Iranian Journal of Fuzzy Systems

Volume 23, Number 5, (2026), pp. 107-127

Original paper

University of Sistan and Baluchestan

Fuzzy logic in large language models: Concepts, applications, and challenges

M. B. Dowlatshahi [1] and S. Beiranvand [2]

[1] *Department of Computer Engineering, Faculty of Engineering, Lorestan University, Khoramabad, Iran*
[2] *Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran.*

dowlatshahi.mb@lu.ac.ir, Sbeyranvand@tvu.ac.ir

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation; however, they continue to suffer from critical limitations, including predictive uncertainty, factual hallucination, and lack of interpretability. Fuzzy logic, with its inherent capacity to model graded membership and reason under imprecision, has emerged as a promising complementary paradigm to mitigate these shortcomings. Despite the rapidly growing body of literature, a systematic review that comprehensively synthesizes integration methodologies, empirical gains, and deployment challenges in fuzzy-LLM systems remains absent. This study conducts a systematic literature review of 70 peer-reviewed studies published between November 2022 and June 2026, organized around five research questions: (1) theoretical frameworks for fuzzy-LLM integration, (2) improvements in reasoning performance, (3) hallucination mitigation mechanisms, (4) fuzzy evaluation metrics, and (5) practical implementation obstacles. Our analysis reveals that fuzzy rule-based systems dominate the integration landscape, comprising 40 studies (57.1%), whereas fuzzy attention mechanisms remain critically underexplored, with only 3 studies (4.3%). Uncertainty reasoning constitutes the predominant application domain, addressed in 54 studies (77.1%), while causal reasoning remains entirely uninvestigated. Empirical findings indicate that fuzzy mechanisms yield measurable accuracy improvements between 1.5% and 3.0%, alongside substantial hallucination reduction. Among mitigation strategies, fuzzy refinement emerges as both the most frequently adopted and effective approach, reported in 19 studies (27.1%). Nevertheless, large-scale deployment continues to face persistent barriers, particularly computational overhead and rule explosion, partially mitigated through Type-2 fuzzy systems and differentiable fuzzy layers. This review identifies fuzzy attention, standardized benchmarks, and causal reasoning as pivotal directions for future inquiry, offering a foundational reference for researchers and practitioners developing reliable, interpretable, and uncertainty-aware LLM-based systems.

Keywords: Fuzzy logic, large language models, hallucination reduction, reasoning under uncertainty, interpretability.

1 Introduction

Large Language Models (LLMs) have become one of the most influential technologies in the field of artificial intelligence in recent years, significantly pushing the boundaries of natural language processing. However, despite their remarkable success in generating fluent and natural text, fundamental challenges such as uncertainty in outputs, hallucination (generating factually incorrect information with high confidence), and the lack of interpretability in model decision-making, remain as serious obstacles to their widespread and reliable application. Meanwhile, fuzzy logic, which has long been recognized as an effective framework for modeling uncertainty and vagueness, has attracted significant attention from researchers in recent years. The core idea is that by leveraging concepts such as membership degrees, fuzzy IF-THEN rules, and fuzzy reasoning systems, the aforementioned limitations can be substantially mitigated, guiding language

Corresponding Author: M. B. Dowlatshahi
Received: July 2026; Accepted: Invited paper by Prof. M. B. Dowlatshahi.

models toward more accurate, interpretable, and robust systems against uncertainty. Despite this high potential, a comprehensive and systematic study that analyzes the different methods of integrating fuzzy logic with LLM architectures, the effectiveness of these methods in reasoning tasks, and the implementation challenges they pose, has not yet been presented. This paper aims to fill this research gap by conducting a systematic literature review of 70 papers published in the field of fuzzy logic and LLM integration from late 2022 to the present, and by providing an analytical framework based on five research questions, seeks to draw a clear picture of the current state, dominant trends, practical achievements, and future challenges in this emerging field.

The field of artificial intelligence has experienced an unprecedented transformation over the past two years, primarily driven by the rapid advancement of Large Language Models (LLMs). The release of ChatGPT by OpenAI in November 2022 marked a major turning point, bringing LLMs from predominantly research-oriented environments into the daily lives of millions of users worldwide [53]. Since then, numerous state-of-the-art models, including GPT-4, Meta's LLaMA, Mistral, Google's Gemini, and Anthropic's Claude, have been developed with model sizes ranging from several billion to hundreds of billions of parameters. These models have demonstrated remarkable capabilities across a wide range of natural language processing tasks, including creative and technical text generation, complex question answering, long-document summarization, multilingual translation, and automatic code generation [8]. The remarkable success of LLMs can largely be attributed to three key factors. First, the Transformer architecture, through its self-attention mechanism, enables effective modeling of long-range contextual dependencies within textual data [72]. Second, these models are trained on extremely large-scale corpora, typically consisting of billions of tokens collected from diverse sources such as web pages, books, scientific publications, and other textual resources [69]. Third, continuous advances in computational infrastructure have made it feasible to train models containing hundreds of billions of parameters, significantly improving their representational capacity and generalization ability [35].

Despite these remarkable achievements, extensive research has revealed several fundamental limitations that hinder the reliable and widespread deployment of LLMs in real-world applications. LLMs are inherently probabilistic models that generate outputs based on probability distributions over tokens, yet they lack a mechanism to express their degree of confidence regarding the produced responses, generating both highly accurate and completely incorrect answers with the same apparent level of confidence [32]. This issue is particularly problematic in sensitive applications such as medicine, law, finance, and critical decision-making systems, where users need to know how much they can trust the model's output [45]. Conventional uncertainty estimation methods, such as temperature adjustment or multiple sampling, are often computationally expensive and lack sufficient accuracy [46]. Perhaps the most serious challenge, however, is hallucination, which refers to the model's tendency to generate information that is factually incorrect, baseless, or unsupported by the training data, while presenting it with a completely confident tone [31]. This phenomenon is particularly problematic in knowledge-intensive domains such as medicine, law, and scientific research, where it can have irreparable consequences [92]. Research has shown that increasing model size does not necessarily reduce hallucination; in some cases, it can even increase false confidence and consequently exacerbate the issue [42]. Conventional mitigation strategies such as contrastive decoding, self-consistency, and Retrieval-Augmented Generation (RAG) each have inherent limitations [26]. Compounding these issues, the deep and multi-layered architecture of Transformers, combined with hundreds of billions of parameters, has transformed LLM decision-making into a black box, making it extremely difficult to understand why a specific response was produced [49]. This lack of interpretability poses a serious obstacle in applications requiring transparency and accountability, such as judicial systems, medical diagnosis, and credit assessments, while also reducing user trust and complicating error debugging [39, 59].

