Iranian Journal of Fuzzy Systems

Volume 23, Number 3, (2026), pp. 141-163

Original paper

University of Sistan and Baluchestan

Allo-Self-RAG: Fuzzy aggregation of internal and external critique signals for improved Self-RAG evaluation

F. Hosseini [1] and M. Eftekhari [2]

[1] *Department of computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran*
[2] *Department of computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran*
[2] *Visiting Researcher at Institute for Applied Computer Science (InfAI), Nature-Inspired Machine Intelligence, Dresden, Germany*

fatemeh.hosseini.g108@gmail.com, m.eftekhari@uk.ac.ir

Abstract

Retrieval-Augmented Generation (RAG) systems play a crucial role in grounding Large Language Models (LLMs) with external knowledge. However, existing architectures such as Self-RAG employ static linear aggregation of internal critique tokens, which requires manual tuning and inadequately models the non-linear interactions underlying retrieval and generation. Moreover, exclusive reliance on self-critique can introduce confirmation bias and hallucinations. To overcome these limitations, this work introduces Allo-Self-RAG, a neuro-symbolic framework that integrates fuzzy logic with RAG. From a dual-process-inspired perspective, Allo-Self-RAG is framed as a structured enhancement over the heuristic Self-RAG baseline: the standard Self-RAG pipeline is closer to System-1-like post-retrieval behavior, whereas Allo-Self-RAG introduces a more System-2-like evaluation layer through structured signal aggregation. A Fuzzy Inference System (FIS) adaptively fuses internal self-critique tokens with external allo-critique signals from an independent reranker, replacing static linear aggregation with rule-guided score integration and a rule-based revision mechanism for conflicting evidence. When this evaluation stage detects ambiguity or conflicting evidence among top-ranked candidates, the framework automatically invokes a synthesis step to reconcile contradictions and produce a more reliable consensus answer. Simulated Annealing (SA) is employed to optimize fuzzy membership functions automatically using a small calibration dataset, eliminating manual parameter tuning. Extensive experimental evaluation demonstrates that Allo-Self-RAG consistently outperforms the Self-RAG baseline, achieving 56.61% accuracy on PopQA (+1.45% improvement), 66.98% on ARC-Challenge (+1.03% improvement), and 67.51% on PubHealth (+1.01% improvement), showing reliable gains across retrieval-augmented question answering benchmarks.

Keywords: Retrieval-augmented generation, fuzzy inference systems, simulated annealing, external evaluator, evidence synthesis, neuro-symbolic approach.

1 Introduction

LLMs have transformed Natural Language Processing (NLP) with advanced capabilities in knowledge retrieval and reasoning. However, deploying these models in critical domains, such as medicine or science, remains risky due to hallucinations, where models often confidently generate plausible but incorrect facts [16]. RAG has emerged as a standard solution to this problem, as it anchors generation to verifiable external documents [25]. Yet, the reliability of a RAG system depends heavily on the quality of the retrieval step. Standard retrievers often return noisy or conflicting information, leading to the lost-in-the-middle phenomenon, where the LLM fails to identify the correct answer amidst irrelevant data [1, 27].

Corresponding Author: M. Eftekhari
Received: June 2026; Accepted: Invited paper by Prof. M. Eftekhari.

Much research has focused on optimizing retrieval algorithms. However, two critical decision-making phases remain underexplored. Many RAG systems unconditionally trigger retrieval without assessing whether external knowledge is necessary to answer the given query. This unconditional triggering leads to unnecessary computational overhead and potential noise injection when the LLM already possesses sufficient parametric knowledge to answer accurately. Beyond this, the post-retrieval decision-making phase—involving filtering, selecting, and aggregating retrieved contexts—constitutes a significant bottleneck. Current systems often implicitly assume that top-k retrieved documents are relevant and factually supportive. This assumption fails in complex reasoning scenarios. Even a high-performing retriever may introduce noise that degrades generation quality. Consequently, if the downstream selection mechanism cannot filter out contradictory evidence, the final output suffers. Robust evaluation protocols are necessary to intervene both before and after retrieval to ensure efficient and accurate knowledge integration.

To address these challenges, recent frameworks have shifted towards Active RAG paradigms, which attempt to internalize the evaluation process within the LLM itself. Self-RAG [2] is a prominent example, introducing a self-reflective mechanism where the generator is fine-tuned to produce special reflection tokens alongside its output. These tokens include `[Retrieval]` to decide whether retrieval is needed, `[IsREL]` to assess relevance of retrieved passages, `[IsSUP]` to verify factual support, and `[IsUSE]` to evaluate overall utility. This mechanism enables the model to perform on-demand retrieval and critique both retrieved documents and generated outputs during inference, allowing dynamic filtering of irrelevant passages and unsupported claims, providing adaptive control over when external knowledge is retrieved.

While Self-RAG introduces a promising paradigm for adaptive retrieval and critique, several critical limitations constrain its robustness and generalizability. Self-critique mechanisms that rely solely on model-generated signals are vulnerable to confirmation bias: when the same model both generates and evaluates its own outputs, it may reinforce existing errors rather than detect them, particularly in adversarial or ambiguous contexts where external validation is essential.

Linear aggregation of critique signals introduces a flawed compensability assumption. Self-RAG combines reflection tokens (e.g., relevance, utility, and support) using static weighted sums of the form $\text{Score} = w_1 \cdot \text{Rel} + w_2 \cdot \text{Util} + w_3 \cdot \text{Sup}$. This approach allows deficiencies in one dimension to be offset by strengths in another. For instance, a retrieved passage that is highly relevant ($\text{Rel} \approx 1.0$) and useful ($\text{Util} \approx 1.0$) but contains factually contradictory evidence ($\text{Sup} \approx 0.0$) would still receive a relatively high score (≈ 0.67), thereby masking the contradiction. In contrast, Allo-Self-RAG does not rely on a purely additive scoring function. Instead, it employs an interpretable fuzzy rule base with a rule-based revision mechanism, allowing adverse evidence to reduce the final evaluation in a structured, non-linear manner rather than being trivially compensated for by strong performance on other dimensions. Moreover, manually tuning aggregation weights (e.g., $w_1 = 0.3, w_2 = 0.4, w_3 = 0.3$) is labor-intensive and domain-specific, and offers no principled mechanism for automatic calibration across diverse reasoning scenarios. Unlike linear score fusion or regression-based weighting schemes, fuzzy inference explicitly models interactions among evaluation criteria through human-interpretable rules, eliminating the need to encode such relationships indirectly through coefficients and thresholds.

Hard numerical thresholds impose artificial boundaries on inherently uncertain linguistic phenomena. Natural language reasoning involves vagueness, polysemy, and context-dependent interpretation, yet Self-RAG's binary decision rules treat a relevance score of 0.49 versus 0.51 as categorically distinct, triggering entirely different retrieval behaviors despite negligible semantic difference. This rigidity is incompatible with the graded, approximate nature of human judgment, which operates on degrees of membership rather than crisp cutoffs—a mismatch that motivates the use of fuzzy logic for modeling such uncertainty. This comparison to human judgment is used here only as an intuitive motivation for graded uncertainty handling, not as a literal cognitive equivalence between fuzzy inference and human reasoning. Finally, per-document critique evaluates each retrieved passage independently without modeling dependencies across multiple sources, which prevents coherent synthesis in complex queries requiring integration of evidence from several documents.

These limitations also reflect broader challenges in building adaptive, interpretable, and efficient machine learning systems. The rigidity of hard thresholds and linear aggregation contrasts with recent advances in neuro-symbolic integration [24], which leverage fuzzy logic and symbolic rules to model uncertainty and enforce rule-constrained and penalty-sensitive decision boundaries. Finally, the inefficiency of unconditional critique aligns with the motivation for dynamic inference [28], where computational resources are allocated adaptively based on input complexity. These connections position RAG critique as part of a broader effort toward adaptive, interpretable, and efficient AI systems.

To address these limitations, this paper introduces Allo-Self-RAG, a neuro-symbolic framework that integrates Fuzzy Inference Systems (FIS) with a dual-stream critique architecture for robust RAG evaluation. Drawing on the integration of fuzzy systems within machine learning [9], Allo-Self-RAG models critique signals (relevance, support, utility) as linguistic variables with graded membership functions, replacing crisp cutoffs (score > 0.5) with smooth boundaries that capture natural language vagueness (e.g., "highly relevant"). From a cognitive perspective, this contribution is

best understood at the level of overall framework behavior: the standard Self-RAG baseline is closer to a System-1-like heuristic post-retrieval decision process, whereas Allo-Self-RAG adds a more System-2-like layer of structured signal aggregation and ambiguity-aware reassessment.

Unlike Self-RAG's self-critique with compensatory linear scoring, Allo-Self-RAG integrates internal self-critique with external allo-critique from an independent reranker, mitigating confirmation bias. Its Mamdani-type FIS uses interpretable fuzzy rules together with an externally grounded revision mechanism, whereby low external relevance acts as a strong adverse signal within the aggregated evaluation process. This design reduces the risk that internally confident but externally weak evidence is scored too favorably, while avoiding the overstatement of a formal hard-veto operator.

Compared with linear regression or static weighted fusion, this fuzzy decision layer offers three practical advantages: (i) it captures non-linear interactions among relevance, support, utility, and external relevance through explicit rules; (ii) it provides greater interpretability by expressing decision behavior in linguistic form rather than opaque fitted coefficients alone; and (iii) it enables adverse evidence to influence the decision more directly through structured revision, instead of relying solely on compensatory averaging.

The membership functions of the fuzzy inference system are calibrated via SA [22] on a small calibration set by optimizing their parameters, thereby reducing the need for manual membership-function tuning across diverse reasoning scenarios.

The proposed framework introduces four key innovations:

- **Hybrid Allo-Self Critique (Dual-Source Critique Integration):** This mechanism breaks the self-reinforcing feedback loop by integrating an independent external evaluator (allo-critique) with the generator's introspection (self-critique). This dual-signal design enables more robust multi-perspective assessment than relying on model-internal critique alone.

- **Revision Mechanism** (Adverse-Evidence Handling): Unlike linear aggregation, which can allow weak factual support or poor external grounding to be offset by high relevance or utility, the fuzzy inference system applies an externally grounded, rule-based revision mechanism to adverse evidence. In particular, low external relevance is treated as a privileged negative signal through the rule base, exerting structured and interpretable downward pressure on the final evaluation. This design preserves the neuro-symbolic nature of the framework while remaining faithful to the behavior of the Mamdani fuzzy inference system.

- **Adaptive Fuzzy Calibration via SA:** Rather than requiring large-scale retraining, the parameters of fuzzy membership functions are calibrated via SA using a small calibration set. This data-efficient optimization reduces the need for manual tuning of membership-function parameters and supports adaptation to domain-specific score distributions.

- **Dynamic Generative Synthesis (Resource-Aware Reasoning):** The system dynamically modulates generative reasoning based on the confidence and ambiguity of retrieved evidence. Complex synthesis is triggered only when necessary, enabling dynamic computation akin to conditional inference mechanisms, ensuring computational resources are utilized efficiently without compromising answer reliability.

Taken together, these four design elements make the proposed framework more structured, ambiguity-aware, and selectively deliberative than the standard Self-RAG pipeline. Conceptually, Allo-Self-RAG can be viewed as more System-2-like in its overall decision behavior, as it supports structured aggregation, ambiguity-sensitive evaluation, and conditional deeper assessment rather than relying primarily on immediate self-critique signals.

