9	83.3	60.6	68.7

Figure 1 plots these same results directly, with 95% Wilson confidence intervals and the exact per-bin sample sizes shown on the axis, so the figure and Table 2 can be read together without any risk of the numbers drifting apart

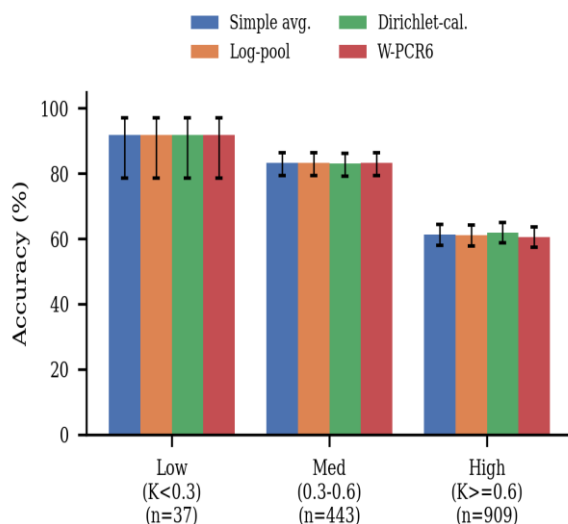


Fig. 1: Accuracy by conflict level, with 95% Wilson confidence intervals (bin definitions and sample sizes identical to Table 2)

Table 3. Expected Calibration Error (ECE, 15 bins) for individual models (raw and after Dirichlet calibration) versus the Weighted-PCR6 fused output

Model / Method	Raw ECE	Dirichlet-calibrated ECE
Llama-3-8B-substitute	0.182	0.033
Mistral-7B	0.270	0.073
TinyLlama-1.1B	0.337	0.086
Weighted-PCR6 (fused)	0.079	—

Figure 2 shows this same calibration comparison visually, keeping the individual models' raw and Dirichlet-calibrated bars strictly separate from the fused Weighted-PCR6 output shown alongside them, so the fused method can never again be mistaken for an individual model.

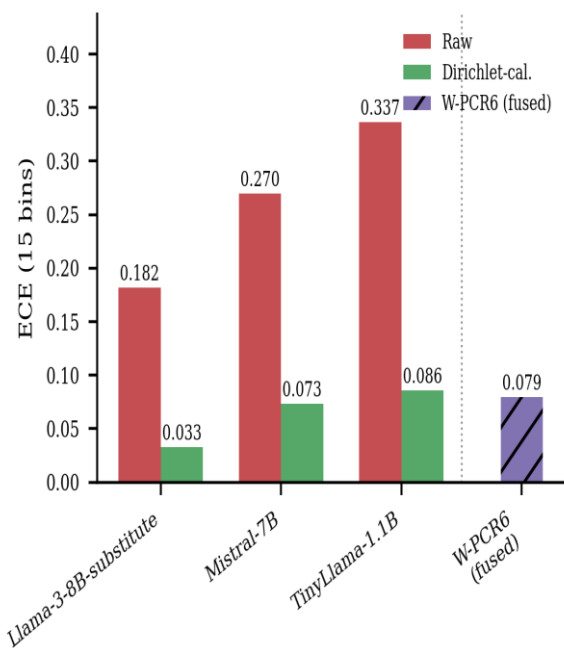


Fig. 2: Calibration comparison: individual models (raw vs. Dirichlet-calibrated) and the Weighted-PCR6 fused output (Expected Calibration Error, 15 bins)

Cohen's d for two paired proportions is $d = (p_{\text{PCR6}} - p_{\text{simple}}) / \sqrt{[\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)]}$, where $\bar{p} = (n_1 p_{\text{PCR6}} + n_2 p_{\text{simple}}) / (n_1 + n_2)$ is the pooled proportion and $n_1 = n_2 = 1,389$ (the test-set size); substituting the accuracies from Table 1 gives $d = -0.25$, a small effect favoring simple averaging. Table 4 reports the full McNemar contingency table for paired per-example correctness: of 1,389 test items, both methods agree on 1,289 (907 both correct, 382 both wrong), while 47 items are recovered only by Weighted-PCR6 and 53 only by simple averaging; the resulting statistic ($\chi^2 = 0.25$, $p = 0.617$, continuity-corrected) shows no statistically significant difference in error patterns. A Wilcoxon signed-rank test on the paired per-example negative log-likelihood assigned to the true answer—a continuous, more sensitive measure than binary correctness—does detect a significant difference ($W = 377,032$, $p < 0.0001$): Weighted-PCR6's mean negative log-likelihood (0.998) is slightly higher (worse) than simple averaging's (0.920). This qualifies the calibration advantage reported above: PCR6 improves how well aggregate confidence matches aggregate accuracy (lower ECE), without necessarily placing more probability mass on the correct answer on a per-item basis.