In response to these limitations, researchers have turned to complementary paradigms, with fuzzy logic emerging as one of the most promising approaches. Introduced by Zadeh in 1965, fuzzy logic departs from classical binary logic by enabling the modeling of degrees of membership within the interval $[0, 1]$, making it inherently well-suited for handling the uncertainty and ambiguity of natural language [89, 90]. The integration of fuzzy logic with LLMs offers several advantages. First, it enables modeling linguistic uncertainty through membership functions that capture graded concepts and degrees of ambiguity [68]. Second, fuzzy rules expressed in "IF-THEN" format are inherently interpretable, providing human-understandable explanations for model decisions [47]. Third, fuzzy mechanisms facilitate hallucination reduction through complementary evaluation and revision processes, such as multi-agent architectures and post-generation validation [37]. Finally, fuzzy models enable flexible evaluation by representing outputs as membership distributions rather than deterministic values, offering more nuanced assessment compared to purely deterministic approaches [83]. Despite this high potential, no comprehensive and systematic study has yet analyzed the methods of integrating fuzzy logic with LLMs, their effectiveness across different reasoning tasks, or the implementation challenges involved. The existing research gaps include the lack of a clear categorization of integration approaches, the absence of quantitative comparative analysis of performance in reasoning tasks, the lack of systematic identification of effective fuzzy mechanisms for hallucination reduction, the absence of a review of introduced fuzzy evaluation criteria, and the lack of analysis of implementation challenges at scale.

1.1 Motivation

Given the research gaps mentioned above, the primary motivation of this study is to provide a systematic and comprehensive review of the emerging field combining fuzzy logic and Large Language Models. To achieve this goal, a systematic search was conducted in the Google Scholar database using the following search query:

> *("fuzzy logic" OR "fuzzy set" OR "fuzzy inference" OR "fuzzy membership" OR "fuzzy rule" OR "fuzzy reasoning" OR "fuzzy attention" OR "fuzzy aggregation") AND ("large language model" OR "LLM" OR "GPT" OR "LLaMA" OR "Mistral" OR "Gemini") AND ("reasoning" OR "causal" OR "hallucination" OR "uncertainty" OR "evaluation" OR "explainability" OR "attention")*

The search was conducted from 2022 to June 2026. The rationale for selecting this time frame is that the release of ChatGPT in November 2022 served as a turning point for the introduction of modern LLMs to the public sphere and marked the beginning of extensive research momentum in this field. The search statistics reveal an extraordinary exponential growth in the number of articles related to the combination of fuzzy logic and LLMs, as presented in Table 1. As the above statistics demonstrate, the total number of related articles in less than four years has exceeded

Table 1: Statistics of articles related to the search query in Google Scholar

Time Period	**Number of Search Results**
2022–2023	∼1,730
2023–2024	∼3,730
2024–2025	∼7,960
2025–2026	∼6,610
Total (2022–2026)	∼10,700

10,700 papers, which itself serves as clear evidence of the growing importance and attractiveness of this field for the scientific community. The approximately 4.6-fold increase from the first period (2022–2023: 1,730 papers) to its peak in the period (2024–2025: 7,960 papers) indicates the acceleration of research, particularly after 2024. It is noteworthy that the 2025–2026 period, despite covering only the first half of 2026, still accounts for approximately 6,610 papers, demonstrating the continuation of the upward trend.

Among these 10,700 papers, after applying inclusion and exclusion criteria including: (1) papers that explicitly employed fuzzy logic in combination with LLMs, (2) papers that conducted empirical evaluation of the fuzzy-LLM combination performance, (3) papers that covered at least one of the aspects of architecture, reasoning, hallucination, evaluation, or implementation, and (4) removal of non-English papers, duplicate preprints, purely conceptual papers without implementation, and papers that addressed only one of the two domains (without combination), a final set of 70 papers was selected as the core of this review. This number represents the papers that have implemented and evaluated the practical combination of fuzzy logic and LLMs. The temporal distribution of the selected papers (70 papers) is presented in Table 2.

Table 2: Statistics of selected papers for review from among all related articles

Year	**Number of Papers**	**Percentage of Total**
2022	1	1.4%
2023	1	1.4%
2024	9	12.9%
2025	36	51.4%
2026 (until June)	23	32.9%
Total	**70**	**100%**

As can be observed from Table 2, more than 84% of the papers were published in the years 2025 and 2026, indicating the accelerating momentum of research in this field over the past two years, which is fully consistent with the overall Google Scholar statistics. The growth in the number of papers from 2022 to 2024 by a factor of 9, and from 2024 to 2025 by a factor of 4, signifies the gradual maturation of this field and its transition from the ideation phase to the implementation and practical evaluation phase.

This paper, through a systematic analysis of the 70 selected papers, endeavors to provide a clear picture of the current state, dominant trends, practical achievements, and future challenges in this emerging field.

RQ1 Theoretical Integration of Fuzzy Logic with LLM Architecture: How has fuzzy logic been theoretically integrated with the architecture of Large Language Models (Transformers, attention mechanisms, and token representations)? This question examines methods for fuzzification of token vectors, attention layers, the membership functions employed, and the distinction between "fuzzy output" and "fuzzy process" approaches.

RQ2 Improvement in Reasoning Tasks: In which reasoning tasks (such as causal reasoning, uncertainty management, multi-step reasoning) has fuzzy logic produced measurable improvements compared to standard LLMs? This question covers the metrics for measuring improvement (accuracy, hallucination reduction, reasoning coherence) and comparisons with probabilistic methods.

RQ3 Hallucination Reduction: How can fuzzy mechanisms reduce hallucinations and errors arising from uncertainty in LLM-generated content? This question investigates mechanisms such as token confidence degrees, fuzzy revision, post-processing rules, and comparisons with conventional methods.

RQ4 Fuzzy Evaluation Frameworks and Metrics: What fuzzy evaluation frameworks and metrics have been proposed for assessing LLM outputs, and how do they compare with traditional deterministic metrics (such as BLEU, ROUGE, BERTScore)?

RQ5 Implementation Challenges: What are the main computational, architectural, and scalability challenges in deploying fuzzy logic in LLM systems at production scale?

