SciELO Preprints

Publication status: This preprint has not been published elsewhere.

Provenance-Aware PCR6 for Conflicting Evidence in Retrieval-Augmented Generation

Florentin Smarandache, Maikel Yelandi Leyva Vázquez, Jean Dezert

https://doi.org/10.1590/SciELOPreprints.18188

Submitted on: 2026-09-26
Posted on: 2026-10-29 (version 1)
(YYYY-MM-DD)

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

Provenance-Aware PCR6 for Conflicting Evidence in Retrieval-Augmented Generation

When Does PCR6 Help Retrieval-Augmented Generation? A Provenance-Aware DSmT Study of Conflicting Evidence, with a Worked Illustration of PCR6+

Florentin Smarandache[1], Maikel Leyva Vázquez[2,3] and Jean Dezert[4]

[1] Mathematics, Physics and Natural Science Division, University of New Mexico, Gallup, New Mexico, USA
[2] Centro de Investigación Institucional, Universidad Bernardo O'Higgins, Santiago, Chile
[3] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Guayaquil, Ecuador
[4] Independent Researcher, ISIF Senior Member, Orléans, France

ORCID: Florentin Smarandache https://orcid.org/0000-0002-5560-5926; Maikel Leyva Vázquez https://orcid.org/0000-0001-7911-5879; Jean Dezert https://orcid.org/0000-0003-3474-9186

Abstract

Retrieval-augmented generation (RAG) must often combine passages that are ambiguous, redundant, or mutually contradictory. This study asks whether proportional conflict redistribution from the Dezert-Smarandache theory (DSmT) provides a measurable advantage over Dempster-Shafer (DS) combination when both rules receive the same evidence and the same control for source dependence. Eight fusion methods were evaluated on 500 RAMDocs (Retrieval with Ambiguity and Misinformation in Documents) questions in three conditions. Under duplicated misinformation, provenance-aware PCR6 (Proportional Conflict Redistribution rule #6) reached 80.6% exact-set accuracy and 92.31% set F1, against 78.2% and 90.98% for provenance-matched DS; the exact-set difference was 2.4 percentage points (95% bootstrap interval [1.2, 3.8]) and the F1 difference was 1.33 percentage points (95% bootstrap interval [0.60, 2.33]). With NLI (natural language inference)-derived masses, PCR6 and DS selected identical answer sets, but PCR6 reduced Brier error. A preregistered 1,000-item multi-source experiment using singletons, proper unions, and total ignorance was followed by external validation on 428 RAMDocs items. In controlled partial-ignorance conflict, PCR6+ reduced BetP (betting probability transformation) Brier error from 0.3185 for PCR6 to 0.2733 and remained exactly neutral to an added vacuous BBA (basic belief assignment), although the DS rule achieved the lowest BetP Brier error of the three rules in that same controlled evaluation (0.1878). On the prespecified 260-item RAMDocs applicability subset, ordinary PCR6 had lower BetP Brier error than PCR6+ (0.2354 versus 0.2629), while both achieved set F1 = 0.7718. Exploratory decision sensitivity showed that maximum belief gave PCR6+ its best hard-decision result (exact set 0.5077; set F1 0.7859), whereas DSmP (Dezert-Smarandache probability transformation) sharpened rankings but worsened the RAMDocs Brier score. The evidence therefore supports a bounded claim: PCR6+ corrects a structural multi-source weakness of PCR6 under broad or vacuous focal evidence, but the best fusion and decision rules depend on the downstream loss and the mapping from text to BBAs.

Keywords: Dezert-Smarandache theory; PCR6; PCR6+; Dempster-Shafer theory; retrieval-augmented generation; information fusion; provenance; conflicting evidence; natural language inference; calibration

1 Introduction

Retrieval-augmented generation (RAG) pairs a user query with passages retrieved from external sources. Retrieval improves factual access, but it does not guarantee agreement among the passages it returns: one query may refer to several entities, correct passages may coexist with misinformation, and a set of apparently independent passages may be copies of a single source. RAMDocs (Retrieval with Ambiguity

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

and Misinformation in Documents) was built to combine these difficulties instead of studying them in isolation, and its authors report a sharp drop in performance once support for misinformation becomes imbalanced [1]. A fusion stage therefore has to decide which claims deserve support, how repeated evidence should be counted, and how the remaining uncertainty should be reported.

Most RAG systems handle conflict through ranking, voting, prompting, or learned aggregation. These procedures work, but they compress disagreement into one score. Evidence theory offers a different representation. A basic belief assignment (BBA) can commit mass to a specific proposition while leaving residual mass on ignorance. Dempster-Shafer (DS) theory combines BBAs on an exclusive frame and normalizes their conflict [2], whereas the Dezert-Smarandache Theory (DSmT, https://fs.unm.edu/DSmT/) extends belief reasoning to non-exclusive frames and defines proportional conflict redistribution (PCR) rules that return each partial conflict to the propositions that generated it [3–7]. PCR6 generalizes that principle to any number of sources and is the rule studied here; a recent theoretical refinement, PCR6+, further restricts the redistribution to focal elements that are genuinely implicated in a given conflict, which is the improvement examined in Section 5.5 [7a].

A concrete frame of discernment is essential for interpreting the fusion equations. Section 3.2 defines the frame used in this study and introduces a worked numerical example before the general methods.

The distinction matters for generative artificial intelligence. Retrieved passages, language-model agents, and external tools may support overlapping or incompatible claims, while duplicated provenance can manufacture a false consensus. DSmT has already proved useful in sensor fusion, target recognition, tracking, and decision support [7,8,17,18]; its value for conflict-aware RAG, however, has not been tested under a matched protocol. Any informative test must compare PCR6 with a DS baseline that shares the frame, the mass assignment, the entity grouping, and the provenance treatment. Otherwise an improvement produced by better preprocessing or by duplicate control can be misattributed to the combination rule.

This paper reports a matched comparison and, to our knowledge, the first controlled characterization of when PCR6 helps RAG. It defines the frame of discernment and the number of fused sources explicitly, analyzes the configurations in which DS and PCR6 are expected to diverge, and evaluates PCR6+ with three or more sources and partial-ignorance BBAs. Three questions guide the study. Does provenance-aware PCR6 improve multi-answer recovery when misinformation is duplicated? Does any difference survive when document support is estimated automatically by natural language inference (NLI) instead of a heuristic score? Under which evidence configurations and decision losses do DS, PCR6, and PCR6+ differ in practice? The study also reports conditions in which no detectable difference is observed, because the practical value of a combination rule depends on knowing where it pays off. The contributions are as follows.

- A matched comparison of PCR6 and Dempster-Shafer combination on RAMDocs in which the frame, the mass assignment, the entity grouping, and the provenance treatment are identical.
- A provenance discounting transformation that makes g copies of one text count as a single source, applied in the same way to both rules.
- An explicit accounting of how many BBAs are fused for each entity group and experimental condition, together with a worked numerical example of the frame of discernment (Section 3.2).
- A theoretical analysis of when DS and PCR6 are expected to agree or diverge, explaining why the empirical PCR6 advantage is concentrated in the duplicated-misinformation condition (Section 5.1).
- An empirical localization of the PCR6 advantage: answer-set recovery under duplicated conflicting evidence, calibration under NLI-derived masses, and no measurable difference in the remaining conditions.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

- A worked, exactly-computed three-source toy example contrasting DS, PCR6, and PCR6+, followed by a preregistered multi-source partial-ignorance experiment and RAMDocs external validation (Section 5.5).
- A negative control on ConflictQA that separates dependence modeling from the choice of combination rule.