2 Related work

RAG has emerged as a foundational paradigm for grounding large language model outputs in external knowledge, reducing hallucinations and improving factual accuracy [25]. Early RAG systems adopted a passive retrieval strategy, retrieving a fixed set of documents once and appending them to the input prompt. While effective in controlled settings, this approach exhibits limited robustness when retrieved contexts are incomplete, noisy, or inconsistent. Subsequent work has improved retrieval quality through dense retrieval methods—DPR [20], ColBERT [21], and Contriever [14], and hybrid approaches combining lexical and neural retrievers [13, 29]. Query optimization techniques [10, 36] and multi-hop retrieval strategies [34, 41] further enhance retrieval effectiveness for complex information needs. Despite these advances, retrieval optimization alone does not address a fundamental challenge: what to do with retrieved content once it has been obtained. Even with perfect retrieval, the generator must still evaluate relevance, resolve

conflicts, and selectively integrate evidence—tasks that require post-retrieval reasoning rather than improved retrieval alone.

To overcome the limitations of passive retrieval pipelines, recent research has shifted toward Active RAG frameworks that regulate retrieval and generation through feedback signals. Self-RAG [2, 12] introduces reflection tokens (`[Retrieval]`, `[IsREL]`, `[IsSUP]`, `[IsUSE]`) enabling the model to decide when retrieval is needed and evaluate relevance, factual support, and utility of retrieved passages. Subsequent work has expanded this paradigm: CRAG [30, 50] triggers corrective retrieval when confidence declines; VERA [3] decomposes claims into atomic statements for fine-grained verification; FLARE [17] performs forward-looking iterative retrieval. Complementary approaches such as Self-Consistency [46], Reflexion [37], and Self-Ask [34] further enhance reasoning through sampling-based consensus, iterative feedback, and sub-question decomposition. A significant challenge for these architectures is their substantial computational overhead due to multiple LLM calls. To mitigate this, orthogonal research focuses on optimizing inference efficiency through adaptive compute strategies [24]. Adaptive-RAG [15] learns to adapt retrieval scope based on question complexity, Speculative RAG [45] proposes a drafting-and-verification mechanism to reduce latency, and RECOMP [48] introduces compression and selective augmentation to reduce context length. These efforts collectively aim to balance performance with computational cost in active RAG pipelines. Despite these advances, self-reflective RAG systems remain constrained by three fundamental limitations. Self-confirmation bias arises when the same model generates and evaluates outputs, potentially reinforcing errors rather than detecting them. Recent work addresses this by integrating deductive reasoning traces to adjudicate conflicting sources and produce citation-linked answers or justified refusals [31], while multi-agent debate frameworks jointly handle ambiguity and misinformation in retrieved documents [42]. In addition, linear aggregation with compensatory scoring can allow deficiencies in one dimension (e.g., factual support) to be offset by strengths in another (e.g., relevance), potentially masking critical failures such as factually contradictory but highly relevant passages. Finally, hard numerical thresholds impose artificial boundaries on inherently uncertain linguistic phenomena, reducing robustness to edge cases and ambiguous queries.

While retrieval optimization focuses on candidate selection, reranking methods aim to refine the ordering of retrieved passages before generation. Traditional approaches employ cross-encoders that score query-document pairs [32], but recent work has shifted toward generative reranking using large language models. RankGPT [38] formulates reranking as a listwise generation task, while RankT5 [55] fine-tunes encoder-decoder models for passage ordering. RankRAG [52] addresses the trade-off between recall and noise by leveraging LLM-based ranking to refine top-k context selection, improving generation accuracy without increasing context length. Context selection has also been reframed as an inference-time data valuation problem, introducing Contextual Influence Value (CI value) to measure each retrieved context's contribution by combining query-aware relevance, list-aware uniqueness, and generator-aware alignment [7]. However, existing reranking methods remain constrained by three key limitations. Compensatory scoring functions can allow high relevance to mask low factual support, making it difficult to express rule-constrained or penalty-sensitive decision behavior when one adverse signal should exert disproportionate influence. Deterministic ranking treats reranking as a fixed ordering task rather than a multi-criteria filtering problem under uncertainty, ignoring the inherent ambiguity in assessing passage quality. Finally, per-document evaluation scores passages independently without modeling dependencies across sources, preventing coherent synthesis when retrieved documents contain conflicting evidence or require multi-hop reasoning to reconcile. Recent approaches like Chain-of-Note [53] enhance robustness by generating contextual notes for each retrieved passage, allowing the model to explicitly reason about relevance and reliability before integration. Similarly, Astute RAG [44] focuses on overcoming imperfect retrieval and knowledge conflicts through sophisticated evidence aggregation and conflict resolution mechanisms. However, these methods still often rely on compensatory scoring or threshold-based decision rules, leaving room for more uncertainty-aware and explicitly rule-structured post-retrieval filtering mechanisms.

The limitations of deterministic ranking and compensatory scoring motivate alternative formalisms that model uncertainty explicitly. Neuro-symbolic approaches merge structured reasoning with neural representations [11]. Among these, fuzzy set theory provides a principled formalism for reasoning over imprecise and linguistically grounded variables, enabling graded decision boundaries rather than rigid thresholds [9]. Recent advances include the Fuzzy Reasoning Chain (FRC) [5], which integrates fuzzy logic within Chain-of-Thought prompting, and MultiRAG [39], which applies fuzzy classifiers to characterize query complexity prior to retrieval. Extensions to multimodal knowledge graph construction employ fuzzy information entropy-driven RAG for industrial process modeling [47], while fuzzy reinforcement learning combined with LLM guidance has been applied to cooperative endovascular robotics [51], illustrating the broader applicability of fuzzy-LLM integration beyond retrieval tasks. Large–small model collaborative frameworks have also employed fuzzy neural decision agents to route uncertain samples between models based on complexity assessment, coupled with self-reflection to refine outputs and achieve substantial efficiency gains in industrial applications [43]. While these approaches demonstrate the utility of soft computing techniques in enhancing pre-retrieval analyses and internal reasoning processes, they primarily focus on early-stage query processing. The post-retrieval decision phase—

where conflicting evidence must be systematically evaluated and integrated under uncertainty—remains comparatively underexplored. Moreover, many existing fuzzy and neuro-symbolic frameworks use static membership functions or fixed rule sets. While such designs support interpretability and implementation simplicity, they may require additional calibration when transferred across domains or exposed to different evidence distributions.

Recent work on efficient language modeling has highlighted the value of selective computation, in which processing effort is allocated adaptively rather than uniformly across inputs [4, 26, 28, 33, 49, 54]. In RAG settings, a related consideration arises in post-retrieval decision-making: candidate contexts may differ substantially in evidential clarity, suggesting the potential value of evaluation mechanisms that respond differently to straightforward and ambiguous cases.

While significant progress has been made in retrieval optimization, active RAG, reranking, and efficient inference, critical challenges remain in integrating uncertainty-aware reasoning into post-retrieval decision-making and mitigating degenerate feedback loops. Existing approaches often rely on deterministic rules, compensatory aggregation, or self-evaluation, failing to capture the nuanced interplay of evidence quality, query context, and conflicting signals.

These limitations are addressed in this work through the introduction of Allo-Self-RAG, a neuro-symbolic framework for uncertainty-aware post-retrieval evaluation. By treating relevance, support, utility, and external relevance as linguistic variables with calibrated membership functions, and by combining adaptive fuzzy inference with external reranking, the framework enables interpretable, rule-constrained, and context-sensitive filtering of retrieved evidence while reducing self-confirmation loops.

3 Methodology

The Allo-Self-RAG framework implements a neuro-symbolic methodology for post-retrieval evaluation. The system operates through two primary phases: parameter optimization and inference execution.

During the optimization phase, fuzzy membership functions are calibrated using an SA procedure. This allows the system to learn optimal decision boundaries based on the noise characteristics and distribution of the target domain, transitioning from heuristic initialization to a data-driven optimal configuration.

Once optimized, the system functions as a robust post-retrieval discriminator, as illustrated in Figure 1. The inference workflow splits into two parallel streams: an external path evaluates documents via a Cross-Encoder to provide objective allo-critique, while an internal path leverages the generator's self-assessment tokens for subjective self-critique. These heterogeneous signals are combined through an FIS engine, producing a unified fuzzy validity score for each candidate answer.

Candidate answers generated during signal extraction are cached for semantic comparison in the synthesis phase, but are excluded from the numerical fuzzy validity scoring to prevent hallucination propagation. Finally, an ambiguity check determines the routing: queries with high consensus proceed to direct output selection, whereas queries with conflicting evidence trigger a Dynamic Generative Synthesis module to reconcile contradictions and produce a final response.

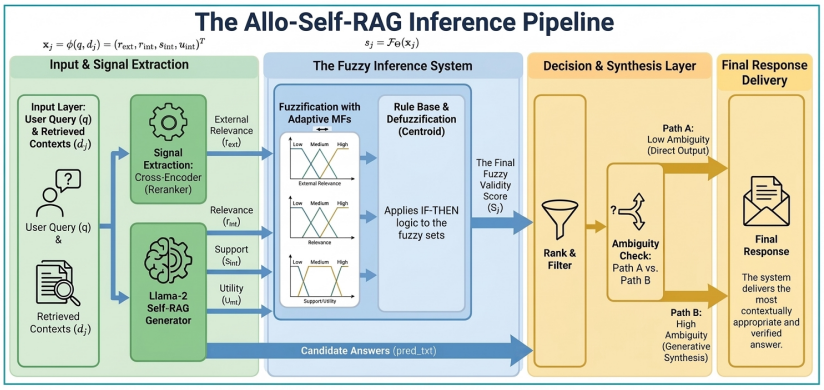

Figure 1: The Allo-Self-RAG Pipeline. Heterogeneous retrieval signals $\mathbf{x}_j = (r_{ext}, r_{int}, s_{int}, u_{int})$ are fused via a Mamdani fuzzy inference system $\mathcal{F}_\Theta$ to produce a validity score s_j for each candidate document d_j. The system dynamically routes candidate answers to either direct selection or generative synthesis based on the fuzzy validity gap between the top-ranked candidates.

3.1 Problem formulation

The task of post-retrieval evaluation is formulated as a document selection and aggregation problem. Let q denote a user query, and let $\mathcal{D}_{\text{ret}} = \{d_1, d_2, \ldots, d_M\}$ represent the initial retrieval pool of M documents obtained via a first-stage retriever. The goal is to identify an optimal subset $\mathcal{D}^* \subseteq \mathcal{D}_{\text{ret}}$ that maximizes end-to-end answer accuracy. This subset is then used to synthesize the final response a.

Standard retrieval systems typically assign a single scalar similarity score to each document. However, a single scalar often fails to capture the multidimensional aspects of document quality. For instance, a document may be topically relevant yet factually unreliable. To address this limitation, each retrieved document d_j is represented by a multi-dimensional feature vector $\mathbf{x}_j \in [0,1]^4$, extracted using a composite evaluation function $\phi(\cdot)$ that integrates signals from both Self-RAG reflection tokens and a Cross-Encoder model:

$$\mathbf{x}_j = \phi(q, d_j) = (r_{ext}, r_{int}, s_{int}, u_{int})^T, \tag{1}$$

where $\phi(q, d_j)$ denotes the feature extraction process conditioned on query q and document d_j (detailed in Section 3.2). The four dimensions capture external relevance (r_{ext}, from Cross-Encoder scoring), internal relevance (r_{int}), support (s_{int}), and utility (u_{int}) from Self-RAG reflection tokens. All signals are normalized to $[0, 1]$ through either inherent probabilistic outputs or post-extraction scaling.

This vector serves as input to an FIS, which applies fuzzy logic rules to evaluate documents while tuning its membership functions. Parameter optimization is performed using SA, allowing the system to adjust its fuzzy boundaries to the specific noise and distribution characteristics of the domain. This approach captures the multidimensional quality of documents and guides the subsequent Dynamic Generative Synthesis module, without implying a conventional neural network for direct prediction.

3.2 Dual-perspective signal extraction

To mitigate the self-correction feedback loop inherent in purely generative evaluation, the Allo-Self-RAG framework aggregates signals from heterogeneous sources. Critique signals are separated into two distinct streams—external and internal—to preserve their individual semantic nuances for downstream fuzzy processing.