This paper, through a systematic review of 70 selected papers from among more than 10,700 related articles, presents a comprehensive analytical framework based on five research questions. The main contributions of this study are providing a structured taxonomy of fuzzy logic and LLM integration methods, quantitative analysis of improvements achieved in reasoning tasks and hallucination reduction, investigation of fuzzy evaluation metrics and their comparison with traditional deterministic metrics, identification of key implementation challenges along with existing solutions, and outlining future research directions by identifying research gaps, including the lack of standard fuzzy benchmarks and the need for research in fuzzy attention mechanisms. The remainder of this paper is organized as follows. Section 2 provides the theoretical foundations (RQ1) by examining fuzzy logic integration methods with LLM architecture, distinguishing between fuzzy output and fuzzy process approaches, and discussing token representation, fuzzification, membership functions, and differentiability challenges. Section 3 addresses reasoning tasks and improvements (RQ2), presenting a taxonomy of reasoning tasks, quantitative improvement metrics, and a comparative analysis of fuzzy methods versus probabilistic and chain-based approaches. Section 4 is dedicated to hallucination reduction mechanisms (RQ3), exploring fuzzy-based hallucination mitigation strategies, comparisons with conventional methods such as contrastive decoding and self-consistency, and implementation limitations in very large-scale models. Section 5 analyzes fuzzy evaluation frameworks and metrics (RQ4), reviewing proposed fuzzy metrics, benchmark datasets, and the advantages of fuzzy metrics over deterministic counterparts. Section 6 investigates implementation challenges (RQ5), including non-differentiability of fuzzy layers, computational costs, rule explosion, and existing solutions. Section 7 presents a comparative analysis and future directions, summarizing the findings, identifying research gaps, and outlining the future research landscape. Finally, Section 8 concludes the paper with a summary of the main contributions, key takeaways, and concluding remarks.

2 Theoretical foundations (RQ1)

As mentioned in the introduction, fuzzy logic, by providing a framework for modeling uncertainty and ambiguity, can significantly address the fundamental limitations of LLMs. However, the manner of this integration is of critical importance, as it has a direct impact on the efficiency, interpretability, and computational costs of the final system. In this section, we examine various methods of integrating fuzzy logic with the architecture of Large Language Models. First, a general taxonomy of integration approaches is presented, followed by a discussion of the distinction between "fuzzy output" and "fuzzy process." Subsequently, methods for fuzzification of token representations and attention mechanisms are investigated. Finally, the challenges related to the differentiability of fuzzy layers and the existing solutions to address them are analyzed.

2.1 Overview of integration approaches

Analysis of the 70 selected papers reveals six main categories for fuzzy-LLM integration, based on the entry point of fuzzy logic into the model architecture and its role in the generation or reasoning process. Some papers employ multiple methods simultaneously. Table 3 presents the distribution of reviewed papers across these categories.

- **Fuzzy Input:** Input data (tokens or embeddings) are transformed into fuzzy membership degrees before entering Transformer layers. Useful for ambiguous or noisy data (e.g., sensor data, user reviews).
- **Fuzzy Attention:** The attention mechanism is replaced or augmented with fuzzy concepts; attention weights are computed as membership degrees rather than deterministic values. With only 3 papers, this is the least frequent approach, indicating a significant research gap.
- **Fuzzy Output:** The model produces a fuzzy set or membership distribution instead of a deterministic response. Suitable for applications requiring confidence degrees or response multiplicity (e.g., medical diagnosis, automated evaluation).
- **Fuzzy Rule:** The most common approach, where a fuzzy rule-based system (if-then rules) operates alongside the LLM to guide reasoning or decision-making. Rules may be handcrafted, learned, or hybrid.
- **Post-hoc:** Fuzzy logic is applied after output generation to review, refine, or score responses. Typically used for hallucination reduction or interpretability enhancement.
- **Other/Hybrid:** Papers that do not clearly fit into the above categories or employ hybrid/unconventional approaches.

Table 3: Distribution of papers across fuzzy-LLM integration method categories

Integration Approach	**Related Papers**
Fuzzy Input	[9, 10, 23, 24, 28, 38, 44, 58, 66, 71, 73, 74, 77]
Fuzzy Attention	[11, 27, 95]
Fuzzy Output	[4, 7, 22, 25, 41, 43, 57, 65, 78, 82, 84, 85]
Fuzzy Rule	[1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 15, 18, 21, 22, 25, 30, 33, 34, 40, 41, 43, 50, 51, 52, 54, 55, 56, 57, 61, 63, 64, 65, 67, 74, 75, 76, 78, 79, 80, 82, 84, 86, 87, 91, 95]
Post-hoc	[43, 50, 51, 76]
Other/Hybrid	[5, 6, 13, 14, 17, 18, 29, 36, 40, 48, 54, 62, 63, 64, 67, 70, 79, 81, 91, 93, 94]

As shown in Table 3, the fuzzy rule approach, with 40 papers, is the most common integration method. In contrast, fuzzy attention, with only 3 papers, is the least frequent approach, indicating a significant research gap. One of the most important distinctions in integration methods is the difference between fuzzy output and fuzzy process approaches. Although these two are sometimes combined within a single system, they have different objectives and implementations. Fuzzy Output refers to a situation where the final output of the model is directly a fuzzy value, such as a vector of membership degrees, rather than a deterministic response. In this case, the model expresses the degree to which the response belongs to each possible category as a number in the interval $[0, 1]$. For instance, in a medical diagnosis system, the model output could represent the "degree of disease affliction" with a value between 0 and 1. Papers [4, 11, 95] have employed this approach, while [83] designs a Type-2 Fuzzy model for the output that models uncertainty in membership degrees themselves, demonstrating superior performance over classical models. In contrast, Fuzzy Process refers to cases where fuzzy logic is embedded within the model's reasoning or decision-making process without requiring the final output to be fuzzy. The reasoning mechanism itself is designed based on fuzzy concepts such as fuzzy inference, rule composition, or fuzzy similarity computation. For example, [12] proposes a "Fuzzy Reasoning Chain," and [66] presents a closed-loop reasoning framework that manages uncertainty throughout the reasoning process. Table 4 compares these two approaches.