2 Related work

2.1 Dempster-Shafer theory and DSmT

Dempster defined lower and upper probabilities induced by multivalued mappings, and Shafer developed the theory of evidence built on them [2,9]. On a frame Θ, a BBA m distributes unit mass over the subsets of Θ, and the mass committed to Θ measures ignorance. DS combination first computes conjunctive agreement and then divides every nonempty mass by $1 - K$, where K is the total conflict, i.e. the mass that the conjunctive fusion of the sources of evidence commits to the empty set. The normalization is coherent under its assumptions, but strongly conflicting evidence can concentrate belief on a proposition that no source supports in isolation, as Zadeh showed [10]. Alternative rules send conflict to ignorance, to unions, or to the open-world empty set [11–13].

A broader body of theoretical analyses, counterexamples, and comparative applications has documented further departures from Bayesian updating, insensitivity in some conjunctive configurations, and problematic behavior under particular high-conflict conditions [21–29]. These results motivate targeted stress tests of combination rules; they do not imply that Dempster's rule fails under all modeling assumptions or operating regimes.
DSmT relaxes the mutual-exclusivity assumption through free and hybrid models and works on the hyper-power set when intersections must remain explicit [5–7]. The PCR (Proportional Conflict Redistribution) family separates the conjunctive step from conflict redistribution. PCR5 (PCR rule #5) returns each two-source partial conflict only to the focal elements involved in it; Martin and Osswald introduced PCR6 (PCR rule #6) as its generalized multi-source counterpart [7d]. PCR6 follows the same local principle for multiple sources [4,6,7d]. The rule therefore keeps information about the origin of disagreement that global normalization discards. That property is precisely what a RAG fusion layer needs when one document contradicts another while a third remains neutral.

The experiment reported here isolates the conflict rule rather than the full expressive capacity of DSmT. The answer frame is exclusive and the focal sets are candidate singletons and total ignorance, a Shafer model in which PCR6 remains applicable [4]. Testing whether the hybrid DSm frame adds a further benefit requires semantically overlapping answer claims, which we return to in Section 6.

2.2 A theoretical refinement: PCR5+ and PCR6+

Dezert et al. identified a limitation of PCR5 and PCR6 that becomes relevant when three or more sources are combined: a partial conflicting product can redistribute mass to a focal element — typically a broad disjunction such as total ignorance Θ — that is not itself in conflict with the disputed propositions, merely because it appears as one factor of the conflicting product [7a]. A direct symptom is that the vacuous BBA can cease to be neutral once three or more BBAs are combined with PCR5 or PCR6. The authors introduce a binary keeping index that retains the inclusion-minimal focal elements genuinely implicated in the conflict and discards a broader element when it properly contains the more specific participants. PCR5+ and PCR6+ coincide with PCR5 and PCR6 for two sources, restore vacuous-BBA neutrality, and avoid inflating broad uncertainty. Section 5.5 presents a worked three-source illustration and relates it to the multi-source evaluation.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

2.3 Conflicting evidence in RAG

ConflictQA evaluates the interaction between the parametric memory of a language model and retrieved counterevidence [14]. RAMDocs extends the problem to ambiguity, misinformation, and noise, and introduces MADAM-RAG (Multi-Agent Debate for Ambiguity and Misinformation in RAG), in which several agents debate the candidate answers [1]. Both benchmarks show that conflict depends on relevance, on the number of documents, and on the form in which those documents are presented. They also leave open the question that concerns us: how a fusion layer should keep support, conflict, ignorance, and dependence apart.

Dempster–Shafer combination has also been applied to text-derived evidence outside RAG. Saha et al. [14a] construct a stance tree from rhetorical parsing, predict stance at the elementary-discourse-unit or sentence level, and aggregate those predictions with DS theory to produce explainable document-level stance decisions. Their objective is discourse-based explanation within a document rather than the fusion of entailment-derived BBAs across conflicting retrieved sources, leaving the latter problem open.
Natural language inference (NLI) provides a practical bridge from text to belief assignments. An NLI model estimates entailment, neutrality, and contradiction between a document and a candidate claim. Such scores are not calibrated belief masses, but they support a transparent mapping whose sensitivity can be examined. DistilBERT [15] fine-tuned on MultiNLI [16] is used here to test whether the behaviour of PCR6 persists when evidence strength is derived from document semantics rather than from a fixed lexical reliability score. Learned evidential models estimate belief masses directly [19], but they need training targets that RAG conflicts do not supply in a comparable form.

3 Methods

3.1 Study design and data

The primary benchmark used the 500 items in the official RAMDocs test file. Each item contains a question, disambiguated entities, a gold answer set, and documents labeled for evaluation as correct, misinformation, or noise. The SHA-256 (Secure Hash Algorithm, 256-bit) digest of the downloaded test file was C67F699C97349F00CF1BD08D1DBF8CA1D0CC38C306715C93A10A4F961DCF28B7. The repository was retrieved on 14 September 2026 under the MIT license.

Three variants were evaluated. Original retained the published documents. Duplicate misinformation added three exact copies of every misinformation document, producing four appearances of that text in total. Misinformation dominant retained one correct document for each distinct correct answer and added two further copies of every misinformation document. The document type label was used only to build these declared stress conditions and to score predictions; it was never shown to a fusion method.

A secondary negative-control experiment used 7,947 records from the ConflictQA-popQA-chatgpt configuration of ConflictQA [14] (https://huggingface.co/datasets/osunlp/ConflictQA), which pairs PopQA questions [14b] with ChatGPT-generated parametric-memory answers as counter-evidence; the dataset authors do not publish the exact ChatGPT snapshot used. It compared external evidence with parametric-memory answers under balanced, repeated, and independent-conflict constructions. Because this binary task did not separate the two rules, we report it as a scope check rather than as a primary claim.

3.2 The frame of discernment

For a RAMDocs question, the frame of discernment Θ is the set of normalized candidate answer strings that appear anywhere in the documents attached to that question's entity group, e.g. $\Theta = \{a_1, a_2, \ldots, a_k\}$ for a question with k distinct candidates. This is an exclusive Shafer frame: the k candidates are treated as mutually exclusive singletons, and the only other focal element used anywhere in the study is Θ itself, which represents “the true answer is one of these candidates, but this document does not say which.” No

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

BBA in this study commits mass to a proper union of two or more named candidates (e.g. $\{a_1, a_2\}$); such partial-ignorance focal elements are outside the frame used for the RAMDocs and NLI experiments and are introduced only in the illustrative Section 5.5.

A concrete instance: consider a RAMDocs-style question "Who directed the film X?" whose documents mention two distinct names, so $\Theta = \{a_1 =$ "Director A", $a_2 =$ "Director B"$\}$. Suppose three documents are retrieved for this entity. Document 1 names Director A with high lexical and entity overlap; its BBA is $m_1(a_1) = 0.70$, $m_1(\Theta) = 0.30$. Document 2 is a misinformation page naming Director B; its BBA is $m_2(a_2) = 0.60$, $m_2(\Theta) = 0.40$. Document 3 is a weaker, partially relevant page that also names Director A; its BBA is $m_3(a_1) = 0.50$, $m_3(\Theta) = 0.50$. Section 5.5 fuses exactly this three-source example with DS, PCR6, and PCR6+ and reports the resulting numbers in full, so the reader can see, with actual arithmetic, what each rule does to this Θ before returning to the aggregate benchmark results in Section 4.