3.2.1 External discrimination (Allo-Critique)

The first stream provides an objective evaluation of relevance, referred to as External Discrimination or Allo-Critique. An independent Cross-Encoder model (`ms-marco-MiniLM-L-6-v2`) directly assesses the interaction between the query and each document. Unlike bi-encoders, which encode inputs separately, the Cross-Encoder processes the query-document pair jointly, capturing fine-grained semantic dependencies. The model produces a raw logit λ_j, which is normalized into the $[0, 1]$ range using a sigmoid function:

$$r_{ext} = \frac{1}{1 + \exp(-\lambda_j)}. \tag{2}$$

The resulting scalar r_{ext} serves as an objective anchor. Because the Cross-Encoder is independent of the generator, this signal remains free from internal biases and hallucinations.

3.2.2 Internal reflection (Self-Critique)

The second stream captures the generator's subjective confidence in the retrieved context, referred to as Internal Reflection or Self-Critique. A fine-tuned generator (e.g., `Llama-2-7b-SelfRAG`) processes each retrieved document d_j and generates a candidate answer c_j along with specialized reflection tokens. During generation, the model emits token-level probabilities for three reflection tokens: [IsREL] (internal relevance), [IsSUP] (factual support), and [IsUSE] (overall utility). These probabilities are directly extracted from the model's output distribution and denoted as r_{int}, s_{int}, and u_{int}, respectively. All three signals are inherently normalized to the $[0, 1]$ range as they represent probability values.

The generated candidate answers are collected into a set $\mathcal{C}_{\text{ret}} = \{c_1, c_2, \ldots, c_M\}$, where each c_j corresponds to document $d_j \in \mathcal{D}_{\text{ret}}$. These candidate answers are produced as a side effect during the reflection token extraction process and are cached internally as a dictionary mapping $\{d_j \to c_j\}$. Although excluded from the quantitative fuzzy scoring to prevent bias, these textual candidates are retained for the final Generative Synthesis phase (Section 3.5).

These four signals—external and three internal—are consolidated by the extraction function $\phi(\cdot)$ into the feature vector $\mathbf{x}_j = [r_{ext}, r_{int}, s_{int}, u_{int}]^\top$. Representing complementary perspectives, this vector is subsequently forwarded to the Fuzzy Inference Engine for rule-constrained, non-linear aggregation and final fuzzy validity scoring.

3.3 The fuzzy inference system

The Fuzzy Inference Engine integrates heterogeneous retrieval signals into a unified, interpretable fuzzy validity score, s_j, enabling robust post-retrieval decision-making. The computation is executed through three sequential phases, as outlined in Algorithm 1.

3.3.1 Phase 1: Adaptive fuzzification

Crisp numerical inputs are mapped to linguistic terms $\mathcal{L} = \{\text{Low}, \text{Medium}, \text{High}\}$ to handle the inherent uncertainty in retrieval signals. A unified parametric representation enables MFs to adapt dynamically based on data characteristics. To provide structural flexibility, a shape parameter $\tau \in [0, 1]$ is introduced to determine whether the corresponding MF is instantiated as a triangular or a trapezoidal function. This shape parameter is optimized during the stochastic learning phase described in Section 3.4, where the SA algorithm learns the compact MF parameters controlling both the decision boundaries and the structural shape from data.

After optimization, the learned compact parameters are converted into the final breakpoint representation used during inference. Accordingly, each instantiated MF is represented by four boundary points (a, b, c, d), where a and d denote the support boundaries and b and c define the core region of the function.

The resulting generalized membership function is:

$$\mu(x; a, b, c, d) = \max\left(\min\left(\frac{x-a}{b-a},\, 1,\, \frac{d-x}{d-c}\right), 0\right). \tag{3}$$

Thus, the SA procedure operates on the tunable compact MF parameters, while the fuzzy inference engine evaluates the resulting MFs using their final (a, b, c, d) breakpoint form. The optimized parameter values are reported in Section 4.

3.3.2 Phase 2: Knowledge base and inference

The fuzzified inputs are evaluated against a predefined Mamdani-type rule base $\mathcal{R}$, summarized in Table 1. Rather than enumerating all possible input combinations, the rule base is designed around the main situations that arise in passage assessment: cases where relevance, support, and utility jointly indicate strong evidence; cases where only moderate or partial support is available; and cases where low support, low utility, or low external relevance call for a conservative revision. Rules R1–R4 promote high-quality evidence when strong positive signals align, distinguishing fully supported cases from those with strong but not uniformly maximal evidence. Rules R5–R6 capture intermediate situations in which only partial or moderate support is available. Rules R7–R9 implement conservative revision: low utility and low support reduce the evaluation directly, while low external relevance introduces an externally grounded adverse signal that can substantially depress the final score without acting as a formal hard veto under centroid defuzzification. Because the membership functions overlap, an input may partially activate multiple neighboring linguistic categories, and the final output is obtained through aggregated graded rule responses rather than exact one-to-one pattern matching. When rule activation is weak overall, the aggregated fuzzy output remains correspondingly conservative, preventing unsupported high-confidence promotion of weak evidence.

Table 1: Mamdani-type fuzzy rule base for evidence quality evaluation.

Rule	**Antecedent (IF)**	**Consequent (THEN)**
R1	r_{int} is High AND s_{int} is High AND u_{int} is High	Evaluation is Excellent
R2	r_{int} is High AND s_{int} is High	Evaluation is Good
R3	r_{int} is High AND u_{int} is High	Evaluation is Good
R4	s_{int} is High AND u_{int} is Medium	Evaluation is Good
R5	r_{int} is Medium AND s_{int} is Medium	Evaluation is Fair
R6	r_{int} is Medium AND r_{ext} is Medium	Evaluation is Fair
R7	u_{int} is Low	Evaluation is Marginal
R8	s_{int} is Low	Evaluation is Poor
R9	r_{ext} is Low	Evaluation is Poor

3.3.3 Phase 3: Defuzzification (centroid strategy).

The aggregated fuzzy output is converted into a crisp scalar fuzzy validity score $s_j \in [0,1]$ for document d_j using the Centroid, or center of gravity, defuzzification method:

$$s_j = \frac{\int_Y y \cdot \mu_{\text{agg}}(y)\, dy}{\int_Y \mu_{\text{agg}}(y)\, dy}, \tag{4}$$

where Y denotes the output universe of discourse, $y \in Y$ represents the output variable, and $\mu_{\text{agg}}(y)$ is the aggregated output membership function. In this work, rule antecedents are evaluated using a selected t-norm operator, while the activated consequent membership functions are combined through the aggregation operator to form $\mu_{\text{agg}}(y)$. The centroid method is selected for its stability and its ability to provide a balanced representation of potentially conflicting rules, often yielding smoother and more robust decisions than alternative defuzzification strategies. Accordingly, adverse rules such as low external relevance influence the output through the aggregated fuzzy surface rather than through an absolute override. The resulting continuous score serves as the validity metric for downstream selection and optimization.

The complete fuzzy inference pipeline—comprising fuzzification, rule evaluation, aggregation, and centroid defuzzification—is denoted as the parameterized scoring function:

$$\mathcal{F}_\Theta : [0,1]^4 \to [0,1], \qquad s_j = \mathcal{F}_\Theta(\mathbf{x}_j), \tag{5}$$

where $\mathbf{x}_j = \phi(q, d_j)$ is the feature vector of document d_j produced by the signal extraction operator $\phi(\cdot)$, and Θ collects all tunable parameters, including membership function boundaries and shape parameters. Throughout this paper, $\mathcal{F}_\Theta(\mathbf{x}_j)$ is used exclusively; the dependence on query q and document d_j enters through the feature vector $\mathbf{x}_j = \phi(q, d_j)$.

Algorithm 1 Fuzzy Inference Procedure

Require: Feature vector $\mathbf{x}_j = (r_{ext}, r_{int}, s_{int}, u_{int})^\top \in [0,1]^4$, parameter set Θ
Ensure: fuzzy validity score $s_j \in [0,1]$
Phase 1: Adaptive Fuzzification
Map inputs to $\{\mu_{\text{Low}}, \mu_{\text{Med}}, \mu_{\text{High}}\}$ using membership shapes defined by Θ
Phase 2: Rule Evaluation
$\mathcal{O}_{\text{agg}} \leftarrow \text{Aggregate}(\text{Rules}, \text{Mamdani})$ ▷ Apply implication and aggregate rule outputs
Phase 3: Defuzzification (Centroid)
$s_j \leftarrow \text{Centroid}(\mathcal{O}_{\text{agg}})$ ▷ Eq. (4)
return s_j

3.4 Adaptive fuzzy calibration and optimization via simulated annealing

Optimizing fuzzy membership function boundaries is challenging due to the non-differentiable and highly non-convex parameter space. SA is employed, a meta-heuristic that enables probabilistic acceptance of suboptimal solutions, facilitating escape from local optima.

3.4.1 Parameterization of membership functions

Each input dimension (external relevance, internal relevance, support, utility) is characterized by three linguistic terms: Low, Medium, and High. Each term is controlled by a *position parameter* that determines its boundary or center location in $[0,1]$, and a *shape parameter* $\tau \in [0,1]$ that governs its geometry. When $\tau < 0.5$, the membership function adopts a triangular shape; when $\tau \geq 0.5$, it becomes trapezoidal with a flat plateau.

The Low and High terms are modeled as shoulder-shaped functions with fixed padding of 0.15 to ensure smooth transitions at domain boundaries. This padding value was selected as a design parameter to balance boundary smoothness and discriminative power; alternative values may be explored for domain-specific tuning. For instance, the Low term is defined by parameters $(0, 0, a, a + 0.15)$, where a is the tunable upper boundary. Similarly, the High term uses $(b - 0.15, b, 1, 1)$, where b is the tunable lower boundary. The Medium term is centered around a tunable midpoint m, with its spread and shape controlled by τ_{med}.

Conversion from (m, τ) to membership function points. For the Medium term, the shape parameter τ_{med} determines the support width $w = 0.3 \times (1 + \tau_{\text{med}})$. If $\tau_{\text{med}} < 0.5$, a triangular function is constructed with vertices at $(m - w, m, m + w)$. If $\tau_{\text{med}} \geq 0.5$, a trapezoidal function is formed with core plateau $[m - 0.05, m + 0.05]$ and support $[m - w, m + w]$, yielding four points $(m - w, m - 0.05, m + 0.05, m + w)$. All coordinates are clipped to $[0, 1]$ to respect domain boundaries. This (m, τ) parametrization implicitly governs both the position and the shape of the membership functions, thereby avoiding direct optimization over all breakpoints (a, b, c, d) for each term.

The parameter vector for each dimension thus comprises six values:

$$\Theta_{\text{dim}} = \{a_{\text{low}}, \tau_{\text{low}}, m_{\text{med}}, \tau_{\text{med}}, b_{\text{high}}, \tau_{\text{high}}\}. \tag{6}$$

Across four input dimensions, the total parameter count is $4 \times 6 = 24$. The output dimension (fuzzy validity score) uses five linguistic terms (Poor, Marginal, Fair, Good, Excellent), each with a position parameter and a shape parameter. While the linguistic labels are fixed, their positions and shapes are optimized by SA to align with the empirical distribution of document quality in the calibration set, contributing an additional $5 \times 2 = 10$ parameters and yielding a total of 34 tunable parameters. This compact parametrization substantially reduces the degrees of freedom of the search space compared to independently tuning all membership breakpoints, which in turn stabilizes simulated annealing by lowering the risk of converging to pathological local optima (e.g., highly distorted or mutually crossing membership functions that violate the intended fuzzy partitioning structure).