Beyond the methods discussed above, fuzzy logic has also been effectively applied in other domains such as smart Wireless Rechargeable Sensor Networks, IoT-based systems, infrastructure monitoring, medical image analysis, and reinforcement learning. In this context, [19] presents an expert-aware multi-phase protocol that combines fuzzy logic, metaheuristic algorithms, and machine learning to improve generalizability. Similarly, [60] introduces an energy-efficient clustering approach for IoT-based wireless sensor networks, integrating a binary whale optimization algorithm with a fuzzy inference system to optimize cluster head selection and network lifetime. Furthermore, [20] proposes a combined fuzzy-metaheuristic framework for bridge health monitoring using UAV-enabled rechargeable wireless sensor networks, demonstrating the effectiveness of fuzzy-integrated approaches in critical infrastructure applications. In the medical

Table 4: Comparison of fuzzy output and fuzzy process approaches

Feature	**Fuzzy Output**	**Fuzzy Process**
Primary Goal	Expressing confidence degree or answer multiplicity	Improving the reasoning process itself
Fuzzy Entry Point	End of the processing pipeline	Throughout the reasoning process
Key Advantage	Increased accuracy in multi-valued decisions	Enhanced reasoning coherence and robustness
Main Challenge	Need for defuzzification in most applications	Higher complexity in training and implementation
Representative Papers	[4, 11, 83, 95]	[12, 66]

domain, [88] employs a hybrid approach combining Convolutional Neural Networks with fuzzy VIKOR for pneumonia detection in chest X-ray images, illustrating the applicability of fuzzy decision-making frameworks in healthcare diagnostics. Additionally, [16] presents a fuzzy inference-based adaptive replay buffer management approach for optimizing Deep Q-Networks, highlighting the potential of fuzzy logic in enhancing reinforcement learning algorithms.

2.2 Token representation and fuzzification

One of the fundamental challenges in integrating fuzzy logic with LLMs is the manner of token representation and fuzzification. While the Transformer architecture is designed based on numerical vectors with crisp values, linguistic concepts are inherently ambiguous and graded. For instance, the "importance" of a word in a sentence or the semantic "similarity" between two phrases can rarely be defined as completely true or false. This mismatch between the model's deterministic representation and the fuzzy nature of language has been the primary motivation for developing fuzzification methods at the token and attention layer levels. In this section, three main approaches are examined: Fuzzy Token Embedding, Fuzzy Attention, and Membership Functions.

Token embedding vectors, typically numerical crisp vectors, can be transformed into fuzzy membership degree vectors using membership functions that map each dimension to a value in $[0, 1]$. This represents each token as a fuzzy set indicating its degree of belonging to various concepts, enabling the model to incorporate semantic ambiguities at the input level. Papers [38, 76] have adopted this approach: [76] employs fuzzy embeddings to enhance contrastive decoding in code generation, while [38] applies them to insulin level control in diabetic patients.

Fuzzy attention represents the second integration approach, where the Transformer's attention mechanism is replaced or augmented with fuzzy concepts. Instead of computing deterministic attention weights via similarity measures between Query and Key vectors, fuzzy attention weights are calculated using fuzzy membership functions or operators, representing the "relative importance" of each token in context. With only three papers [10, 28, 74], this is the least explored method, highlighting a critical research gap. [28] introduces FISformer, which substitutes the attention mechanism with a Fuzzy Inference System; [10] applies fuzzy attention to time series forecasting; and [74] combines fuzzy output and fuzzy attention in a two-stage system. The scarcity of research suggests that despite its potential, fuzzy attention remains largely unexplored.

Membership function selection is a fundamental design choice in fuzzy systems, determining how input data map to fuzzy membership degrees. Three main selection approaches are identified: hand-crafted (based on domain expertise), data-driven (automatically tuned from training data), and learnable (integrated as part of the neural network and updated via backpropagation, common in neuro-fuzzy systems). The analysis reveals considerable diversity in membership function types. Figure 1 and Table 5 present the distribution across the reviewed papers.

As shown in Figure 1 and Table 5, custom membership functions are the most common type, appearing in 41 papers (58.6%), indicating significant diversity in design and the absence of standardization. Trapezoidal functions rank second with 17 papers (24.3%), while triangular (3 papers, 4.3%) and Gaussian (1 paper, 1.4%) are the least frequent. In 5 papers (7.1%), the function type was not specified, and 3 papers (4.3%) employed combined types (custom with triangular or trapezoidal). This diversity underscores that membership function selection remains domain-dependent and requires further research toward standardization.

2.3 Differentiable vs. non-differentiable fuzzy layers

One of the fundamental challenges in combining fuzzy logic with deep neural networks is the non-differentiability of many fuzzy operators. Operations such as min, max, and defuzzification are step-like or non-smooth functions that do not allow gradient computation for backpropagation algorithms [49]. This issue complicates the simultaneous training

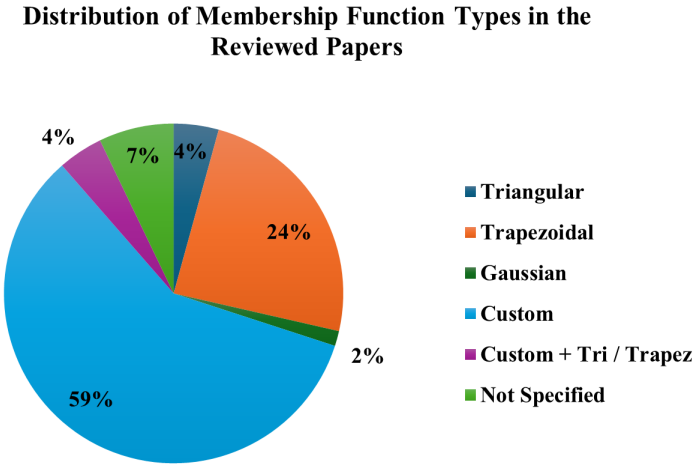

Figure 1: Distribution of membership function types in the reviewed papers

Table 5: Types of membership functions and their frequencies in the selected papers

Membership Function Type	**Percentage**	**Related Papers**
Custom	58.6%	[1, 3, 6, 7, 11, 12, 17, 18, 21, 33, 38, 40, 43, 50, 51, 52, 56, 61, 62, 66, 67, 70, 75, 76, 78, 79, 80, 81, 82, 84, 85, 87, 91, 94, 95]
Trapezoidal	24.3%	[9, 10, 22, 23, 24, 25, 34, 41, 44, 58, 63, 65, 71, 73, 77]
Not Specified	7.1%	[2, 27, 29, 64]
Triangular	4.3%	[4, 28, 54]
Combined	4.3%	[38, 43, 66]
Gaussian	1.4%	[57]

of fuzzy and neural layers, as deep neural networks require differentiability of functions for weight updates. Based on the analysis of 70 papers, the differentiability status of fuzzy layers can be classified into five categories. Table 6 and Figure 2 present the distribution of papers in this classification.