3.3 Candidate answers and entity grouping

Because the study targets the fusion layer, the document-level answer field supplied the output of a perfect claim extractor. Answer strings were normalized to lowercase alphanumeric tokens and deduplicated. This is an oracle assumption: free-form answer extraction is not measured here. The required number of outputs k was the number of disambiguated entities supplied by RAMDocs.

Documents were assigned to an entity by lexical overlap, exact entity-phrase presence, and parenthetical qualifiers. Fusion then ran separately within each entity group — the procedure this paper labels a "factorized" fusion method, as opposed to a "global" method that pools all of a query's documents into a single fusion regardless of entity. The highest-probability answer from each group was selected, duplicate answers were removed, and any missing positions were filled from the global fused distribution. All factorized methods received the same entity assignments.

3.4 Number of sources fused per query

Fusion is not restricted to two sources. In the oracle RAMDocs experiment, S equals the number of document BBAs assigned to an entity group. Exact copies remain as separate BBAs, but each copy receives the discounted confidence r* of Eq. (2), so their combined residual ignorance equals that of one original source. Consequently, adding copies can increase the enumerated S without giving them the evidential strength of independent confirmations. Across the 500 RAMDocs items there were 2,766 document BBAs before entity grouping, with 1–12 documents per item (median 5; mean 5.53); the entity-specific S is a subset of that item total. In the NLI experiments, by contrast, exact text duplicates are collapsed to one provenance source before fusion. The PCR6+ RAMDocs extension retains at most the six most committed provenance-distinct BBAs. The DS rule is applied sequentially because it is associative, whereas PCR6 and PCR6+ use simultaneous S-source enumeration because they are not associative.

The NLI model evaluated 7,802 document-candidate pairs, but a pair is an NLI scoring operation rather than a separate fusion source. After duplicate collapse, each retained document contributes one BBA to its entity group. The number of fused BBAs is therefore the number of retained provenance-distinct documents, never an assumption that S = 2.

3.5 Basic belief assignments

In the oracle fusion experiment, each document i supported one normalized candidate answer a_i with confidence r_i and assigned the remaining mass to total ignorance Θ.

$$m_i(\{a_i\}) = r_i, m_i(\Theta) = 1 - r_i \quad (1)$$

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

The two-focal-element BBA serves as a deliberately simple baseline because it permits DS and PCR6 to be compared on identical ground with a minimal, auditable mass model. The extension in Section 3.9 adds proper candidate unions and Θ while preserving the baseline analyses in Tables 1 and 2.

Confidence was computed deterministically from question-token overlap, entity-token overlap, and exact entity-phrase presence. The unclipped score was 0.50 plus 0.16 times question overlap, plus 0.18 times entity overlap, plus 0.08 times phrase presence; values were clipped to the interval from 0.50 to 0.90. This mapping never uses the correct, misinformation, or noise label.

3.6 Provenance discounting

Exact normalized document text defined a provenance identifier. When g documents shared the same identifier, their individual confidence was replaced by r* so that combining g copies preserved the residual ignorance of one original source.

$$r^* = 1 - (1 - r)^{\frac{1}{g}} \tag{2}$$

The transformation prevents exact copies from being counted as independent confirmation, and it was applied identically to the provenance-aware DS and PCR6 methods. The matched design is essential, because the largest improvement over methods without provenance control comes from dependence modeling itself. Provenance discounting also fixes the realized value of S discussed in Section 3.4: g exact copies of one document contribute a single discounted source rather than g independent sources.

3.7 DS rule and PCR6 fusion for S sources

For focal sets X_1, …, X_s from S sources, the conjunctive mass was computed by summing products whose intersections equal A.

$$m_{\cap}(A) = \sum_{X_1 \cap \cdots \cap X_S = A} \prod_{i=1}^{S} m_i(X_i) \tag{3}$$

The DS rule removed the empty-set conflict K and normalized each nonempty mass by 1 − K. The rule is mathematically undefined at complete conflict (K = 1); any software return of total ignorance in that case is an explicitly labeled engineering fallback, not Dempster's rule. When S documents support one entity and K < 1, the DS rule is applied sequentially over all S sources because it is associative.

$$m_{\mathrm{DS}}(A) = \frac{m_{\cap}(A)}{1-K}, A \neq \emptyset \tag{4}$$

The multi-source PCR6 implementation enumerated every S-tuple of focal elements. A compatible tuple contributed the product of its source masses to the intersection. For a conflicting tuple (X_1,…,X_s) with product P, PCR6 redistributed P only to the focal elements of that tuple, in proportion to the sum of the source masses that equal each distinct focal element [4,6].

$$\Delta m_{\mathrm{PCR6}}(X|X_1, \ldots, X_S) = P \frac{\sum_{i:X_i = X} m_i(X)}{\sum_{j=1}^{S} m_j(X_j)} \tag{5}$$

The resulting mass was normalized for numerical precision and converted into answer probabilities by the pignistic transformation (the betting probability, BetP); the highest-probability answer in each entity group was then selected.

BetP is the primary decision transform because it is parameter-free and yields probabilities suitable for proper scoring. The exploratory sensitivity analysis also evaluates DSmP with ε = 0.001, maximum singleton belief, and maximum singleton plausibility on the same fused BBAs. DSmP redistributes each

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

compound mass in proportion to the singleton masses already present plus ε [20a]. Maximum belief and maximum plausibility are ranking rules rather than calibrated probability forecasts.

Beyond these transforms, the belief-interval-distance method exploits the full interval [Bel(A), Pl(A)] of each focal element. It selects the decision $\hat{X}$ that minimizes dBI(m,mX), the distance between the fused BBA m and the categorical BBA mX focused on X. Han, Dezert, and Yang define Euclidean- and Chebyshev-family belief-interval distances and prove that they are strict distance metrics [7b]; Dezert, Han, Tacnet, Carladous, and Yang develop the corresponding minimum-distance decision scheme and its quality indicator $q(\hat{X})$ [7c]. The present post-confirmatory implementation uses the Euclidean-family distance, restricts admissible X to singletons to match the RAMDocs answer task, and enumerates every nonempty subset of the frame. Minimum belief-interval distance, maximum belief, and maximum plausibility are ranking rules, so Brier scores are not reported for them. Results appear in Section 4.4.

3.8 NLI-derived masses

The NLI experiment processed 7,802 document-candidate pairs with the quantized Xenova/distilbert-base-uncased-mnli model, using 8-bit quantization, Transformers.js 4.2.0, ONNX Runtime, and a maximum sequence length of 256 tokens. For each document-candidate pair, one NLI forward pass returns the probability triplet (p_E,p_N,p_C), where p_E, p_N, and p_C denote entailment, neutral, and contradiction, respectively, and $p_E + p_N + p_C = 1$. Candidate support was $p_E(1 - p_C)$, and the candidate with the largest support was selected as that document's supported answer a_i.