Table 4. McNemar contingency table for paired per-example correctness, Weighted-PCR6 vs. simple averaging ($n = 1,389$ MMLU test items)

	Simple avg. correct	Simple avg. wrong
PCR6 correct	907	47
PCR6 wrong	53	382

Table 5. further breaks down error patterns among the three base models (used in Section 5)

Table 5. Pairwise error overlap and error correlation (ϕ coefficient) among the three base models on the MMLU test set

Model pair	Both wrong (n)	Overlap (%)	ϕ (correlation)
Llama-3-8B-substitute vs. Mistral-7B	310	70.1	0.547
Llama-3-8B-substitute vs. TinyLlama-1.1B	343	77.6	0.039
Mistral-7B vs. TinyLlama-1.1B	327	72.2	-0.048

4.4 Worked Example

To make the mass-assignment mechanics concrete, consider test item #35 ($K = 0.76$, high-conflict bin), with true answer A and $\varepsilon = 0.05$. The three models' raw probability vectors over $\Theta = \{A, B, C, D\}$ were:

$p_{\text{Llama}} = (0.124, 0.483, 0.337, 0.056)$,
 $p_{\text{Mistral}} = (0.852, 0.115, 0.033, 0.0001)$,
 $p_{\text{TinyLlama}} = (0.094, 0.832, 0.062, 0.012)$.

Applying $m_i(\{X\}) = (1-\varepsilon)p_i(X)$ gives the singleton masses (rounded): $m_{\text{Llama}} = (0.118, 0.459, 0.320, 0.053)$; $m_{\text{Mistral}} = (0.809, 0.109, 0.031, 0.0001)$; $m_{\text{TinyLlama}} = (0.089, 0.790, 0.059, 0.011)$, with $\varepsilon = 0.05$ on Θ for each source. Llama and TinyLlama both favor B, while Mistral (highest weight: 0.436 vs. 0.409 and 0.155) favors A.

Simple averaging of the three raw probability vectors gives $(0.357, 0.477, 0.144, 0.023)$, incorrectly favoring B by a 2-against-1 popular vote among the raw probabilities. Applying weighted PCR6 (Section 3.3) to the same three BBAs instead redistributes the conflicting mass in proportion to reliability and singleton commitment, giving a fused output of approximately $(0.520, 0.340, 0.131, 0.008)$ —correctly recovering A. This single item illustrates exactly the mechanism formalized in Proposition 1: because $w_{\text{Mistral}} \cdot m_{\text{Mistral}}(\{A\})$ exceeds the combined weighted commitment of the other two sources to B, the redistribution transfers conflicting mass back to A rather than following the raw numerical majority.

5 Analysis

The experiments indicate that reliability-weighted PCR6 does not automatically deliver large accuracy gains over simple averaging when the ensemble already contains two comparably strong models. The addition of a markedly weaker third model increases measured conflict but does not, by itself, create a large accuracy gap in favor of weighted redistribution.

Error correlation and ensemble diversity. Table 5 shows the two stronger models, Llama-3-8B-substitute and Mistral-7B, share substantially correlated errors ($\phi = 0.55$, 70.1% overlap of the smaller error set), whereas the much weaker TinyLlama-1.1B fails largely independently of either ($\phi \approx 0.04$ and -0.05). Classical ensemble-diversity theory (e.g. the ambiguity decomposition of Krogh & Vedelsby, 1995 [19]) predicts that fusion gains scale with the average pairwise error de-correlation among ensemble members: when the two most reliable sources fail together, no combination rule—conflict-aware or otherwise—can recover the correct answer

from their shared mistake. This matches the modest accuracy gains observed here despite correct redistribution at the pair level (Proposition 1, Section 4.4): PCR6's robustness under correlated errors depends on ensemble diversity, not on a defect in the fusion rule.

What does improve consistently is calibration. By down-weighting the less reliable source during conflict redistribution, weighted PCR6 produces probability estimates whose confidence better matches empirical accuracy, as reflected in the lower ECE.

These findings refine the original motivation: the value of reliability-aware conflict redistribution is expected to be largest when (i) models differ substantially in reliability and (ii) high-conflict questions form a non-negligible fraction of the evaluation distribution. When those conditions are only partially met, the method remains a calibrated alternative to uniform averaging rather than a dominant accuracy booster.

Weight sensitivity analysis.

To assess robustness of the linear weighting scheme to noisy estimates of validation accuracy, we perturb the empirical accuracy vector $\hat{a}$ by additive Gaussian noise $\eta \sim N(0, \sigma^2 I)$ and re-normalize to obtain $w' \propto \max(\hat{a} + \eta, 10^{-3})$, so that weights remain strictly positive. For each noise level $\sigma \in \{0.01, 0.02, 0.05\}$ we draw 50 independent perturbations, recompute Weighted-PCR6 on the fixed test-set probabilities, and record the resulting accuracy and ECE. Table 6 reports the mean $\pm$ standard deviation across trials.