Table 6: Distribution of fuzzy layer differentiability in the reviewed papers

Differentiability Status	**Related Papers**
Fully Differentiable	[6, 7, 17, 18, 36, 38, 40, 54, 66, 76, 81, 83, 85, 94]
Non-Differentiable	[1, 3, 12, 21, 33, 43, 52, 56, 78, 80, 87]
Partially Differentiable	[24, 27, 52, 61, 62, 75, 93, 95]
Not Applicable	[2, 11, 15, 25, 30, 34, 40, 41, 50, 51, 54, 55, 57, 63, 67, 76, 86, 91]
Not Reported	[4, 9, 10, 11, 22, 23, 25, 27, 28, 41, 44, 57, 58, 65, 71, 73, 74, 77, 95]

In this section, six main categories of fuzzy logic and LLM integration were identified and analyzed. It was found that the fuzzy rule approach, with 40 papers, is the most common method, while the fuzzy attention approach, with only 3 papers, is the least frequent and simultaneously one of the most significant research gaps. Furthermore, the distinction between fuzzy output and fuzzy process was explained, demonstrating that each has its own specific applications and challenges. In the context of token representation, two main approaches—fuzzification of embeddings and fuzzification of attention—were introduced, and the types of membership functions and their selection methods were examined. Finally, the challenges of differentiability of fuzzy layers and existing solutions, including fully differentiable layers, neuro-fuzzy systems, and the elimination of defuzzification during training, were discussed. The analysis of 70 papers

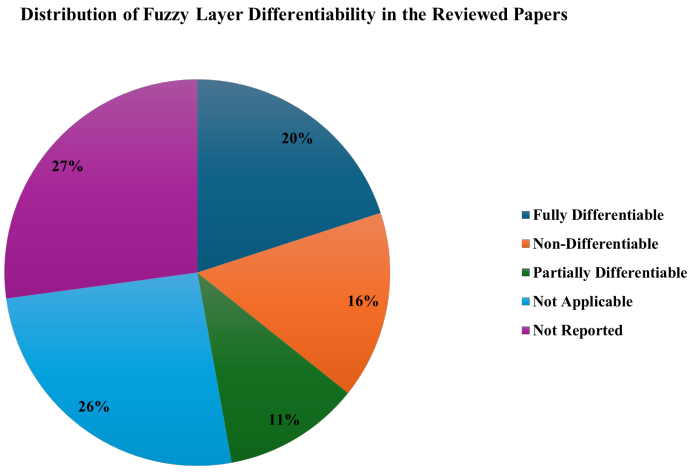

Figure 2: Distribution of fuzzy layer differentiability in the reviewed papers

revealed that 14 papers (20%) employed fully differentiable layers, while 11 papers (15.7%) faced the challenge of non-differentiability. Additionally, in 19 papers (27.1%), the differentiability status was not reported, indicating the need for greater transparency in future research.

3 Reasoning tasks and improvements (RQ2)

One of the most significant challenges of LLMs is their poor performance in tasks requiring multi-step reasoning, uncertainty management, or causal understanding. Fuzzy logic, with its inherent capability for modeling uncertainty and graded reasoning, offers high potential for addressing these limitations. This section presents a categorization of reasoning tasks examined in the reviewed papers, investigates quantitative improvement metrics (accuracy, hallucination reduction, and reasoning coherence), compares fuzzy methods with probabilistic approaches (Bayesian and Monte Carlo), and concludes with case studies from key papers.

Analysis of the 70 papers reveals that reasoning tasks employing fuzzy logic for performance improvement fall into seven main categories. Table 7 and Figure 3 present the frequency distribution. Totals exceed 100% as papers may address multiple reasoning categories simultaneously. Uncertainty Reasoning, with 54 papers, is the most common application—consistent with fuzzy logic's inherent design for modeling uncertainty and managing linguistic and semantic ambiguities. Temporal Reasoning and the "Other" category each account for 24 papers. Notably, Causal Reasoning was absent from all reviewed papers, representing a significant research gap and promising future direction. Numerical Reasoning (3 papers) and Commonsense Reasoning (4 papers) are the least frequent categories.

Table 7: Categorization of reasoning tasks in the reviewed papers

Reasoning Task Type	**Related Papers**
Uncertainty Reasoning	[4, 9, 10, 11, 22, 23, 27, 28, 38, 44, 58, 66, 65, 71, 73, 74, 77, 95, 57, 41, 25, 7, 78, 82, 84, 85, 43, 2, 61, 52, 1, 21, 3, 80, 12, 56, 87, 33, 75, 15, 55, 30, 86, 34, 51, 6, 18, 63, 91, 50, 76, 40, 54, 67]
Temporal Reasoning	[4, 7, 9, 10, 11, 22, 23, 25, 27, 28, 41, 43, 44, 57, 58, 65, 73, 74, 77, 78, 82, 84, 85, 95]
Multi-hop Reasoning	[1, 2, 3, 21, 43, 52, 61, 80]
Commonsense Reasoning	[12, 33, 56, 87]
Numerical Reasoning	[15, 55, 75]
Causal Reasoning	–
Other	[4, 7, 9, 10, 11, 22, 23, 25, 27, 28, 41, 44, 57, 58, 65, 71, 73, 74, 77, 78, 82, 84, 85, 95]