The committed mass r_{NLI} assigned to a_i is a single scalar, not a vector. It equals p_E discounted multiplicatively by one minus half the neutral probability and by one minus half the contradiction probability, and is then clipped to [0.05,0.95] as shown in Eq. (6).

$$r_{\mathrm{NLI}} = \mathrm{clip}[p_E(1 - 0.5p_N)(1 - 0.5p_C),0.05,0.95] \tag{6}$$

Here clip(x,lo,hi) = min(max(x,lo),hi) is the standard clipping or saturation function: it truncates x to the closed interval [lo,hi] and leaves it unchanged when $lo \leq x \leq hi$. The bounds prevent this heuristic mapping from producing a completely uncommitted source at $r_{NLI} = 0$ or a fully dogmatic singleton source at $r_{NLI} = 1$. Thus r_{NLI} becomes the document's committed mass $m_i(\{a_i\})$, while $m_i(\Theta) = 1 - r_{NLI}$, exactly as in Eq. (1). All 7,802 probability triplets were valid and summed to one within floating-point precision. One RAMDocs item had no usable document candidate, leaving 499 items with NLI pairs, although evaluation retained 500 rows by assigning an empty prediction to that case.

Exact duplicate documents were collapsed into one provenance source before factorized DS or PCR6 fusion, exactly as in Section 3.6. Both methods received identical NLI records, candidates, entity assignments, and provenance groups, so the comparison isolates the fusion rule and not the quality of the mass mapping.

3.9 PCR6+ confirmatory extension

Before inspecting PCR6+ outcomes, we froze the generator, endpoints, seed, and RAMDocs applicability criteria. The controlled benchmark generated 1,000 four-candidate decisions. Each decision fused five BBAs in the clean and partial-conflict conditions and six after vacuous or near-vacuous source injection. The BBAs included singleton focal elements, proper candidate unions, and Θ. PCR6+ used the published binary keeping index and simultaneous multi-source redistribution [7a].

For external validation, the same fixed NLI outputs were mapped to generalized BBAs. Let $s_j = p_E(a_j)[1 - p_C(a_j)]$ and let a1 and a2 be the two candidates with greatest support. Equation (6) supplies commitment c, while $d = \mathrm{clip}[(s_1 - s_2)/\max(s_1,10^{-12}),0,1]$ measures separation. Each document assigns $m(\{a_1\}) = cd$, $m(\{a_1,a_2\}) = c(1 - d)$, and $m(\Theta) = 1 - c$. Thus a decisive document supports one answer, an ambiguous document can support the union of its two leading candidates, and unresolved evidence remains on Θ. Exact duplicates were collapsed and at most six provenance-distinct sources were fused.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

The prespecified applicability subset required at least three retained BBAs, at least three candidates, and at least one proper non-singleton focal element.
The primary endpoint was multiclass Brier score of BetP (i.e. the mean squared error of the BetP probability forecast relative to the correct outcome). Secondary endpoints were exact-set accuracy, set F1, negative log loss, expected calibration error, Hartley nonspecificity, and vacuous-source invariance in BetP total-variation distance. Paired 95% intervals used 10,000 bootstrap replicates; directional confirmatory hypotheses used 100,000 paired sign-flip permutations with Holm correction. The decision-rule sensitivity analysis was conducted afterward and is therefore exploratory.

3.10 Baselines and outcomes

The RAMDocs comparison included majority top-k, relevance-weighted top-k, global DS rule top-one and top-k, factorized DS rule with and without provenance, and factorized PCR6 with and without provenance (factorized fusion is defined in Section 3.3). Exact-set accuracy was one only when the predicted and gold answer sets were identical. Set precision, recall, and F1 measured partial recovery. Expected calibration error (ECE) used exact-set correctness and the reported joint confidence. The NLI experiment also reported Brier error and negative log-likelihood [20].

Paired bootstrap intervals resampled per-item differences. The oracle RAMDocs analysis used 2,000 replicates with seed 20260914. The NLI analysis used 4,000 replicates, with seeds 20260915 and 20260916 for F1 and Brier error respectively. All intervals are percentile intervals. Secondary comparisons were exploratory and were not adjusted for multiplicity.

4 Results

4.1 Duplicate misinformation reveals the PCR6 advantage

The duplicate-misinformation condition favored PCR6 over the provenance-matched DS baseline on both reported answer-recovery metrics. Provenance-aware PCR6 reached 80.6% exact-set accuracy and 92.31% set F1, while provenance-aware DS reached 78.2% and 90.98%. The exact-set difference was 2.4 percentage points, with a 95% bootstrap interval from 1.2 to 3.8 points (2,000 paired bootstrap replicates, seed 20260914, the same protocol used for the F1 contrast). The paired F1 difference was 1.33 percentage points, with a 95% bootstrap interval from 0.60 to 2.33 points. Both intervals exclude zero, providing interval-based evidence for the exact-set and F1 advantages alike.

This answer-recovery gain did not improve calibration under the same heuristic mass mapping: ECE was higher for provenance-aware PCR6 than for the matched DS rule (0.193 versus 0.160). By contrast, PCR6 reduced both Brier error and ECE under the NLI-derived mapping in Section 4.2. Calibration conclusions are therefore conditional on how textual evidence is mapped to BBAs.
The same condition shows why the matched comparison matters. Without provenance control, the DS rule obtained 46.4% exact-set accuracy and PCR6 obtained 46.0%. With provenance control, both rose above 78%. Most of the resilience to exact repetition therefore comes from modeling dependence, and PCR6 contributes a smaller additional gain on top of it. The two mechanisms are complementary rather than interchangeable: discounting removes the false weight of repeated text, and PCR6 redistributes the conflict that remains.

Table 1. RAMDocs results across stress conditions

Condition	Method	Exact set	Set F1	ECE
Original	Majority top-k	64.8	84.61	0.241
Original	DS rule, factorized + provenance	75.4	89.18	0.161

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

Original	PCR6, provenance-aware	75.0	89.05	0.133
Duplicate misinformation	Majority top-k	42.4	69.08	0.490
Duplicate misinformation	DS rule, factorized	46.4	71.91	0.428
Duplicate misinformation	PCR6, factorized	46.0	71.78	0.347
Duplicate misinformation	DS rule, provenance-aware	78.2	90.98	0.160
Duplicate misinformation	PCR6, provenance-aware	80.6	92.31	0.193
Misinformation dominant	DS rule, provenance-aware	67.8	85.01	0.156
Misinformation dominant	PCR6, provenance-aware	67.6	84.58	0.132

Exact-set and set-F1 values are percentages. ECE is expected calibration error, and lower values are better. Selected rows are shown; the full grid, including the relevance-weighted and global-fusion baselines and, per Section 3.4, the per-condition distribution of S, is included in the released results files.

4.2 NLI masses preserve a calibration advantage

With NLI-derived masses, provenance-aware DS rule and PCR6 selected the same answer sets in every condition, so their exact-set accuracy and set F1 are identical. PCR6 nevertheless produced lower Brier error and lower ECE. In the original condition, Brier error fell from 0.3504 to 0.3152; using the unrounded item-level values (0.350446 and 0.315183), the paired difference was −0.0353 with a 95% interval from −0.0464 to −0.0251. Duplicate misinformation produced the same values because exact copies were collapsed by provenance.

Under misinformation dominance, Brier error fell from 0.3504 for the DS rule to 0.3314 for PCR6, a paired difference of −0.0190 with an interval from −0.0267 to −0.0119. Both intervals excluded zero. PCR6 also improved F1 over unstructured NLI pooling by 1.61 points in the original condition and by 6.33 points under misinformation dominance. The latter comparison measures the whole structured fusion pipeline rather than PCR6 alone, because the matched DS rule reached the same answer decisions.