3.4.2 Optimization objective

Let $\mathcal{D}_{\text{calib}} = \{(q_i, a_i^{\text{gt}})\}_{i=1}^{N}$ denote a calibration dataset with $N = 600$ query–answer pairs, where a_i^{gt} is the ground-truth answer for query q_i. For each query q_i, the system first retrieves an initial pool $\mathcal{D}_{\text{ret},i}$ of M documents. This retrieval is performed once per query and remains fixed throughout the SA optimization process. The system then extracts feature vectors $\mathbf{x}_j = \phi(q_i, d_j)$ for each document $d_j \in \mathcal{D}_{\text{ret},i}$, and simultaneously generates candidate answers $\mathcal{C}_{\text{ret},i} = \{c_j : d_j \in \mathcal{D}_{\text{ret},i}\}$ via the cached mapping $\{d_j \to c_j\}$ established during signal extraction (Section 3.2). The fuzzy inference engine then computes fuzzy validity scores $s_j = \mathcal{F}_\Theta(\mathbf{x}_j)$ (Eq. 5), selects the top-$k = 2$ documents by score to form $\mathcal{D}_{\text{top},i}$, retrieves their corresponding candidate answers $\mathcal{C}_{\text{top},i} = \{c_j : d_j \in \mathcal{D}_{\text{top},i}\}$, and synthesizes the final answer via Synthesize$(\cdot)$.

The optimization objective maximizes end-to-end accuracy:

$$\Theta^* = \arg\max_{\Theta} \sum_{(q_i, a_i^{\text{gt}}) \in \mathcal{D}_{\text{calib}}} \mathbb{I}\Big(\text{Synthesize}\big(q_i,\ \mathcal{D}_{\text{top},i},\ \mathcal{C}_{\text{top},i}\big) = a_i^{\text{gt}}\Big), \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the synthesized answer exactly matches the ground-truth answer a_i^{gt} and 0 otherwise. The operator $\mathcal{D}_{\text{top},i} = \text{Top}_k(\mathcal{D}_{\text{ret},i})$ denotes the top-$k = 2$ documents selected by fuzzy validity scores $\{s_j\}$, and $\mathcal{C}_{\text{top},i}$ represents the corresponding candidate answers retrieved from the cache. The dependence on Θ enters through the fuzzy scoring function $\mathcal{F}_\Theta$, which determines the ranking and thus the selection of $\mathcal{D}_{\text{top},i}$.

3.4.3 SA optimization procedure

Algorithm 2 describes the complete optimization procedure. Membership function parameters are initialized as follows: position parameters (a, m, b) are set manually based on domain expertise to semantically meaningful values, while shape parameters (τ) are randomly initialized from $\mathcal{U}(0, 1)$ to enable exploration of both triangular and trapezoidal forms.

At each iteration, candidate parameters are generated by adding Gaussian noise:

$$\Theta_{\text{new}} = \Theta_{\text{curr}} + \mathcal{N}(0, \sigma^2), \quad \sigma = 0.05. \tag{8}$$

All parameters are clipped to $[0, 1]$ after perturbation. Notably, shape parameters are also perturbed: since the Gaussian noise has zero mean, it can both increase and decrease τ, enabling smooth transitions between membership function geometries (triangular $\leftrightarrow$ trapezoidal) during optimization. When τ crosses the threshold of 0.5 due to perturbation, the corresponding membership function automatically switches between triangular and trapezoidal forms.

The Metropolis criterion determines acceptance: improvements ($\Delta J > 0$) are always accepted, while degradations ($\Delta J < 0$) are accepted with probability $P = \exp(\Delta J / T)$, where T is the current temperature. The cooling schedule begins at $T_0 = 1.0$, applies a cooling rate $\alpha = 0.90$ after each epoch, and terminates when T falls below $T_{\min} = 0.001$.

Algorithm 2 Simulated Annealing Optimization Procedure

Require: Calibration set $\mathcal{D}_{\text{calib}}$, number of top documents $k = 2$, initial temperature $T_0 = 1.0$, cooling rate $\alpha = 0.90$, minimum temperature $T_{\min} = 0.001$, perturbation scale $\sigma = 0.05$
Ensure: Optimized parameters Θ^*

Initialize position parameters (a, m, b) in Θ_{curr} manually
Initialize shape parameters (τ) in Θ_{curr} from $\mathcal{U}(0, 1)$ ▷ Random tri/trap
$\Theta^* \leftarrow \Theta_{\text{curr}}, \quad J_{\text{best}} \leftarrow 0, \quad J_{\text{curr}} \leftarrow 0$
$T \leftarrow T_0$
while $T > T_{\min}$ **do**
  $\Theta_{\text{new}} \leftarrow \Theta_{\text{curr}} + \mathcal{N}(0, \sigma^2)$ ▷ Perturb all parameters (position & shape)
  Clip all parameters in Θ_{new} to $[0, 1]$ ▷ Perturbation may cause τ to cross 0.5
  $N_{\text{correct}} \leftarrow 0$
  for each $(q_i, a_i^{\text{gt}}) \in \mathcal{D}_{\text{calib}}$ **do**
    for each $d_j \in \mathcal{D}_{\text{ret},i}$ **do**
      $\mathbf{x}_j \leftarrow \phi(q_i, d_j)$ ▷ Extract signals; c_j cached as $\{d_j \to c_j\}$
      $s_j \leftarrow \mathcal{F}_{\Theta_{\text{new}}}(\mathbf{x}_j)$ ▷ Compute fuzzy validity score
    end for
    $\mathcal{D}_{\text{top},i} \leftarrow \text{Top}_k(\{(d_j, s_j)\})$ ▷ Select top-$k = 2$ by score
    $\mathcal{C}_{\text{top},i} \leftarrow \{c_j : d_j \in \mathcal{D}_{\text{top},i}\}$ ▷ Retrieve from cache $\{d_j \to c_j\}$
    $\hat{a} \leftarrow \text{Synthesize}(q_i, \mathcal{D}_{\text{top},i}, \mathcal{C}_{\text{top},i})$
    if $\hat{a} = a_i^{\text{gt}}$ **then**
      $N_{\text{correct}} \leftarrow N_{\text{correct}} + 1$
    end if
  end for
  $J_{\text{new}} \leftarrow N_{\text{correct}} \,/\, |\mathcal{D}_{\text{calib}}|$
  $\Delta J \leftarrow J_{\text{new}} - J_{\text{curr}}$
  if $\Delta J > 0$ **or** $\text{Rand}(0, 1) < \exp(\Delta J / T)$ **then** ▷ Metropolis criterion
    $\Theta_{\text{curr}} \leftarrow \Theta_{\text{new}}, \quad J_{\text{curr}} \leftarrow J_{\text{new}}$
    if $J_{\text{new}} > J_{\text{best}}$ **then**
      $\Theta^* \leftarrow \Theta_{\text{new}}, \quad J_{\text{best}} \leftarrow J_{\text{new}}$
    end if
  end if
  $T \leftarrow T \times \alpha$ ▷ Exponential cooling
end while
return Θ^*

3.5 Dynamic generative synthesis and conflict resolution

After fuzzy inference and document ranking (Algorithm 2), the top-$k = 2$ documents $\mathcal{D}_{\text{top}} = \{d_1, d_2\}$ along with their pre-computed validity scores $s_1 \geq s_2$ and cached candidate answers $\mathcal{C}_{\text{top}} = \{c_1, c_2\}$ are forwarded to the Generative Synthesis module. The value $k = 2$ was determined empirically: restricting synthesis to the top-2 documents captures the majority of potential conflicts while maintaining efficient computational performance. Ablation studies validating this choice are presented in Section 4. The candidate answers, generated during the signal extraction phase (Section 3.2), are now utilized for semantic comparison and conflict resolution.

Selecting only the highest-scoring document can ignore potential contradictions or marginal score differences. To address this, a dynamic routing mechanism directs the query to either Direct Output (fast path) or Generative Synthesis (slow path) based on two quantitative criteria. The routing decision is performed dynamically at inference time for each query according to the fuzzy validity score disparity between top candidates and their semantic agreement. The score gap is computed as

$$\Delta s = s_1 - s_2, \tag{9}$$

which measures the decisiveness of the leading document. In parallel, semantic consensus is evaluated through cosine similarity between the sentence embeddings of the candidate answers:

$$\text{Sim}_{\text{txt}} = \text{CosineSim}(c_1, c_2). \tag{10}$$

The thresholds $\delta_{\text{ambiguity}} = 0.15$ and $\delta_{\text{consensus}} = 0.85$ were set empirically as fixed routing parameters. Here, $\delta_{\text{ambiguity}}$ defines the minimum score gap required to regard the top-ranked document as decisively preferred over the second-ranked one, while $\delta_{\text{consensus}}$ specifies the level of semantic agreement required between the candidate answers to permit direct output without synthesis. The selected values were chosen conservatively so that small score differences or weak semantic alignment still trigger synthesis, whereas direct output is reserved for cases with both clear score separation and strong semantic agreement.

As described in Algorithm 3, the system follows the fast path if the top candidate has a decisive advantage ($\Delta s > \delta_{\text{ambiguity}}$) and high semantic agreement ($\text{Sim}_{\text{txt}} > \delta_{\text{consensus}}$). Otherwise, the Conflict Resolution Module is activated. The Synthesize($\cdot$) function is formally defined as a deterministic prompt-based LLM call: it constructs a structured prompt containing the query, both documents, their candidate answers, and their fuzzy validity scores, then invokes the generator (e.g., `Llama-2-7b-SelfRAG`) for deterministic output. The fuzzy validity scores are explicitly incorporated into the prompt so that the LLM can prioritize higher-reliability evidence during synthesis.

Algorithm 3 Dynamic Generative Synthesis and Conflict Resolution

Require: Query q, top-2 documents $\mathcal{D}_{\text{top}} = \{d_1, d_2\}$ (pre-ranked by s_j), candidate answers $\mathcal{C}_{\text{top}} = \{c_1, c_2\}$, fuzzy validity scores $s_1 \geq s_2$
Ensure: Final answer a
$\Delta s \leftarrow s_1 - s_2$ ▷ fuzzy validity score gap
$\text{Sim}_{\text{txt}} \leftarrow \text{CosineSim}(c_1, c_2)$ ▷ Semantic consensus
if $\Delta s > \delta_{\text{ambiguity}}$ **and** $\text{Sim}_{\text{txt}} > \delta_{\text{consensus}}$ **then**
    Fast Path: $a \leftarrow c_1$ ▷ Direct output from top candidate
else
    Slow Path (Conflict Resolution):
    Construct prompt $P \leftarrow \text{Format}(q, d_1, d_2, c_1, c_2, s_1, s_2)$
    $a \leftarrow \text{LLM}(P)$ ▷ Synthesize via deterministic generation
end if
return a

3.6 Cognitive perspective

Allo-Self-RAG can be viewed through a dual-process-inspired perspective on decision-making [19]. In this sense, the comparison is intended as a functional analogy at the level of the overall retrieval-to-response pipeline.

From this perspective, the standard Self-RAG baseline is closer to a System-1-like post-retrieval evaluation style, relying on relatively heuristic first-pass critique signals, whereas Allo-Self-RAG introduces a more System-2-like mode of evidence assessment through structured signal integration, ambiguity-aware reassessment, and conditional allocation of additional computation.

The Fuzzy Inference System serves as an interpretable computational mechanism for rapid post-retrieval assessment under uncertainty. Unlike standard neural rerankers that rely solely on semantic similarity, this module aggregates heterogeneous signals ($r_{\text{ext}}, r_{\text{int}}, s_{\text{int}}, u_{\text{int}}$) through the scoring function $\mathcal{F}_\Theta(\mathbf{x}_j)$ to execute a structured multi-signal evaluation. This process enforces filtration through prioritization: by assigning low validity scores to ambiguous or unsupported documents, the system demotes weak candidates to the bottom of the ranking stack, effectively excluding them from the limited context window (top-k) fed to the generator. Simultaneously, the module applies rule-based revision–for example, reducing the final evaluation when factual support is absent or when external relevance is low–thereby acting as an interpretable quality-control layer before generation begins.