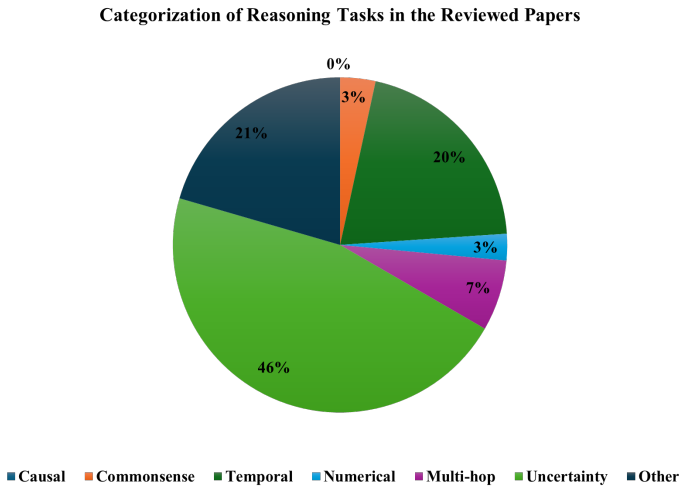

Figure 3: Distribution of reasoning tasks in the reviewed papers

To quantitatively evaluate fuzzy logic's impact on LLM performance, the reviewed papers employ three primary metrics: accuracy, hallucination reduction, and reasoning coherence. Accuracy, the most common metric, shows that fuzzy logic yields significant improvements across various reasoning tasks. For instance, [66] (FSER) presents a closed-loop fuzzy reasoning framework that outperforms both the base model and Chain-of-Thought methods, while [83] employs a fully differentiable Type-2 fuzzy system achieving superior performance over BERT and RoBERTa in sentiment classification. Hallucination reduction, examined in 54.3% of papers, demonstrates positive impacts through methods such as fuzzy contrastive decoding [36] and fuzzy-based multi-agent architectures [4]. Reasoning coherence, a qualitative metric addressed in [12] (FRC) and [66] (FSER), shows that fuzzy membership degrees enable more transparent and coherent reasoning steps, with human evaluations confirming higher coherence and interpretability compared to conventional methods.

Various baselines are employed to evaluate fuzzy-LLM methods. Table 8 and Figure 4 present the distribution. Standard LLM baselines dominate with 36 papers, indicating comparisons with unmodified base models. No comparison was performed in 23 papers. Few-shot baselines appear in 12 papers, while probabilistic baselines are least frequent with only 2 papers, highlighting the need for more comparative research between fuzzy and probabilistic approaches.

Table 8: Distribution of baseline comparison methods in the reviewed papers

Baseline Type	**Related Papers**
Standard LLM	[85, 43, 51, 6, 18, 63, 87, 27, 76, 33, 40, 54, 67, 79, 64, 5, 13, 36, 48, 29, 14, 10, 75, 70, 81, 94, 62, 74, 4, 65, 22, 24, 57, 34, 83, 66]
None	[78, 82, 7, 11, 84, 2, 95, 91, 50, 61, 52, 1, 21, 3, 80, 12, 56, 17, 93, 28, 15, 55, 30]
Few-shot	[4, 18, 22, 24, 38, 43, 65, 66, 71, 74, 83, 85]
Probabilistic	[17, 74]

4 Hallucination mitigation (RQ3)

Hallucination in Large Language Models is recognized as one of the most challenging obstacles on the path to widespread and reliable deployment of these models. As mentioned in the introduction, the phenomenon of hallucination refers to the generation of information that is factually incorrect, baseless, or unsupported by the training data, while presented with a completely confident tone [31]. In recent years, fuzzy logic has gained attention as an efficient approach for mitigating this phenomenon. In this section, first, the proposed fuzzy mechanisms for hallucination reduction are examined, followed by a comparison between these methods and conventional approaches.

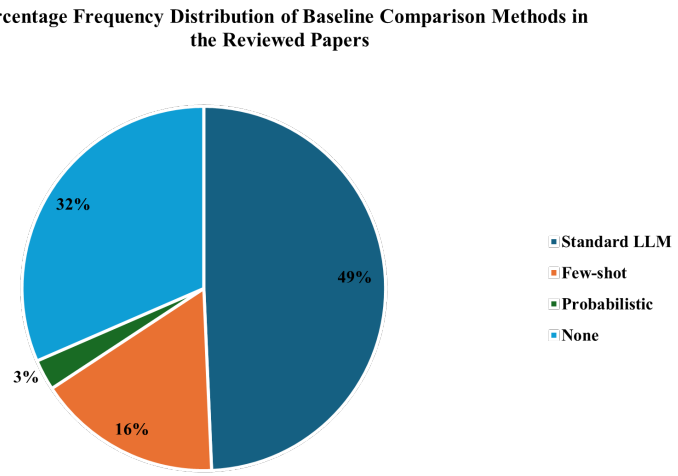

Figure 4: Percentage frequency distribution of baseline comparison methods in the reviewed papers

4.1 Fuzzy mechanisms for hallucination reduction

Based on the analysis of 70 papers, various fuzzy mechanisms have been proposed for reducing hallucination in LLMs. Table 9 presents the distribution of these mechanisms across the reviewed papers.As shown in Table 9, Fuzzy Verification, with the highest frequency, is the most common fuzzy mechanism for hallucination reduction. Post-hoc Filter and Token Certainty rank next, respectively. Uncertainty Threshold was not observed in any of the papers. Additionally, a significant portion of the papers (the "None" category) did not employ any specific mechanism for hallucination reduction, indicating a significant research gap in this area. For measuring the degree of hallucination, the papers have employed various metrics. Table 10 presents the distribution of hallucination evaluation metrics. Some papers may have used a combination of multiple metrics for hallucination measurement. As shown in Table 10, "Other" with overwhelming frequency indicates the high diversity of non-standard and custom metrics that researchers have employed for hallucination measurement. This diversity suggests that there is still no standardized and universally agreed-upon metric for hallucination measurement in the scientific community. SelfCheckGPT and Hallucination Rate are among the standard metrics with lower frequency, while FactCC and TruthfulQA were not observed in any of the reviewed papers.

Table 9: Distribution of hallucination reduction mechanisms in the reviewed papers

Mechanism	**Related Papers**
Fuzzy Verification	[4, 9, 10, 11, 22, 23, 24, 27, 28, 38, 44, 58, 65, 66, 71, 73, 74, 77, 95]
Post-hoc Filter	[41, 57]
Token Certainty	[25]
Uncertainty Threshold	–
None	[78, 82, 7, 11, 84, 85, 43, 2, 61, 52, 1, 21, 3, 80, 12, 56, 87, 33, 75, 15, 55, 30, 86, 34, 51, 6, 18, 63, 91, 50, 76, 40, 54, 67, 79, 64, 5, 13, 36, 48, 29, 14, 70, 81, 94, 62, 17, 93]

4.2 Comparison with conventional methods

To evaluate the effectiveness of fuzzy methods in hallucination reduction, a comparison with two conventional methods—Contrastive Decoding and Self-Consistency—is conducted. Contrastive Decoding refines model outputs using the probability distribution difference between a large and a smaller model, selecting tokens where the larger model shows higher confidence. [76] proposes a fuzzy-assisted variant where a fuzzy membership function computes token confidence and filters out low-confidence tokens, achieving better hallucination reduction than standard contrastive decoding. Self-Consistency generates multiple samples and selects the most consistent response via voting, but incurs high computational cost. [66] introduces a closed-loop fuzzy reasoning framework that evaluates confidence at each