Table 2. NLI-assisted fusion results

Condition	Method	Exact set	Set F1	Brier	ECE
Original	NLI pooling (unstructured)	54.0	79.33	0.4211	0.4218
Original	DS rule, provenance-aware	59.4	80.95	0.3504	0.2950
Original	PCR6, provenance-aware	59.4	80.95	0.3152	0.2398
Duplicate misinformation	DS rule, provenance-aware	59.4	80.95	0.3504	0.2950
Duplicate misinformation	PCR6, provenance-aware	59.4	80.95	0.3152	0.2398
Misinformation dominant	NLI pooling (unstructured)	40.0	72.95	0.5672	0.5710
Misinformation dominant	DS rule, provenance-aware	56.2	79.28	0.3504	0.3062
Misinformation dominant	PCR6, provenance-aware	56.2	79.28	0.3314	0.2664

Exact-set and set-F1 values are percentages. Lower Brier and ECE values indicate better probabilistic performance.

4.3 ConflictQA negative control

ConflictQA produced no unique PCR6 advantage. In the balanced condition, the evidence methods achieved 100% accuracy. When a false parametric-memory answer was repeated as copies of one provenance source, methods without dependence control failed, whereas both provenance-aware DS rule and PCR6 reached 100%. When the false memories were treated as independent sources, neither method recovered the correct answer. These results confirm that provenance modeling is necessary on this frame, and they show that a binary decision task offers too little structure to separate the two combination rules.

4.4 PCR6+ confirmatory and decision-sensitivity results

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

In controlled partial-ignorance conflict, PCR6+ reduced BetP Brier error from 0.3185 for PCR6 to 0.2733, a paired difference of −0.0452 (95% CI −0.0457 to −0.0447; Holm-adjusted one-sided p = 0.00003). Both PCR rules achieved 100% top-1 accuracy, compared with 98.5% for the DS rule. Adding a vacuous BBA changed ordinary PCR6 by a mean BetP total-variation distance of 0.0803 but left PCR6+ unchanged to numerical precision. The DS rule nevertheless had the lowest BetP Brier error in the controlled conflict conditions, so the result establishes a structural PCR6+ advantage over PCR6 rather than universal dominance over the DS rule.

External validation produced 428 eligible RAMDocs items; 260 met the prespecified PCR6+ applicability criteria. In that subset, ordinary PCR6 had lower BetP Brier error than PCR6+ (0.2354 versus 0.2629), while both reached set F1 = 0.7718. Both PCR rules had lower Brier error than the DS rule (0.3263). Thus, although the broad focal mass that PCR6+ removes is structurally undesirable in the vacuous-source test, retaining that mass on Θ under ordinary PCR6 happened to improve the BetP Brier score (a proper score reflecting both calibration and resolution, not calibration alone) in the RAMDocs applicability subset for this fixed NLI-to-BBA mapping.

Decision sensitivity further qualified the conclusion. On the RAMDocs applicability subset, DSmP increased PCR6+ exact-set accuracy from 0.4885 to 0.4962 and set F1 from 0.7718 to 0.7788, but worsened Brier error from 0.2629 to 0.4279. Maximum belief produced the best exploratory PCR6+ hard-decision result (exact set 0.5077; set F1 0.7859). The minimum belief-interval-distance criterion reproduced the BetP answer sets for the DS rule, PCR6, and PCR6+ in this subset. A different decision transform can therefore alter ranking, but the effect is method- and dataset-dependent and may come at the expense of probabilistic calibration. Figure 1 summarizes the primary BetP Brier comparisons across the controlled benchmark and RAMDocs external validation.

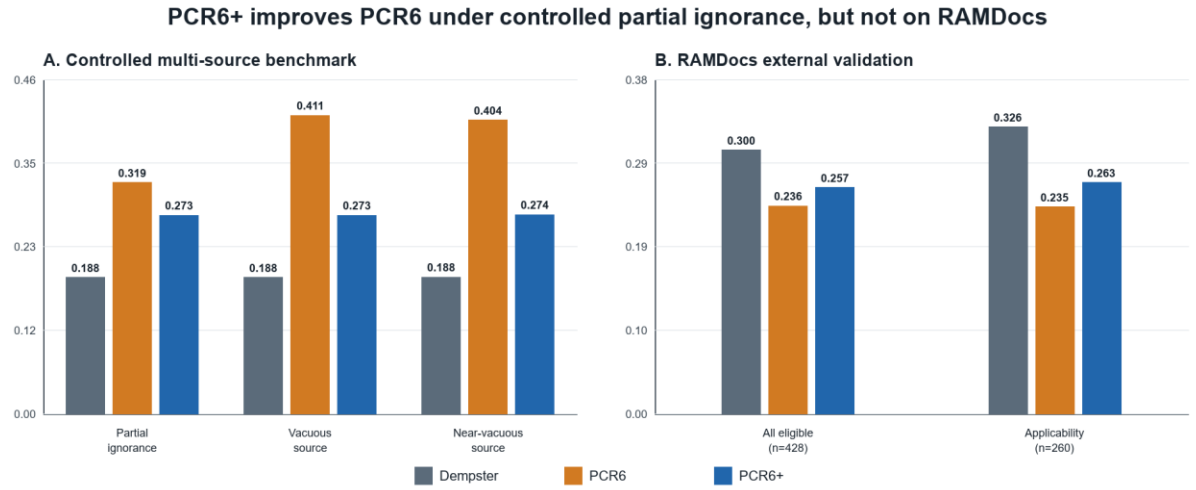

Figure 1. BetP Brier error in the controlled multi-source benchmark and RAMDocs external validation. Lower is better.

Table 3. Confirmatory multi-source PCR6+ results

Evaluation	Rule	Accuracy / exact set (%)	Set F1 (%)	BetP Brier
Controlled partial conflict	DS rule	98.50	—	0.1878
Controlled partial conflict	PCR6	100.00	—	0.3185
Controlled partial conflict	PCR6+	100.00	—	0.2733
RAMDocs applicability	DS rule	48.85	77.05	0.3263

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

Evaluation	Rule	Accuracy / exact set (%)	Set F1 (%)	BetP Brier
RAMDocs applicability	PCR6	49.23	77.18	0.2354
RAMDocs applicability	PCR6+	48.85	77.18	0.2629

BetP Brier is a proper probabilistic score; lower values are better. The controlled column reports top-1 accuracy and the RAMDocs column exact-set accuracy.

Table 4. Exploratory decision-rule sensitivity on the 260-item RAMDocs applicability subset

Fusion	Decision	Exact set (%)	Set F1 (%)	Brier
DS rule	BetP	48.85	77.05	0.3263
DS rule	DSmP	48.46	77.12	0.4690
DS rule	Minimum BI distance	48.85	77.05	—
PCR6	BetP	49.23	77.18	0.2354
PCR6	DSmP	49.23	77.76	0.4389
PCR6	Minimum BI distance	49.23	77.18	—
PCR6+	BetP	48.85	77.18	0.2629
PCR6+	DSmP	49.62	77.88	0.4279
PCR6+	Minimum BI distance	48.85	77.18	—
PCR6+	Maximum belief	50.77	78.59	—
PCR6+	Maximum plausibility	49.62	77.50	—

DSmP uses $\varepsilon = 0.001$. Minimum belief-interval distance, maximum belief, and maximum plausibility are ranking rules, so Brier scores are not applicable. Every comparison shown in this table is exploratory. The displayed rows are a selected sensitivity analysis rather than a complete factorial grid, and no multiplicity adjustment was applied; none of these values should be read as a confirmatory endpoint.