The Generative Synthesis module operates conditionally, triggered when the validity gap $\Delta s = s_1 - s_2$ is narrow or the semantic consensus Sim_{txt} is low, indicating that the post-retrieval assessment process cannot resolve the query with sufficient confidence. When the top-ranked candidate exhibits a decisive validity gap ($\Delta s > \delta_{\text{ambiguity}}$) and high semantic consensus ($\text{Sim}_{\text{txt}} > \delta_{\text{consensus}}$), the system infers low ambiguity and bypasses further processing to output the retrieved text directly. Conversely, when these conditions are not met, the system activates the LLM for Generative Synthesis. This design enables a deeper evaluation path only when needed, making the overall framework functionally reminiscent of dual-process accounts of fast versus more deliberate decision behavior.

4 Experimental results

The performance of Allo-Self-RAG is evaluated against competitive baselines across four benchmarks. The impact of retrieval augmentation strategies on accuracy is analyzed, with particular attention to domain-specific robustness and generalization to rare-entity queries.

4.1 Datasets and tasks

To assess the generalizability and robustness of Allo-Self-RAG, four diverse benchmarks were selected, representing distinct domains and reasoning patterns. All evaluations were conducted on the standard test splits to measure performance in diverse retrieval scenarios.

Fact Verification and Scientific Reasoning.

- **PubHealth** [23]: A domain-specific fact verification dataset consisting of expert-curated public health claims (e.g., "Vaccines cause autism"). The task requires classifying claims as *Supported*, *Refuted*, or *Not Enough Info* based on biomedical evidence. The standard test split is utilized to evaluate veracity prediction accuracy in a high-stakes medical context.
- **ARC-Challenge** [6]: A multiple-choice question-answering benchmark derived from grade-school science exams. The test set (N = 1,022) contains questions that require complex reasoning and cannot be solved by simple retrieval or co-occurrence statistics. This benchmark serves as a stress test for the framework's ability to handle multi-step scientific inference.

Open-Domain Question Answering.

- **PopQA** [35]: A large-scale open-domain benchmark comprising 14k factual queries. Following prior work, evaluation focuses on the long-tail subset (N = 1,249) of rare-entity questions (e.g., "What is the capital of Burkina Faso?"). This subset is selected to assess the system's capacity to retrieve external knowledge rather than relying on facts memorized during pre-training.
- **TriviaQA** [18]: A reading comprehension dataset requiring multi-hop reasoning over textual evidence. To rigorously evaluate the framework's capacity to synthesize answers from retrieved documents—rather than relying on parametric memory—a subset of 2,850 rare-entity questions was sampled. This filtering ensures that performance gains are attributable to the retrieval pipeline rather than the LLM's pre-training knowledge.

Calibration Protocol for Membership Function Tuning. To facilitate the fine-tuning of fuzzy membership functions described in Section 3, a balanced calibration set was constructed separately from the final test sets. This set consists of 600 data points, assembled via stratified sampling of 150 instances from each of the four datasets (PubHealth, ARC-Challenge, PopQA, TriviaQA). This balanced composition ensures that the SA algorithm optimizes the fuzzy membership functions to handle diverse noise profiles—ranging from scientific ambiguity to rare factual lookup—without overfitting to a single domain.

4.2 Baselines

To rigorously assess the contributions of the Allo-Self-RAG framework, three categories of baselines are considered.

1. No Retrieval (Closed-Book): Standard LLaMA2-7B [40] and Alpaca-7B [8] are evaluated without access to external documents. These baselines quantify the inherent parametric knowledge capacity of the base models and highlight their susceptibility to hallucinations in the absence of retrieved context.

2. Standard RAG: A conventional retrieval pipeline prepends the top-k retrieved passages to the prompt without any filtration or relevance assessment. This unfiltered augmentation establishes an upper bound on retrieval noise and demonstrates the risks of incorporating irrelevant evidence into generation.

3. Self-RAG-7B: As a state-of-the-art adaptive retrieval baseline [2], Self-RAG employs reflection tokens and linear score aggregation to dynamically rank passages. This baseline enables a direct evaluation of the advantages conferred by Allo-Self-RAG over conventional linear scoring.

4.3 Implementation details

Allo-Self-RAG is implemented as an inference-time enhancement, requiring no additional gradient updates or fine-tuning. The subjective reflection tokens (r_{int}, s_{int}, u_{int}) are generated via the selfrag_llama2_7b checkpoint, and the objective Allo-Critique signal (r_{ext}) is derived from the ms-marco-MiniLM-L-6-v2 Cross-Encoder.

All experiments are conducted on a dual NVIDIA T4 GPU node (15 GB VRAM each), with high-throughput inference supported by the vLLM engine alongside Hugging Face Transformers. For the consensus computation, cosine similarity between candidate answers, CosineSim(c_1, c_2), is computed using sentence embeddings from the all-MiniLM-L6-v2 model within the SentenceTransformers framework. Membership-function tuning via SA is performed on a stratified held-out calibration set of 600 samples (150 per dataset). This calibration set is kept separate from the examples used for final test evaluation and is not included in the reported test results. All final accuracies are computed only on non-calibration evaluation samples.

4.4 Membership function tuning via simulated annealing

To examine how the proposed calibration procedure reshapes the behavior of the fuzzy inference engine, Figures 2 and 2 compare the MFs of all linguistic variables before and after SA optimization on the calibration set. Specifically, the figures show the learned fuzzy partitions for *External Relevance* (r_{ext}), *Internal Relevance* (r_{int}), *Support* (s_{int}), *Utility* (u_{int}), and the output *fuzzy validity score*. Before tuning, all MFs are initialized as symmetric and uniformly spaced functions, reflecting a neutral and domain-agnostic prior. After optimization, however, the learned MFs become noticeably asymmetric, sharper, and more selective. This indicates that SA does not merely adjust numerical boundaries; rather, it reconfigures the semantic decision geometry of the fuzzy system to match the empirical reliability structure of the signals.

- **External Relevance Membership Functions:** The strongest redistribution is observed for *External Relevance*. After tuning, the *Low* MF extends substantially toward higher values, while the *High* MF is shifted rightward and activates later than in the initial configuration. At the same time, the *Medium* region collapses into a narrow band centered around a small transition zone. This pattern shows that the calibrated system has learned to treat a broad range of moderate cross-encoder scores as still insufficiently trustworthy, rather than prematurely interpreting them as strong evidence. In effect, SA makes the external relevance signal more conservative in its contribution to high-validity decisions, which is especially important in open-domain retrieval settings where superficially related but unreliable passages are common.

- **Internal Relevance Membership Functions:** In contrast to the external signal, the tuned MFs for *Internal Relevance* preserve a broader intermediate region, while still shifting the decision boundaries away from the original symmetric layout. The *Low* MF decays earlier, the *Medium* MF becomes left-centered, and the *High* MF activates only after sufficiently strong internal relevance is observed. This suggests that the optimizer learns a more graded treatment of the model's self-assessed relevance signal: moderate internal confidence is retained as informative, but it is not allowed to dominate the decision process too early. This behavior is desirable because internal relevance is useful but potentially biased, and therefore should contribute smoothly rather than act as a decisive signal on its own.

- **Support Membership Functions:** The tuned *Support* MFs exhibit one of the sharpest structural changes. The *Medium* region contracts into a narrow peak, while the *High* MF rises rapidly and the *Low* MF drops off quickly. This indicates that the support signal is treated in a near-polarized manner after calibration: evidence is learned to be either genuinely supportive or insufficient, with only a small margin for ambiguous middle cases. Such a sharpened partition is particularly meaningful for grounded generation, since support is closely tied to factual consistency. By compressing the uncertain middle zone, the fuzzy system reduces the risk that weakly grounded passages receive overly favorable evaluations.

- **Utility Membership Functions:** The tuned *Utility* MFs display a different behavior from *Support*. Here, the *Medium* region shifts toward higher values and remains comparatively broad, while the *High* MF activates much later than in the initial configuration. This suggests that the optimizer interprets utility as a more gradual and noisier signal than support or external relevance. In practical terms, the system becomes more conservative about labeling a passage as highly useful, while still preserving a meaningful intermediate range for partially useful evidence. This is consistent with the subjective nature of utility judgments, which often vary more smoothly than binary support relations.

- **Fuzzy Validity Score Membership Functions:** The output MFs also undergo a clear conservative recalibration. Compared with the pre-tuned configuration, the *Good* and *Excellent* categories are shifted rightward, while the intermediate categories become more concentrated. This means that higher-quality final labels are assigned only when stronger aggregated evidence is present. Importantly, this shows that SA reshapes not only the interpretation of the input signals, but also the semantics of the final decision space itself. As a result, the fuzzy engine becomes less permissive in promoting candidates into high-validity regions, which directly supports the framework's objective of reducing weakly grounded answer selection and limiting hallucination propagation.

Overall, the learned MF transformations reveal that fixed manually specified fuzzy partitions are insufficient for reliable post-retrieval evaluation. The optimized shapes are clearly signal-specific: external relevance contributes more conservatively to high-validity decisions, support becomes more decisive and polarized, utility remains graded but more selective, and the output validity scale raises the bar for strong acceptance decisions. These changes demonstrate that SA serves as more than a numerical tuning procedure; it performs an interpretable semantic calibration of the fuzzy decision boundaries. This adaptive restructuring is a key component of the proposed framework, as it enables the neuro-symbolic evaluator to align its reasoning behavior with the empirical uncertainty patterns of real retrieval outputs. The downstream impact of this calibration is further supported by the ablation study in Section 4.7.

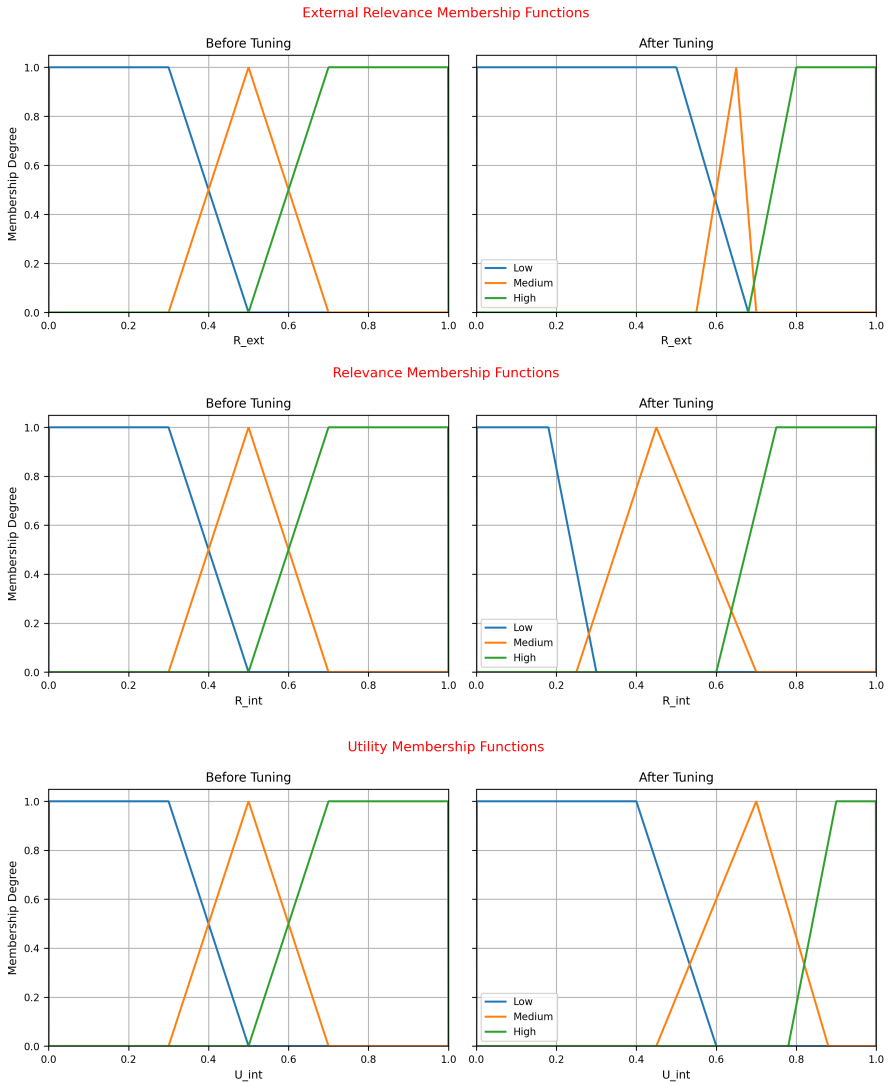

Figure 2: Membership functions before (left) and after (right) SA optimization (Part 1).