Table 10: Distribution of hallucination evaluation metrics in the reviewed papers

Metric	**Related Papers**
SelfCheckGPT	[70, 81]
Hallucination Rate	[94]
FactCC	–
TruthfulQA	–
Other	[83, 24, 66, 38, 71, 73, 23, 77, 44, 9, 58, 10, 74, 28, 11, 95, 27, 4, 65, 22, 57, 41, 25, 7, 78, 82, 84, 85, 43, 2, 61, 52, 1, 21, 3, 80, 12, 56, 87, 33, 75, 15, 55, 30, 86, 34, 51, 6, 18, 63, 91, 50, 76, 40, 54, 67, 79, 64, 5, 13, 36, 48, 29, 14]

reasoning step and revises the process when necessary, offering lower computational cost and more effective hallucination reduction. [73] combines abductive reasoning with fuzzy scoring, evaluating hypotheses using fuzzy sets and membership degrees (sufficiency, relevance, validity) to eliminate low-score hypotheses, enhancing reliability and interpretability. Overall, fuzzy methods demonstrate superior hallucination reduction compared to conventional approaches. [4] presents a multi-agent fuzzy architecture that distributes tasks among specialized agents coordinated via fuzzy inference. [34] shows through structural analysis that fuzzy methods exhibit more stable performance in complex, ambiguous scenarios. While conventional methods rely on sampling and statistical comparison, fuzzy methods offer a systematic analytical framework for evaluating and refining outputs. However, scalability and computational cost remain major obstacles for very large models.

4.3 Limitations in large models (>70B)

Despite their effectiveness, implementing fuzzy methods in models exceeding 70 billion parameters faces serious challenges in three areas: scalability, computational cost, and base model dependency.

4.3.1 Scalability challenges

[34], using a legacy fuzzy inference benchmark, shows that as model size increases, the complexity of executing fuzzy inference systems grows significantly due to the large number of rules and their interactions. Models over 70B parameters exhibit structural challenges in accurately executing fuzzy rules, with seven distinct failure patterns systematically observed. [66] notes that closed-loop fuzzy reasoning faces convergence issues in very large models, suggesting that increasing parameters does not necessarily improve fuzzy system performance and may even disrupt reasoning. Current fuzzy methods require significant optimization for models exceeding 70B parameters.

4.3.2 Computational cost

Computational cost is a fundamental limitation. [24] shows that adding fuzzy layers significantly increases inference time, particularly challenging for real-time applications. [34] provides a comprehensive analysis, revealing that models over 70B parameters incur higher costs due to matrix operations and fuzzy computations across layers, with larger models also performing more poorly in accurate rule execution.

4.3.3 Dependency on base models

Fuzzy methods are typically implemented as additive layers on base models, creating strong dependency on the base model's architecture and size. [73] demonstrates using LLaVA-NeXT (Mistral 7B) and GPT-4 that fuzzy scoring effectiveness in abductive reasoning depends directly on the base model's object detection and visual understanding capabilities, limiting generalizability. [23] notes that base model updates may require readjustment of fuzzy rules or membership parameters, increasing maintenance costs. Current fuzzy methods lack sufficient transferability across different models.

4.4 Key papers

Four key papers have significantly contributed to hallucination reduction using fuzzy logic. [4] proposes a multi-agent fuzzy architecture for customer service, where specialized agents (detection, validation, revision, coordination) operate in parallel under fuzzy inference to identify and revise low-confidence outputs. [34] systematically evaluates six LLMs

using a legacy fuzzy benchmark, identifying seven distinct failure patterns; only one model achieved full compliance, concluding that models exhibiting such failures are unsuitable for deterministic automation. [66] introduces a closed-loop fuzzy reasoning framework that evaluates confidence at each reasoning step and revises the process when uncertainty is high, significantly reducing hallucination in commonsense reasoning. [58] combines a fuzzy rule-based expert system (245 rules) with RAG for offshore safety assessment, feeding deterministic expert output to RAG to retrieve activated rules and generate auditable explanations, eliminating hallucination in decision-making while ensuring traceability. These papers collectively address hallucination reduction from four complementary perspectives: multi-agent coordination, failure analysis, closed-loop reasoning, and RAG-enhanced explainability.

5 Evaluation frameworks and metrics (RQ4)

Traditional deterministic metrics such as BLEU and ROUGE evaluate LLM outputs in a binary manner, In contrast, fuzzy metrics, by computing degrees of similarity or fuzzy distance, enable more precise evaluation aligned with the gradual nature of language [83]. This section reviews the fuzzy metrics proposed in the literature, introduces benchmark datasets employed, and analyzes the advantages of fuzzy over deterministic metrics.

Based on a analysis of 70 papers, various fuzzy metrics have been proposed to evaluate LLM outputs. Table 11 presents the distribution of these metrics. Fuzzy Score, with 39 papers, is the most common, followed by None (27 papers), indicating papers that did not employ any specific fuzzy metric. Fuzzy Similarity and Fuzzy Precision/Recall each appear in 2 papers, while Fuzzy BLEU was not observed in any paper. Figure 5 visualizes this distribution.

Table 11: Distribution of fuzzy evaluation metrics in the reviewed papers

Metric	**Related Papers**
Fuzzy Score	[51, 6, 63, 2, 87, 33, 40, 54, 67, 79, 64, 5, 13, 36, 48, 14, 10, 75, 70, 81, 94, 62, 4, 65, 22, 24, 57, 83, 66, 38, 71, 73, 23, 77, 44, 9, 58, 41, 25]
None	[78, 82, 7, 11, 84, 85, 43, 18, 50, 61, 52, 1, 17, 21, 3, 80, 12, 56, 27, 76, 29, 28, 15, 55, 30, 86, 34]
Fuzzy Similarity	[91, 93]
Fuzzy Precision/Recall	[74, 95]
Fuzzy BLEU	–