5 Discussion

5.1 Why DS and PCR6 usually agree, and where they should not

On the two-focal-element frame used here ($\{a_i\}$, Θ), a conflicting product arises only when two or more sources commit their singleton mass to different candidates. If the contributing masses are numerically close, DS and PCR6 redistribute the conflict almost identically because both rules split it in proportion to essentially the same competing masses. The rules can diverge more clearly when the sources contributing to a conflicting product carry unequal weight, as when several near-duplicate documents push a disproportionate share of mass toward one candidate. Once provenance discounting is applied, most remaining conflicts in the original and misinformation-dominant conditions involve a small number of roughly balanced sources. In those two conditions, PCR6 point estimates were slightly below those of the matched DS rule (75.0% versus 75.4% exact set in the original condition; 67.6% versus 67.8% exact set and 84.58% versus 85.01% set F1 under misinformation dominance). No paired intervals were computed for these differences, so they are reported descriptively rather than interpreted as evidence of a disadvantage. Without provenance control, the duplicate-misinformation condition produced 46.4% versus 46.0% exact-set accuracy; this residual difference is also small and is not the headline comparison. Table 1 therefore shows large gains from provenance discounting and a small additional PCR6 F1 gain under provenance control in the duplicated condition that is supported by a positive paired interval. This pattern is consistent with the theoretical expectation of modest divergence outside high-conflict configurations.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

The NLI results (Section 4.2) show a second, distinct source of divergence: even when PCR6 and DS select the same answer, their intermediate mass distributions differ enough to change the resulting Brier score and ECE, because the pignistic probability is sensitive to how conflicting mass was distributed even when it does not change which candidate has the largest share. This calibration-level effect does not require a pathological conflict configuration and was observed in every condition tested.

5.2 Where PCR6 pays off

The experiment identifies a specific and defensible domain in which PCR6 improves evidence fusion for RAG. Multiple valid answers create several entity-specific decisions; duplicated misinformation creates severe local conflict; and provenance discounting prevents copies from dominating through sheer frequency. Under that combination, PCR6 recovered more complete answer sets than the DS rule from the same evidence representation, and the positive F1 interval shows that the result was not driven by a handful of questions.

The NLI experiment sharpened the interpretation of the contribution. PCR6 did not alter the ranking of candidate answers relative to the DS rule, but it improved proper probabilistic scores in every condition. That difference matters whenever a system must decide whether to answer, to present several possibilities, to request verification, or to abstain. Two rules can agree on the answer and still disagree on whether their confidence reflects observed correctness, and the second disagreement is what determines the cost of acting on the output.

Future work should calibrate entailment scores on held-out RAG conflicts and compare several mappings, including masses on candidate unions, explicit contradiction focal sets, and open-world ignorance.

The present study tests PCR6 on an exclusive answer frame. The next theoretical step is a hybrid DSm frame induced by semantic relations among claims. Generated answers are often partially overlapping rather than strictly identical or disjoint. One answer may name a correct entity but omit a qualifier; another may combine two compatible interpretations. A semantic graph could classify candidate pairs as equivalent, incompatible, overlapping, or unresolved, and only justified intersections would remain nonempty in the hybrid model.

5.3 Toward an auditable fusion pipeline

A practical DSmT-RAG system should keep conflict, ignorance, and dependence as separate quantities. Conflict records incompatible support. Ignorance records insufficient specificity. Dependence records shared provenance, prompts, training data, or tool outputs. Maintaining these distinctions allows the system to produce an auditable explanation that names the claims and sources responsible for uncertainty instead of returning a single opaque confidence score.

5.4 Limitations

- Candidate answers came from the RAMDocs document annotations, so the main experiment measures the fusion layer after perfect candidate extraction; the fully automatic pipeline is the subject of the next stage of this work.
- Entity labels were provided by RAMDocs and assigned with a deterministic lexical heuristic rather than learned entity linking.
- The answer frame was exclusive, so the experiment does not yet test the computational or representational advantage of the DSm hyper-power set or of partial-ignorance focal elements.
- The confidence mapping was heuristic, and the NLI model was compact, quantized, and truncated to 256 tokens.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

- The duplicate stress test used exact copies. Paraphrases, syndicated text, common upstream sources, and correlated model agents require more general provenance graphs. The exact-invariance argument for provenance discounting (Section 3.6) is proved for the DS rule, which is associative; it has not been shown, and should not be assumed, that discounting g exact copies to residual-ignorance-equivalent confidence r* yields an equivalent input for simultaneous multi-source PCR6, which is not associative. A companion analysis in this program (Leyva-Vazquez et al., lineage-invariant fusion, in preparation) isolates this question formally; until it is resolved, the provenance-aware PCR6 results in Table 1 should be read as empirically matched, not as proved invariant to the representation of duplicate evidence.
- Bootstrap intervals describe these benchmark items; the secondary comparisons were exploratory and were not corrected for multiplicity.
- Tables 1 and 2 retain the original two-focal-element model ($\{a_i\}$, Θ). The confirmatory extension adds proper candidate unions, but its NLI-to-BBA mapping was fixed rather than calibrated on held-out RAG conflicts; different defensible mappings may change the PCR6 versus PCR6+ calibration ordering.
- BetP remains the primary transform. DSmP, minimum belief-interval distance, maximum belief, and maximum plausibility were evaluated in an exploratory sensitivity analysis; their hard-decision comparisons require independent validation.

5.5 A worked PCR6+ illustration and the multi-source experiment

The worked example provides a transparent comparison of DS, PCR6, and PCR6+ for three sources. The toy computation is explanatory, whereas the separate empirical evaluation is reported in Section 4.4.

Recall the three documents from Section 3.2 on $\Theta = \{a_1, a_2\}$: $m_1(a_1) = 0.70$, $m_1(\Theta) = 0.30$; $m_2(a_2) = 0.60$, $m_2(\Theta) = 0.40$; $m_3(a_1) = 0.50$, $m_3(\Theta) = 0.50$. With $S = 3$, the conjunctive rule (Eq. 3) gives $m\cap(a_1) = 0.34$, $m\cap(a_2) = 0.09$, $m\cap(\Theta) = 0.06$, and total conflict $K = 0.51$. Three products are conflicting: (a_1, a_2, a_1) with mass 0.21, (a_1, a_2, Θ) with mass 0.21, and (Θ, a_2, a_1) with mass 0.09 — note that the second and third products both involve Θ as one factor even though Θ is not, in itself, in conflict with either a_1 or a_2.

Table 5. Exact fusion results for the three-source, two-focal-element toy example

Rule	$m(a_1)$	$m(a_2)$	$m(\Theta)$
Dempster (DS)	0.693878	0.183673	0.122449
PCR6	0.593810	0.268571	0.137619
PCR6+	0.633986	0.306014	0.060000

Computed exactly (rational arithmetic) for the BBAs of Section 3.2. DS renormalizes the whole conflict $K = 0.51$ onto a_1, a_2, and Θ in fixed proportion to their conjunctive masses. PCR6 redistributes each conflicting product to every distinct focal element that appears in it, including Θ whenever Θ is one of the three factors. PCR6+ first discards Θ from any conflicting product in which Θ is not implicated in the actual disagreement — formally, whenever Θ contains every other focal element present in that product — and redistributes only to a_1 and a_2 in that case.