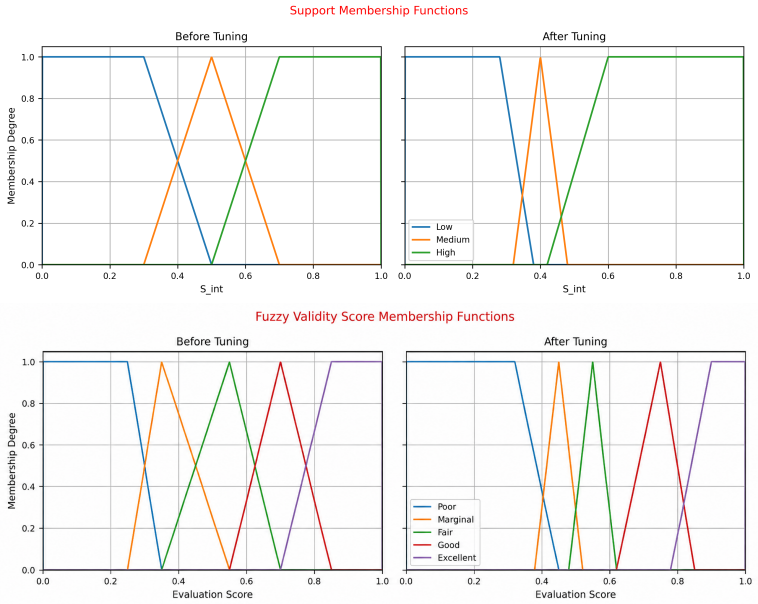

Figure 2: Membership functions before (left) and after (right) SA optimization (Part 2).

4.5 Case study: Interpretable fuzzy inference examples

While benchmark-level results quantify the overall effectiveness of the proposed framework, they provide limited visibility into how the fuzzy inference engine translates heterogeneous critique signals into a final validity score. To make this decision process more transparent, we present three representative case studies that illustrate how crisp inputs are fuzzified, which rules are fired, and how competing rule consequents are aggregated through Mamdani inference and centroid defuzzification. The selected examples cover three qualitatively distinct retrieval scenarios: *(i)* internally strong but externally weak evidence, *(ii)* full agreement among all signals, and *(iii)* relevance without sufficient utility. For clarity, each figure shows only the fired rules, i.e., rules with non-zero activation strength.

Table 2: Representative crisp input configurations and their resulting fuzzy validity scores.

Case	r_{int}	s_{int}	u_{int}	r_{ext}	**Fuzzy Validity Score**
C1	0.80	0.75	0.55	0.30	0.3258
C2	0.92	0.88	0.90	0.85	0.8433
C3	0.70	0.65	0.20	0.60	0.4070

Interpretive perspective. The goal of the following examples is not to restate the rule base introduced earlier, but to show how it behaves under different signal configurations. In particular, the cases highlight how positive and revising rules interact during inference, how some consequents dominate others under conflict, and how these interactions shape the final defuzzified validity score.

Case C1: Strong internal confidence revised by weak external evidence. Case C1 corresponds to a conflict scenario in which the internal signals appear favorable ($r_{int} = 0.80$, $s_{int} = 0.75$, $u_{int} = 0.55$), but the external relevance signal is clearly weak ($r_{ext} = 0.30$). As shown in Figure 3, positive rules are indeed activated: Rule 2 fires at strength 1.00 and Rule 4 contributes an additional *Good*-oriented activation with strength 0.40. However, this positive evidence is counterbalanced by two revision mechanisms. First, Rule 7 activates a *Marginal* consequent with strength 0.25. More importantly, the external revision rule, Rule 9, fires at full strength 1.00 and contributes a dominant *Poor* output region.

This example shows that high internal confidence alone is not sufficient to obtain a high final score. Even though multiple rules support a favorable interpretation, the fuzzy engine treats weak external relevance as a strong reliability warning and shifts the aggregated output toward lower validity regions. The final score of 0.3258 therefore reflects a

deliberate conservative bias: the system down-ranks evidence that appears self-consistent internally but lacks independent external corroboration. This behavior is especially important in retrieval settings where internally plausible but weakly grounded evidence may otherwise reinforce hallucination-prone reasoning.

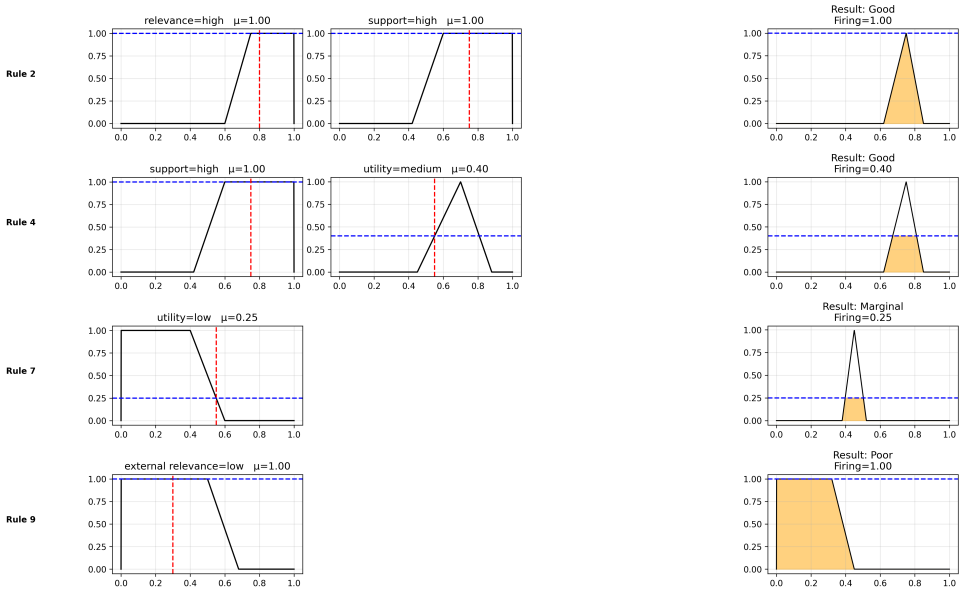

Figure 3: Rule-level activation and aggregation for Case C1. Despite strong internal signals, the external revision associated with low external relevance substantially suppresses the final score. Only fired rules are visualized.

Case C2: Coherent high-confidence evidence across all signals. Case C2 represents the most favorable configuration, where all four inputs are high ($r_{int} = 0.92$, $s_{int} = 0.88$, $u_{int} = 0.90$, $r_{ext} = 0.85$). Under this condition, the rule base behaves oppositely to Case C1: instead of suppressing the decision, it strongly reinforces it. As illustrated in Figure 4, the primary high-confidence rule, Rule 1, fires at strength 1.00 and produces an *Excellent* consequent. Additional supportive rules, Rules 2 and 3, also fire at strength 1.00, contributing strong *Good* regions that broaden the high-quality portion of the aggregated output.

The resulting fuzzy validity score of 0.8433 confirms that the inference engine is not merely restrictive; rather, it is selectively conservative. When internal reasoning signals and external validation are simultaneously strong, the system decisively promotes the candidate into the upper validity range. This is an important property of the framework, since an effective post-retrieval evaluator should not only reject weakly grounded evidence, but also confidently preserve high-quality evidence when cross-signal agreement is present.

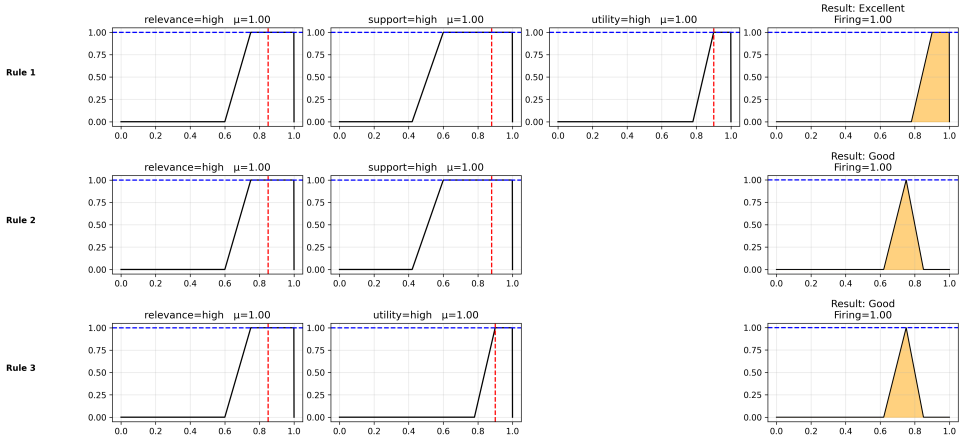

Figure 4: Fully aligned high-confidence evidence in Case C2. Rule 1 produces an *Excellent* consequent, while Rules 2 and 3 reinforce the high-scoring decision.

Case C3: Utility-driven demotion under partial relevance support. Case C3 is more subtle and demonstrates the importance of utility as a distinct decision factor. Here, internal relevance and support remain moderately strong ($r_{int} = 0.70$, $s_{int} = 0.65$), and the external signal is not fully weak ($r_{ext} = 0.60$). Nevertheless, the utility signal is very low ($u_{int} = 0.20$), leading to a final score of only 0.4070. As shown in Figure 5, Rule 2 still contributes a *Good* consequent with firing strength 0.67, indicating that the candidate retains some favorable properties. However, Rule 7, which revises low utility, fires at full strength 1.00 and introduces a dominant *Marginal* region. At the same time, residual low external relevance remains partially active, causing Rule 9 to contribute an additional *Poor* component with firing strength 0.44.

This case highlights an important aspect of the fuzzy design: a passage can be relevant and even partially supported, yet still be demoted if it lacks practical usefulness for answer construction. In other words, the inference engine distinguishes between *semantic relatedness* and *decision value*. The relatively modest final score is therefore not explained by a single failing signal, but by the non-linear aggregation of partial positive evidence with stronger utility- and reliability-based revision effects. This behavior is difficult to reproduce with simple additive score fusion and illustrates the advantage of a rule-based revision mechanism in ambiguous retrieval conditions.

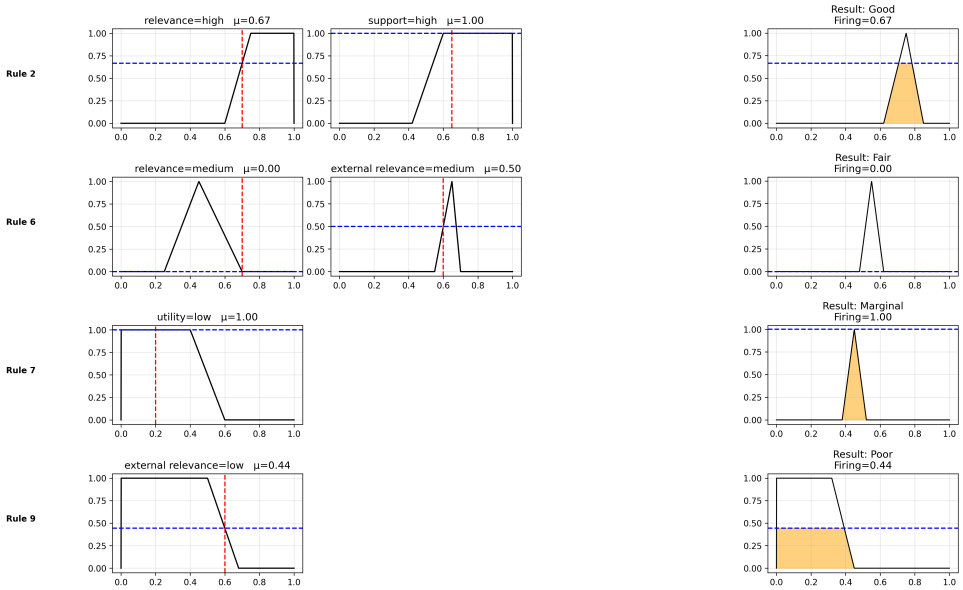

Figure 5: Rule activation pattern for Case C3. Although relevance and support remain moderately strong, low utility and residual external unreliability suppress the final score.