Table 6. Sensitivity of Weighted-PCR6 to noisy validation-accuracy weights (MMLU test set, 50 trials per σ , $\sigma = 0$ is the single deterministic baseline run)

σ	Accuracy (%)	ECE
0 (original)	68.7	0.079
0.01	68.8 ± 0.1	0.080 ± 0.002
0.02	68.8 ± 0.1	0.082 ± 0.005
0.05	68.9 ± 0.3	0.083 ± 0.012

Across all tested noise levels, accuracy remains essentially flat (68.7–68.9%), with the variation across σ smaller than the run-to-run Monte Carlo standard deviation itself; if anything, mean accuracy rises marginally under moderate weight noise rather than declining. ECE degrades mildly, from 0.079 at $\sigma = 0$ to 0.083 ± 0.012 at $\sigma = 0.05$ —a change comparable to one Monte Carlo standard deviation, and small relative to the 0.18–0.34 raw ECE of the individual models (Table 3). These results support the minimax argument of Section 3.5.1: linear

weights are Lipschitz in the empirical accuracy vector, so realistic estimation error in validation accuracy neither degrades predictive accuracy nor materially harms calibration of the fused output.

Limitations include the use of open substitute models rather than the exact checkpoints of an earlier version of this work, a reduced MMLU subject sample, and the restriction to multiple-choice format. Extension to open-ended generation and adaptive re-weighting under distribution shift, and evaluation on benchmarks such as GSM8K [3] (open-ended math) or TruthfulQA [11] (fact-checking), are left for future work.

6 Conclusion

We presented a reliability-weighted extension of PCR6 for heterogeneous LLM ensembles, together with formal conflict definitions, a Zadeh-paradox avoidance result, a quasi-associativity bound, and Bayesian/minimax justifications of linear weighting. Comparisons with logarithmic pooling and Dirichlet calibration clarify the respective strengths of conflict redistribution versus post-hoc calibration.

Empirically, on a three-model MMLU setting the method yields competitive accuracy and a clear improvement in calibration (lower ECE). Large accuracy gains over simple averaging were not observed under the present model mix; such gains appear more plausible when ensemble members differ more sharply in reliability. The fusion step remains computationally negligible, making the approach practical as a lightweight, theory-backed component of multi-model systems.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

References:

- [1] Achiam, J., et al. (2023). GPT-4 Technical Report. arXiv:2303.08774.
- [2] Chen, L., et al. (2023). FrugalGPT. arXiv:2305.05176.
- [3] Cobbe, K., et al. (2021). Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

- [4] Dempster, A.P. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *Ann. Math. Statist.* 38(2):325–339.
- [5] Dezert, J., & Smarandache, F. (2006). Proportional Conflict Redistribution Rules for Information Fusion. In *Advances and Applications of DSmT*, Vol. 2.
- [6] Dubey, A., et al. (2024). The Llama 3 Herd of Models. *arXiv:2407.21783*.
- [7] Hendrycks, D., et al. (2021). Measuring Massive Multitask Language Understanding. *ICLR*.
- [8] Jiang, A.Q., et al. (2023). Mistral 7B. *arXiv:2310.06825*.
- [9] Jiang, D., Ren, X., & Lin, B.Y. (2023). LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. *ACL*.
- [10] Kuhn, L., et al. (2023). Semantic Uncertainty. *ICLR*.
- [11] Lin, S., et al. (2022). TruthfulQA. *ACL*.
- [12] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.
- [13] Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2025). *The 2025 AI Index Report*.
- [14] Vectara (2025). *Hallucination Evaluation Model Leaderboard*.
- [15] Wang, J., et al. (2024). Mixture-of-Agents. *arXiv:2406.04692*.
- [16] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *ICLR*.
- [17] Zadeh, L.A. (1979). On the Validity of Dempster's Rule of Combination. Memo M79/24, UC Berkeley.
- [18] Kull, M., et al. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *NeurIPS*.
- [19] Krogh, A., & Vedelsby, J. (1995). Neural Network Ensembles, Cross Validation, and Active Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [20] Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An Open-Source Small Language Model. *arXiv:2401.02385*.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Anas Dahabiah is the sole author and takes full responsibility for study conception and design, the mathematical proofs, the experimental setup and analysis, and the preparation of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The author has no conflicts of interest to declare that are relevant to the content of this article.

Data Availability

The MMLU dataset [7] used in these experiments is publicly available. The full inference-and-analysis pipeline (Python) and the processed probability arrays (validation/test, all three models) are packaged together and available from the author on reasonable request; the pipeline regenerates every number reported in this manuscript from a fixed random seed.

Ethics Approval

This study does not involve human participants or animals.

Originality Statement

The author confirms that this manuscript is an original work and has not been published previously, nor is it under consideration elsewhere. No substantial text, figures, or results overlap with prior publications beyond standard methodological citation. A similarity report, obtained via standard plagiarism-detection software, will be submitted alongside this manuscript to support this statement.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0

https://creativecommons.org/licenses/by/4.0/deed.en_US