The comparison makes the structural distinction directly visible. DS assigns the least mass to the misinformation-leaning candidate a_2 in this toy case, but does so by global normalization. PCR6 redistributes each conflicting product locally but allows Θ to absorb part of products such as (a_1,a_2,Θ), even though Θ is not itself opposed to either singleton. PCR6+ removes that leakage and leaves $m(\Theta)$ at its non-conflicting conjunctive contribution of 0.06. PCR6+ is slightly more decisive than PCR6 by the singleton gap (0.328 versus 0.325), whereas DS has the largest gap (0.510). The example therefore demonstrates the structural correction made by PCR6+, not predictive superiority over DS.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

This example also shows why the effect was invisible in Tables 1 and 2: those BBAs use only $\{a_i\}$ and Θ and were not designed to isolate conflicting tuples containing broad focal elements. A PCR6 versus PCR6+ difference requires at least three sources and a conflicting tuple with positive mass on a broad set such as Θ or a proper union; the source need not be completely vacuous. The confirmatory extension creates and measures precisely these conditions.

Confirmatory multi-source experiment

The confirmatory experiment combines four design elements: (i) three or more provenance-distinct sources; (ii) generalized BBAs containing proper candidate unions; (iii) matched DS, PCR6, and PCR6+ fusion; and (iv) sensitivity to BetP, DSmP, minimum belief-interval distance, maximum belief, and maximum plausibility. The controlled results verify PCR6+ neutrality to a vacuous BBA and improved Brier error relative to PCR6 under BetP. RAMDocs establishes the boundary: PCR6 retained better BetP calibration, while PCR6+ combined with maximum belief produced the strongest exploratory answer-recovery values.

6 Future research

The most immediate experiment removes the two oracle components. Candidate claims should be extracted directly from passages and linked to automatically discovered entities. A held-out calibration set should map entailment, contradiction, source quality, temporal validity, and answer completeness to generalized BBAs. The system should then choose among a direct answer, several contextualized answers, a request for one more source, and abstention.

A subsequent experiment should test the central representational feature of DSmT. Candidate claims should be organized in a constrained hybrid frame that permits only semantically justified overlaps. The study should compare five ablations: exclusive versus hybrid frame, the DS rule versus PCR6 versus PCR6+, provenance control present versus absent, two-source versus three-or-more-source fusion, and heuristic versus learned mass mapping. That design would reveal whether any gain comes from non-exclusivity, from conflict redistribution, from its PCR6+ refinement, from dependence control, or from better evidence scores.

A third program should extend the source model beyond retrieved documents. Language-model agents, search tools, knowledge graphs, images, and sensor streams can share hidden dependencies. Provenance-aware DSmT can fuse their outputs while preserving a trace of the partial conflicts. Evaluations should report calibration, selective risk, abstention utility, contradiction localization, latency, memory, and focal-set growth in addition to answer accuracy.

Sparse approximations will be needed for larger frames. Candidate pruning, focal-set pruning, clustered PCR, and bounded-error approximations should be evaluated against exact PCR6 and PCR6+. A public benchmark with controlled overlap, conflict, ignorance, and dependence would support independent replication and a rule-selection guide based on observable properties of the fusion problem.

7 Conclusion

This study evaluated the contribution of PCR6 to evidence fusion for retrieval-augmented generation through a matched comparison with Dempster–Shafer combination. It defines the frame of discernment concretely, makes the number of fused sources explicit, analyzes the configurations in which DS and PCR6 are expected to diverge, and evaluates the PCR6+ refinement. On 500 RAMDocs questions, the results identified a context-dependent benefit whose form varied with the evidence representation and the treatment of source dependence.

With heuristic masses and the tested provenance-discounting procedure, PCR6 improved answer-set recovery under duplicated misinformation. Exact-set accuracy increased from 78.2% to 80.6%, while set

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

F1 increased from 90.98% to 92.31%; the paired F1 difference was 1.33 percentage points, with a 95% bootstrap interval of [0.60, 2.33]. The larger improvement relative to methods without provenance control was associated with dependence modeling, establishing the importance of controlling repeated evidence when assessing a combination rule. No answer-recovery advantage was detected in the original distribution, under misinformation dominance, or in the ConflictQA negative control, which Section 5.1 shows is consistent with theory rather than a null result requiring explanation.

With NLI-derived masses, PCR6 and Dempster–Shafer combination selected identical answer sets, but PCR6 achieved lower Brier error and ECE. The paired Brier reductions were 0.0353 in the original condition and 0.0190 under misinformation dominance, with both reported intervals excluding zero. The duplicated condition reproduced the original NLI results because exact copies were collapsed before fusion. These findings show that differences between fusion rules can affect the quality of probability estimates even when answer selection remains unchanged.

The exact three-source toy computation in Section 5.5 shows how PCR6+ prevents over-attribution of conflict to total ignorance. The completed 1,000-item benchmark and 428-item RAMDocs validation then show that this structural benefit is real but loss-dependent: PCR6+ improves PCR6 under controlled broad/vacuous focal evidence, whereas ordinary PCR6 gives better BetP calibration on RAMDocs under the frozen NLI mapping. Exploratory maximum-belief decision support gives PCR6+ the best hard-decision values on the applicability subset.

The contribution is therefore a controlled empirical characterization of when PCR6 and PCR6+ add value within an auditable RAG fusion pipeline. Its scope remains bounded by annotation-derived candidates, supplied entity information, an exclusive answer frame, exact-text provenance identification, and a fixed NLI-to-BBA mapping. The next stage should calibrate mass estimation on held-out conflicts, extend provenance from hashes to lineage graphs, test semantically justified hybrid frames, and validate decision transformations against explicit operational losses.

Data availability statement

The experimental package includes source code, frozen configuration and protocol files, implementation tests, aggregate metrics, item-level predictions, NLI scores, bootstrap outputs, confirmatory sign-flip tests, and decision-sensitivity results. It is publicly available at https://github.com/mleyvaz/dsmt-rag-pcr6. RAMDocs is available from https://github.com/HanNight/RAMDocs and ConflictQA from https://huggingface.co/datasets/osunlp/ConflictQA.

Author contributions

F. Smarandache: conceptualization, DSmT theoretical perspective, methodology, validation, writing (review and editing). J. Dezert: theoretical review of the fusion methodology and frame of discernment, formulation of the PCR6+ comparison and the three-source, partial-ignorance experiment, writing (review and editing). M. Leyva-Vázquez: software, formal analysis, investigation, data curation, visualization, writing (original draft). All three authors discussed the results and approved the final manuscript.

Competing interests

The authors declare no competing financial interests. F. Smarandache and J. Dezert are among the originators and principal contributors to the DSmT and PCR research program evaluated in this study, including the PCR6+ refinement discussed in Section 5.5. The protocol was designed so that PCR6, PCR6+, and Dempster-Shafer combination receive identical masses, identical entity grouping, and identical provenance treatment, and all comparative claims are limited to the conditions and losses actually

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

evaluated. The illustrative computation in Section 5.5 and the confirmatory benchmark in Section 4.4 are reported separately; confirmatory and exploratory analyses are explicitly distinguished.

Use of AI tools

The authors disclose the following uses of AI-based tools (OpenAI Codex and Anthropic Claude): support for software development; copy-editing and formatting of the manuscript; and machine-assisted checking of the reported numbers against the released code and outputs. The study design, the experiments, the analysis, and the interpretation of the results are the authors' own. All AI-assisted output was reviewed and verified by the authors, who take full responsibility for the content of this preprint.