Overall, these case studies show that the proposed fuzzy inference engine functions as an interpretable semantic discriminator rather than a black-box scorer. The examples reveal three complementary behaviors: *(i)* strong internal confidence can be revised by weak external evidence, *(ii)* cross-signal agreement is rewarded with decisively high validity outputs, and *(iii)* utility acts as a meaningful corrective factor even when relevance remains moderately strong. Taken together, these behaviors show that the fuzzy layer performs structured conflict resolution over heterogeneous critique signals, making the final document selection process both more transparent and more robust. This rule-level interpretability complements the benchmark results by showing *why* the proposed framework improves post-retrieval evaluation, not merely *that* it does.

4.6 Main performance comparison

Table 3 reports the comparative evaluation on PopQA, TriviaQA, ARC, and PubHealth.

4.6.1 Evaluation methodology

The performance of the proposed framework and baselines is evaluated using a binary matching criterion. Under this protocol, a generated prediction is classified as correct if it corresponds to at least one of the provided ground-truth answers. This approach explicitly accounts for instances where multiple valid responses exist. The final accuracy is computed as the ratio of correctly matched predictions to the total number of evaluated samples.

Table 3: Accuracy comparison of different methods across PopQA, TriviaQA, ARC, and PubHealth. Results are reported as the average of 5 independent runs.

Category	Model	PopQA	TriviaQA	ARC	PubHealth
Without Retrieval	Alpaca 7B	23.9	54.8	45.0	49.3
Without Retrieval	LLaMA2 7B	15.1	30.5	21.4	33.7
With Retrieval	Alpaca 7B + Retrieval	47.1	64.1	47.5	39.7
With Retrieval	LLaMA2 7B + Retrieval	38.5	42.5	47.5	29.4
With Retrieval	SELF-RAG 7B	55.16	**70.24**	65.95	66.50
With Retrieval	**Allo-Self-RAG (Ours)**	**56.61**	68.83	**66.98**	**67.51**

4.6.2 Analysis.

Several observations emerge from Table 3. First, closed-book models show limited performance on most knowledge-intensive benchmarks, especially on PopQA and ARC, confirming the importance of incorporating external evidence. For example, adding retrieval improves Alpaca 7B from 23.9 to 47.1 on PopQA and LLaMA2 7B from 21.4 to 47.5 on ARC. However, naive retrieval is not uniformly beneficial. On PubHealth, retrieval decreases the accuracy of Alpaca 7B from 49.3 to 39.7 and LLaMA2 7B from 33.7 to 29.4. This degradation suggests that unfiltered retrieved evidence can introduce noisy, misleading, or weakly grounded context, particularly in domains where factual reliability is critical.

Second, compared with the strong SELF-RAG baseline, Allo-Self-RAG achieves higher accuracy on three of the four benchmarks: PopQA (56.61 vs. 55.16, +1.45), ARC (66.98 vs. 65.95, +1.03), and PubHealth (67.51 vs. 66.50, +1.01). These gains indicate that the proposed framework is effective not only in open-domain factual question answering, as reflected by PopQA, but also in reasoning-oriented and verification-sensitive settings such as ARC and PubHealth. This pattern is consistent with the intended role of the proposed critique-and-fusion mechanism, which aims to filter weak or contradictory evidence and promote retrieved content that is both internally plausible and externally grounded.

A different trade-off appears on TriviaQA, where SELF-RAG remains slightly stronger (70.24 vs. 68.83). One plausible explanation is that the more conservative fuzzy gating strategy in Allo-Self-RAG may occasionally suppress marginally useful retrieved context in relatively simpler fact-retrieval settings. In particular, the external revision induced by Rule 9 can down-rank evidence whose external relevance is insufficient, even if it may still contain partially useful information. This behavior introduces a modest recall-oriented trade-off, but it is aligned with the framework's broader objective of favoring reliability and stronger grounding in verification-heavy scenarios.

Overall, the results indicate that Allo-Self-RAG provides a favorable balance between retrieval utilization and post-retrieval evidence control. The framework outperforms SELF-RAG on three out of four datasets and achieves the best results on PopQA, ARC, and PubHealth. In particular, the improvement on PubHealth is notable because naive retrieval substantially degrades performance on this dataset, whereas Allo-Self-RAG achieves the highest accuracy. This suggests that explicitly modeling relevance, support, utility, and external relevance can improve answer selection beyond naive retrieval and beyond a strong retrieval-aware baseline.

4.7 Ablation study

To rigorously quantify the contribution of the principal components of Allo-Self-RAG, an ablation study is performed on TriviaQA and ARC. The analysis focuses on three modules that define the core novelty of the framework: (1) the external relevance estimation stream (*Allo-Critique*), (2) the simulated-annealing-based adaptive fuzzy calibration mechanism, and (3) the Dynamic Generative Synthesis module for ambiguity-aware conflict resolution. For each ablation, one component is removed from the full system, and the resulting change in end-to-end answer accuracy is measured. The results are reported in Table 4.

Table 4: Ablation results on TriviaQA and ARC. "Drop" denotes the absolute decrease in accuracy (percentage points) relative to the full Allo-Self-RAG configuration.

Setting	**TriviaQA**		**ARC**		**Primary Function**
	Acc.	**Drop**	**Acc.**	**Drop**	
Allo-Self-RAG (Full)	**68.83**	–	**66.98**	–	Integrated neuro-symbolic framework
w/o Dynamic Gen. Synthesis	67.14	-1.69	66.63	-0.35	Ambiguity-aware conflict resolution
w/o SA Calibration	63.86	-4.97	66.46	-0.52	Adaptive fuzzy partition learning
w/o External Reranker	63.47	-5.36	65.36	-1.62	Independent external signal

Overall, the ablation results reveal a clear hierarchy of component importance. The largest degradations arise when removing the external relevance estimation stream or the SA-based fuzzy calibration mechanism, showing that the main gains of Allo-Self-RAG stem from *better evidence discrimination* and *adaptive decision calibration*. Dynamic Generative Synthesis contributes a smaller but consistent improvement, indicating that it serves as a targeted refinement layer for ambiguous cases.

4.7.1 Impact of external relevance estimation (Allo-Critique).

The most pronounced performance drop is observed when the External Reranker is removed, particularly on TriviaQA, where accuracy decreases by 5.36 points. In this ablation, the framework loses the independent Cross-Encoder-based external relevance signal r_{ext} and must rely primarily on the generator's internal critique signals $(r_{int}, s_{int}, u_{int})$. The resulting degradation highlights one of the central design contributions of Allo-Self-RAG: document evaluation should not be based solely on self-generated confidence cues. By introducing an external grounding signal that is independent of the generator, Allo-Self-RAG reduces the risk that internally plausible but weakly grounded evidence will dominate the ranking process. The consistent drop across both datasets therefore confirms that Allo-Critique is not a peripheral addition, but a structurally important component of the framework's reliability.

4.7.2 Impact of SA-based adaptive fuzzy calibration.

Removing SA-based calibration also causes a major decline, especially on TriviaQA, where accuracy falls by 4.97 points. Under this ablation, the system reverts from optimized membership functions to fixed heuristic partitions, thereby losing its ability to adapt fuzzy decision boundaries to the empirical distribution of critique signals. This result strongly supports the value of the proposed adaptive fuzzy calibration strategy. Rather than relying on manually fixed partitions, Allo-Self-RAG learns data-sensitive membership structures that better reflect the noise profile and score variability of the target domain. The effect is especially strong on TriviaQA, suggesting that adaptive fuzzy partitioning is particularly beneficial in open-domain settings where retrieval quality and evidence consistency fluctuate substantially across examples. The much smaller drop on ARC further suggests that the value of calibration is most visible in environments with higher signal variability, where rigid heuristic boundaries are less effective.

4.7.3 Impact of dynamic generative synthesis.

The removal of Dynamic Generative Synthesis leads to a smaller but still consistent loss, especially on TriviaQA (-1.69 points). In this setting, the system can no longer route ambiguous cases to the slow path for conflict resolution and must instead rely on direct output from the top-ranked candidate. This result clarifies the role of the synthesis module within the overall architecture. Its primary contribution is not to replace the fuzzy ranking mechanism, but to complement it when the top candidates are close in fuzzy validity score or semantically inconsistent. In such cases, the synthesis stage acts as a final reconciliation layer that integrates partially complementary or conflicting evidence into a more coherent answer. The relatively smaller drop therefore reinforces an important point: the main strength of Allo-Self-RAG comes from its upstream neuro-symbolic evidence evaluation, while Dynamic Generative Synthesis provides an additional accuracy gain precisely where ambiguity persists after ranking. Taken together, the ablation study demonstrates that the effectiveness of Allo-Self-RAG does not arise from any single isolated component, but from the coordinated interaction of external grounding, adaptive fuzzy calibration, and ambiguity-aware synthesis. Among these, the two most influential factors are the independent external relevance signal and the learned fuzzy decision boundaries, which together constitute the core methodological contribution of the proposed framework.

5 Conclusion

This paper presented Allo-Self-RAG, a neuro-symbolic framework that rethinks post-retrieval verification in retrieval-augmented generation through adaptive fuzzy reasoning and externally grounded critique. The approach was motivated by three structural weaknesses of existing Self-RAG-style pipelines: self-confirmation bias when the same model both generates and evaluates, compensatory linear aggregation of heterogeneous critique signals, and rigid threshold-based decisions over inherently uncertain linguistic judgments. Allo-Self-RAG addresses these limitations by replacing static score combination with an interpretable fuzzy inference process that supports uncertainty-aware, logically constrained, and rule-guided evidence evaluation.

Methodologically, the framework integrates three complementary innovations: *(i)* Allo-Critique, which introduces an independent external relevance signal to complement internal self-critique and reduce self-reinforcing evaluation bias; *(ii)* SA-based adaptive fuzzy calibration, which learns dataset-sensitive membership functions instead of relying on fixed heuristic partitions; and *(iii)* Dynamic Generative Synthesis, which provides an ambiguity-aware extended evaluation path when the top-ranked candidates cannot be cleanly separated. Together, these components establish a principled alternative to weighted-sum ranking by allowing fuzzy rule-based interactions among relevance, support, and utility, including rule-based revision effects in cases where contradictory evidence should not be overly offset by strength in other dimensions.

Empirical evaluation on PopQA, TriviaQA, ARC-Challenge, and PubHealth shows that Allo-Self-RAG achieves strong performance and frequently improves upon the Self-RAG baseline, particularly in settings where retrieved evidence is noisy, partially conflicting, or uneven in quality. The ablation study further shows that the external relevance signal and adaptive fuzzy calibration are the two most influential contributors, while Dynamic Generative Synthesis provides a smaller but consistent refinement gain for ambiguous cases. Taken together, these results indicate that a substantial part of RAG reliability depends not only on what is retrieved, but on how retrieved evidence is evaluated, calibrated, and selectively trusted after retrieval.

More broadly, the significance of this work lies in showing that post-retrieval verification should not be treated as a simple scoring or reranking step. Instead, it can be formulated as a structured decision problem in which neural signals are governed by interpretable fuzzy rules and external grounding constraints. In this sense, Allo-Self-RAG is more than a task-specific modification of Self-RAG: it provides a concrete neuro-symbolic alternative for building more transparent and controllable RAG pipelines.