References

1. Wang H, Prasad A, Stengel-Eskin E, Bansal M. Retrieval-augmented generation with conflicting evidence. In: Conference on Language Modeling (COLM); 2025. arXiv:2504.13079. doi:10.48550/arXiv.2504.13079.

2. Shafer G. A Mathematical Theory of Evidence. Princeton, NJ: Princeton University Press; 1976.

3. Smarandache F, Dezert J. Proportional conflict redistribution rules for information fusion. 2004. arXiv:cs/0408064. https://fs.unm.edu/PCRrules.pdf

4. Smarandache F, Dezert J. Information fusion based on new proportional conflict redistribution rules. In: Proceedings of the 7th International Conference on Information Fusion (Fusion 2005). Philadelphia, PA: IEEE; 2005. doi:10.1109/ICIF.2005.1591955. https://fs.unm.edu/PCRrules.pdf

5. Dezert J, Smarandache F. Combining uncertain and paradoxical evidences for DSm hybrid models. 2003. arXiv:math/0310135. https://fs.unm.edu/DSmT2003.htm

6. Dezert J, Smarandache F. An introduction to DSmT. 2009. arXiv:0903.0279. https://fs.unm.edu/IntroductionToDSmT.pdf

7. Smarandache F, Dezert J, Tchamova A, editors. Advances and Applications of DSmT for Information Fusion, Collected Works Volume 5. 2023. https://digitalrepository.unm.edu/math_fsp/620.

7a. Dezert T, Dezert J, Smarandache F. Improvement of Proportional Conflict Redistribution Rules of Combination of Basic Belief Assignments. Journal of Advances in Information Fusion. 2021;16(1):48–74.

7b. Han D, Dezert J, Yang Y. Belief Interval-Based Distance Measures in the Theory of Belief Functions. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2018;48(6):833–850. doi:10.1109/TSMC.2016.2628879.

7c. Dezert J, Han D, Tacnet J-M, Carladous S, Yang Y. Decision-Making with Belief Interval Distance. In: Proceedings of the 4th International Conference on Belief Functions (BELIEF 2016); Prague, Czech Republic; September 21–23, 2016. p. 87–94.

7d. Martin A, Osswald C. A new generalization of the proportional conflict redistribution rule stable in terms of decision. In: Smarandache F, Dezert J, editors. Advances and Applications of DSmT for Information Fusion. 2006. p. 69–88. arXiv:0806.1797. doi:10.48550/arXiv.0806.1797.

8. Smarandache F, Dezert J. Uncertainty modeling and evidence combination: DST and DSmT in perspective. HyperSoft Set Methods in Engineering. 2026;5:15–26. doi:10.61356/j.hsse.2026.5644.

9. Dempster AP. Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics. 1967;38(2):325–339. doi:10.1214/aoms/1177698950.

10. Zadeh LA. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. AI Magazine. 1986;7(2):85–90. doi:10.1609/aimag.v7i2.542.

11. Yager RR. On the Dempster-Shafer framework and new combination rules. Information Sciences. 1987;41(2):93–137. doi:10.1016/0020-0255(87)90007-7.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

12. Smets P, Kennes R. The transferable belief model. Artificial Intelligence. 1994;66(2):191–234. doi:10.1016/0004-3702(94)90026-4.

13. Dubois D, Prade H. Representation and combination of uncertainty with belief functions and possibility measures. Computational Intelligence. 1988;4(3):244–264. doi:10.1111/j.1467-8640.1988.tb00279.x.

14. Xie J, Zhang K, Chen J, Lou R, Su Y. Adaptive chameleon or stubborn sloth: revealing the behavior of large language models in knowledge conflicts. In: Proceedings of the 12th International Conference on Learning Representations (ICLR); 2024. arXiv:2305.13300.

14a. Saha RR, Lakshmanan LVS, Ng RT. Stance Detection with Explanations. Computational Linguistics. 2024;50(1):193–235. doi:10.1162/coli_a_00501.

14b. Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL); 2023. p. 9802–9822. arXiv:2212.10511.

15. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC²), NeurIPS; 2019. arXiv:1910.01108.

16. Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of NAACL-HLT; 2018. p. 1112–1122. doi:10.18653/v1/N18-1101.

17. Pietkiewicz T. Fusion of identification information from ESM sensors and radars using Dezert-Smarandache theory rules. Remote Sensing. 2023;15(16):3977. doi:10.3390/rs15163977.

18. Dong Y, Li X, Dezert J, Khyam MO, Rahim MNA, Ge SS. Dezert-Smarandache theory-based fusion for human activity recognition in body sensor networks. IEEE Transactions on Industrial Informatics. 2020;16(11):7138–7149. doi:10.1109/TII.2020.2976812.

19. Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems; 2018. vol. 31.

20. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 2007;102(477):359–378. doi:10.1198/016214506000001437.

20a. Dezert J, Smarandache F. A new probabilistic transformation of belief mass assignment. In: Proceedings of the 11th International Conference on Information Fusion; 2008. arXiv:0807.3669.

Additional studies on limitations and counterexamples involving Dempster's rule

21) Dezert J, Tchamova A, Han D, Tacnet JM. Why Dempster's Rule doesn't behave as Bayes rule with Informative Priors. In: Advances and Applications of DSmT for Information Fusion, Collected Works Volume 4. American Research Press (ARP); 2015. p. 189–193. http://fs.unm.edu/WhyDempsterFusion.pdf

22) Smarandache F, Dezert J, Kroumov V. Examples where the Conjunctive and Dempster's Rules are Insensitive. In: Advances and Applications of DSmT for Information Fusion, Collected Works Volume 4. American Research Press (ARP); 2015. p. 211–213. http://fs.unm.edu/ExampleConjunctive.pdf

23) Dezert J, Wang P, Tchamova A. On The Validity of Dempster-Shafer Theory. In: Proceedings of the 15th International Conference on Information Fusion; Singapore; 9–12 July 2012. Republished in: Advances and Applications of DSmT for Information Fusion, Collected Works Volume 4. American Research Press; 2015. p. 163–168.

24) Dezert J, Tchamova A. On the behavior of Dempster's rule of combination. BFTA 2011, Autrans, France, April 2011.

SciELO Preprints - This document is a preprint and its current status is available at: https://doi.org/10.1590/SciELOPreprints.18188

25) Dezert J, Smarandache F, Khoshnevisan M. Counter-examples to Dempster's rule of combination. In: Smarandache F, Dezert J, editors. Advances and Applications of DSmT for Information Fusion, Collected Works Volume I. American Research Press; 2004. Chapter V, p. 105–121. http://fs.unm.edu/CounterExamplesToDempstersRuleOfCombination.pdf

26) Dezert J, Tchamova A, Smarandache F, Konstantinova P. Target Type Tracking with PCR5 and Dempster's Rules: A Comparative Analysis. Fusion 2006 Int. Conf. on Information Fusion, Firenze, Italy, July 2006.

27) Tchamova A, Dezert J. On the Behavior of Dempster's Rule of Combination and the Foundations of Dempster-Shafer Theory. 2012 IEEE 6th Int. Conf. on Intelligent Systems, Sofia, Bulgaria, Sept. 2012.

28) Dezert J, Tchamova A. A fundamental contradiction in Dempster-Shafer Theory. Panel Discussion on Uncertainty, Fusion 2012, Singapore.

29) Dezert J. On the inconsistency of Dempster's rule of combination with the probability calculus. Journée Fusion, Télécom Bretagne, Brest, France, November 17, 2011.

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.