Limitations and future work

The current implementation introduces additional latency through the Cross-Encoder-based external reranking stage and, when triggered, the synthesis module. Although the fuzzy inference layer itself is lightweight, end-to-end efficiency remains influenced by these upstream components. In addition, the present framework relies on a manually specified fuzzy rule base with learned calibration, which preserves interpretability but may limit flexibility in some specialized settings.

Several natural extensions could be explored in future work. One possible direction is to apply fuzzy decision mechanisms to broader control problems in LLM systems, such as routing, response calibration, or selective verification. Another is to adapt the framework to more complex retrieval settings, including multi-document or multi-hop reasoning, where evidence interactions become more structured and conflict resolution more critical. It may also be useful to incorporate additional signals—such as source reliability, citation consistency, or explicit uncertainty estimates—while preserving the interpretability of the rule-based layer. More generally, this work suggests that adaptive fuzzy reasoning may offer a useful middle ground between fully first-pass heuristic control and fully opaque end-to-end scoring in RAG systems.

In summary, Allo-Self-RAG demonstrates that externally grounded critique, adaptive fuzzy calibration, and ambiguity-aware synthesis can be unified into a coherent framework for reliable post-retrieval verification. More broadly, the results suggest that fuzzy reasoning is not merely an auxiliary component, but a viable design principle for improving the reliability, controllability, and transparency of LLM- and RAG-based systems.

References

[1] F. Abdolinejad, M. Eftekhari, *Augmenting RAG with nonnegative matrix factorization-driven semantic chunking in embedding space*, The Journal of Supercomputing, **82** (2026), 224. https://doi.org/10.1007/s11227-026-08370-3

[2] A. Asai, Z. Wu, et al., *Self-RAG: Learning to retrieve, generate, and critique through self-reflection*, arXiv, (2023). https://arxiv.org/abs/2310.11511

[3] N. A. Birur, T. Baswa, et al., *VERA: Validation and enhancement for retrieval augmented systems*, arXiv, (2024). https://arxiv.org/abs/2409.15364

[4] W. Cai, J. Jiang, et al., *A survey on mixture of experts in large language models*, IEEE Transactions on Knowledge and Data Engineering, **37**(7) (2025), 3896-3915. https://doi.org/10.1109/TKDE.2025.3554028

[5] P. Chen, X. Liu, et al., *Fuzzy reasoning chain (FRC): An innovative reasoning framework from fuzziness to clarity*, Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, (2025), 10230-10240. https://doi.org/10.18653/v1/2025.findings-emnlp.541

[6] P. Clark, et al., *Think you have solved question answering? Try ARC, the AI2 reasoning challenge*, arXiv, (2018). https://arxiv.org/abs/1803.05457

[7] J. Deng, Y. Shen, et al., *Influence guided context selection for effective retrieval-augmented generation*, arXiv, (2025). https://arxiv.org/abs/2509.21359

[8] Y. Dubois, et al., *AlpacaFarm: A simulation framework for methods that learn from human feedback*, arXiv, (2024). https://arxiv.org/abs/2305.14387

[9] M. Eftekhari, A. Mehrpooya, et al., *How fuzzy concepts contribute to machine learning*, Springer, 2022. https://doi.org/10.1007/978-3-030-94066-9

[10] L. Gao, X. Ma, J. Lin, J. Callan, *Precise zero-shot dense retrieval without relevance labels*, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, (2023), 1762-1777. https://doi.org/10.18653/v1/2023.acl-long.99

[11] A. Garcez, L. C. Lamb, *Neurosymbolic AI: The 3rd wave*, Artificial Intelligence Review, **56** (2023), 12387-12406. https://doi.org/10.1007/s10462-023-10448-w

[12] F. Hosseini, M. Eftekhari, *PFE-SELF-RAG: Balancing self-RAG evaluation metrics via Pareto efficiency*, Journal of Mahani Mathematical Research, (2026), 179-208. https://doi.org/10.22103/jmmr.2026.25661.1841

[13] Y. Huang, J. Xiangji Huang, *A survey on retrieval-augmented text generation for large language models*, ACM Computing Surveys, **58**(12) (2026). https://doi.org/10.1145/3805774

[14] G. Izacard, M. Caron, et al., *Unsupervised dense information retrieval with contrastive learning*, Transactions on Machine Learning Research, (2022). http://dblp.uni-trier.de/db/journals/tmlr/tmlr2022.html#IzacardCHRBJG22

[15] S. Jeong, J. Baek, et al., *Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity*, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), (2024), 7036-7050. https://doi.org/10.18653/v1/2024.naacl-long.389

[16] Z. Ji, N. Lee, et al., *Survey of hallucination in natural language generation*, ACM Computing Surveys, **55**(12) (2023), 1-38. https://doi.org/10.1145/3571730

[17] Z. Jiang, F. Xu, et al., *Active retrieval augmented generation*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, (2023), 7969-7992. https://doi.org/10.18653/v1/2023.emnlp-main.495

[18] M. Joshi, E. Choi, D. Weld, L. Zettlemoyer, *TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension*, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, (2017), 1601-1611. https://doi.org/10.18653/v1/P17-1147

[19] D. Kahneman, *Thinking, fast and slow*, Macmillan, 2011.

[20] V. Karpukhin, B. Oğuz, et al., *Dense passage retrieval for open-domain question answering*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, (2020), 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[21] O. Khattab, M. Zaharia, *ColBERT: Efficient and effective passage search via contextualized late interaction over BERT*, SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, (2020), 39-48. https://doi.org/10.1145/3397271.3401075

[22] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Optimization by simulated annealing*, Science, **220**(4598) (1983), 671-680.

[23] N. Kotonya, F. Toni, *Explainable automated fact-checking for public health claims*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, (2020), 7740-7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

[24] J. Lesatod, J. Rivera, et al., *An adaptive compute approach to optimize inference efficiency in large language models*, Wiley, 2024. https://doi.org/10.22541/au.172851214.47069639/v1

[25] P. Lewis, E. Perez, et al., *Retrieval-augmented generation for knowledge-intensive NLP tasks*, arXiv, (2021). https://arxiv.org/abs/2005.11401

[26] E. Liu, et al., *Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs*, arXiv, (2024). https://doi.org/10.48550/arXiv.2407.00945

[27] N. F. Liu, K. Lin, et al., *Lost in the middle: How language models use long contexts*, Transactions of the Association for Computational Linguistics, **12** (2024), 153-173. https://doi.org/10.1162/tacl_a_00638

[28] J. Liu, P. Tang, et al., *A survey on inference optimization techniques for mixture of experts models*, ACM Computing Surveys, **58**(10) (2026), 1-37. https://doi.org/10.1145/3794845

[29] X. Lyu, S. Grafberger, S. Biegel, et al., *Improving retrieval-augmented large language models via data importance learning*, arXiv, (2023). https://arxiv.org/abs/2307.03027

[30] N. Masoumi, O. Davar, M. Eftekhari, *MG-CRAG: Fusion of multi-granular retrieval evaluators in corrective RAG with weakly supervised fine-tuning*, Knowledge and Information Systems, **68**(1) (2026), 149. https://doi.org/10.1007/s10115-026-02778-2

[31] S. Mishra, et al., *From facts to conclusions: Integrating deductive reasoning in retrieval-augmented LLMs*, arXiv, (2025). https://arxiv.org/abs/2512.16795

[32] R. Nogueira, K. Cho, *Passage re-ranking with BERT*, arXiv, (2020). https://arxiv.org/abs/1901.04085

[33] B. Pan, Y. Shen, et al., *Dense training, sparse inference: Rethinking training of mixture-of-experts language mModels*, arXiv, (2024). https://arxiv.org/abs/2404.05567

[34] O. Press, M. Zhang, et al., *Measuring and narrowing the compositionality gap in language models*, Conference: Findings of the Association for Computational Linguistics: EMNLP, (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.378

[35] A. Rogers, J. Boyd-Graber, N. Okazaki, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, (2023). https://aclanthology.org/2023.acl-long.0/

[36] W. Shi, S. Min, et al., *REPLUG: Retrieval-augmented black-box language models*, arXiv, (2023). https://arxiv.org/abs/2301.12652

[37] N. Shinn, F. Cassano, et al., *Reflexion: Language agents with verbal reinforcement learning*, NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, (2023), 8634-8652.

[38] W. Sun, L. Yan, et al., *Is ChatGPT good at search? Investigating large language models as re-ranking agents*, arXiv, (2024). https://arxiv.org/abs/2304.09542

[39] P. Tamhankar, N. R. Patel, M. C. Kolla, *MultiRAG: A fuzzy logic-driven multi-granularity framework for legal document generation*, 2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI), (2025), 313-318. https://doi.org/10.1109/IRI66576.2025.00065

[40] H. Touvron, et al., *Llama 2: Open foundation and fine-tuned chat models*, arXiv, (2023). https://arxiv.org/abs/2307.09288

[41] H. Trivedi, N. Balasubramanian, et al., *Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions*, arXiv, (2023). https://arxiv.org/abs/2212.10509

[42] H. Wang, A. Prasad, et al., *Retrieval-augmented generation with conflicting evidence*, arXiv, (2025). https://arxiv.org/abs/2504.13079

[43] H. Wang, L. Ren, T. Zhao, L. Jiao, *CoLLM: Industrial large–small model collaboration with fuzzy decision-making agent and self-reflection*, IEEE Transactions on Fuzzy Systems, **34**(4) (2026). https://doi.org/10.1109/TFUZZ.2025.3594229

[44] F. Wang, X. Wan, et al., *Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models*, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, (2025), 30553-30571. https://doi.org/10.18653/v1/2025.acl-long.1476

[45] Z. Wang, Z. Wang, et al., *Speculative RAG: Enhancing retrieval augmented generation through drafting*, arXiv, (2025). https://arxiv.org/abs/2407.08223

[46] X. Wang , J. Wei, et al., *Self-consistency improves chain of thought reasoning in language models*, arXiv, (2023). https://arxiv.org/abs/2203.11171

[47] S. Xie, T. Yang, et al., *LLM-driven multimodal knowledge graph construction for industrial process with prompt optimization and fuzzy RAG*, IEEE Transactions on Fuzzy Systems, **99** (2026), 1-14. https://doi.org/10.1109/TFUZZ.2026.3665172

[48] F. Xu, W. Shi, E. Choi, *RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation*, arXiv, (2023). https://arxiv.org/abs/2310.04408

[49] F. Xue, Z. Zheng, et al., *Openmoe: An early effort on open mixture-of-experts language models*, ICML'24: Proceedings of the 41st International Conference on Machine Learning, (2024), 55625-55655.

[50] S. Q. Yan, J. C. Gu, et al., *Corrective retrieval augmented generation*, arXiv, (2024). https://arxiv.org/abs/2401.15884

[51] T. Yao, et al., *Multiagent fuzzy reinforcement learning with LLM for cooperative navigation of endovascular robotics*, IEEE Transactions on Fuzzy Systems, **34** (2026), 1109-1119. https://doi.org/10.1109/TFUZZ.2025.3585934

[52] Y. Yu, et al., *RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs*, arXiv, (2024). https://arxiv.org/abs/2407.02485

[53] W. Yu, H. Zhang, et al., *Chain-of-Note: Enhancing robustness in retrieval-augmented language models*, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, (2024), 14672-14685. https://doi.org/10.18653/v1/2024.emnlp-main.813

[54] D. Zhang, J. Song, et al., *Mixture of experts in large language models*, arXiv, (2025). https://doi.org/10.48550/arXiv.2507.11181

[55] H. Zhuang, et al., *RankT5: Fine-tuning T5 for text ranking with ranking losses*, SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, (2022), 2308-2313. https://doi.org/10.1145/3539618.3592047