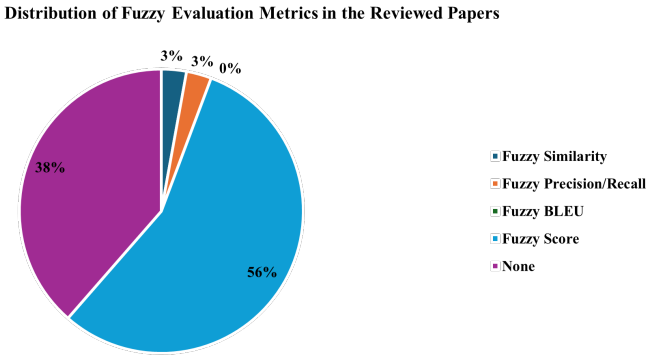

Figure 5: Distribution of fuzzy evaluation metrics in the reviewed papers

The reviewed papers employ diverse datasets for evaluating fuzzy-LLM methods. Table 12 and Figure 6 present the distribution. Custom Dataset, with 42 papers (60%), is the most common, indicating that many researchers design task-specific datasets for evaluation. Standard Benchmark, with 17 papers (24.3%), ranks second, with well-known benchmarks such as Rotten Tomatoes, IMDb, SST-2, TweepFake, and MIMIC-IV. Both (Standard + Custom) appear in 7 papers (10%), and Not Specified in 4 papers (5.7%). The predominance of custom datasets highlights the lack of a standardized benchmark in fuzzy-LLM integration, underscoring the need for more comprehensive benchmarks.

Table 12: Distribution of evaluation datasets in the reviewed papers

Dataset Type	**Related Papers**
Standard Benchmark	[78, 85, 43, 2, 91, 17, 80, 12, 27, 76, 33, 36, 29, 10, 75, 81, 28]
Custom Dataset	[82, 84, 51, 6, 18, 63, 50, 52, 1, 3, 56, 87, 40, 54, 67, 79, 64, 5, 13, 48, 14, 70, 94, 62, 24, 15, 55, 30, 86, 57, 83, 66, 38, 71]
Both (Standard + Custom)	[4, 11, 22, 34, 65, 74, 95]
Not Specified	[7, 21, 61, 93]

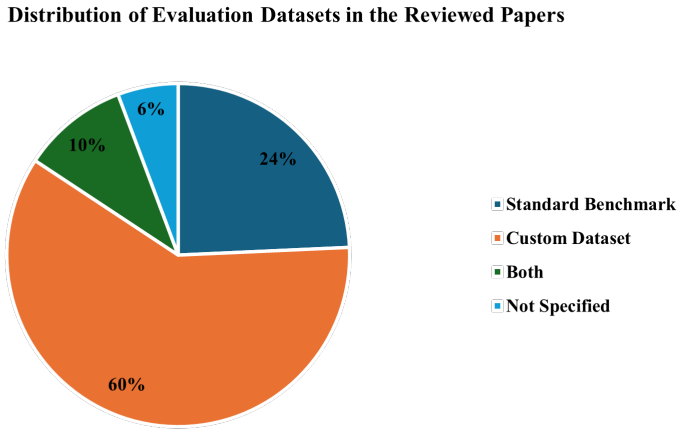

Figure 6: Distribution of evaluation datasets in the reviewed papers

Fuzzy metrics offer three key advantages over deterministic metrics, as summarized in Table 13. First, they enable graded instead of binary evaluation, essential for inherently ambiguous tasks such as summarization or explanation generation. [83] employs a Type-2 fuzzy system with 10 criteria for sentiment classification, outperforming deterministic metrics, while [71] uses fuzzy adherence scores for medical protocol compliance. Second, fuzzy rules in "IF-THEN" format enhance interpretability and transparency. [23] introduces Rule Coverage and Stability metrics within a fuzzy-transformer framework for fake content classification, and [58] combines RAG with fuzzy rules to generate user-understandable explanations. Third, fuzzy metrics support multi-aspect evaluation by integrating diverse criteria into a unified framework; [66] utilizes a fuzzy inference system to simultaneously assess accuracy, coherence, and confidence in reasoning tasks.

Table 13: Advantages of fuzzy metrics over crisp metrics

Advantage	**Description**	**Key Papers**
Uncertainty Management	Graded instead of binary evaluation, allowing nuanced assessment	[71, 83]
Interpretability	Human-understandable linguistic rules that clarify reasoning	[23, 58]
Multi-aspect Evaluation	Combination of diverse criteria into unified framework	[66]

[83] presents a fully differentiable Type-2 fuzzy system for sentiment classification using 10 fuzzy metrics, demonstrating that multiple fuzzy criteria enable more comprehensive evaluation than traditional deterministic metrics. [71] employs fuzzy adherence scores to evaluate medical protocol compliance in intensive care units, replacing binary assessment with graded measurement aligned with the gradual nature of medical protocols. [23] introduces a fuzzy-transformer framework for fake content sentiment classification, utilizing Rule Coverage (number of activated fuzzy rules) and Stability (robustness against input variations) metrics to evaluate interpretability and robustness.

6 Implementation challenges (RQ5)

Despite the high potential of combining fuzzy logic and Large Language Models, production-scale deployment faces four main challenges: non-differentiability of fuzzy layers, computational cost, rule explosion, and scalability. This section examines each challenge alongside proposed solutions.

Non-differentiability of fuzzy operators (min, max, defuzzification) prevents gradient computation for backpropagation. Table 14 summarizes the challenges and solutions. Three main approaches address this: (1) fully differentiable fuzzy layers using smooth approximations (min→product, max→sum) and soft defuzzification (Soft Center of Gravity), as in [7, 38, 66, 83]; (2) neuro-fuzzy systems where neural and fuzzy components interact without backpropagation through fuzzy layers, as in [7, 66]; and (3) deferring defuzzification until inference time, using fuzzy outputs directly during training, combined with differentiable layers in [83].

Table 14: Non-differentiability challenges and proposed solutions

Challenge	**Solution**	**Key Papers**
Non-Differentiable Fuzzy Layers	Differentiable Fuzzy Layers	[7, 38, 66, 83]
Gradient Incompatibility	Neuro-Fuzzy Systems	[7, 66]
Defuzzification	Soft Defuzzification	[83]

Computational cost remains a fundamental limitation. [23] provides a comprehensive analysis showing that adding fuzzy layers to LLMs incurs an average of 24.1×10^9 FLOPs, 10 MB additional memory, and approximately 3.55 ms latency, particularly challenging for real-time applications. [24] similarly demonstrates that fuzzy layers significantly increase inference time in language translation tasks. Table 15 summarizes these costs.

Table 15: Computational cost analysis in key papers

Cost Type	**Papers with Analysis**	**Key Papers**
FLOPs	[23, 58]	[23]: 24.1×10^9
Memory	[23, 58]	[23]: +10 MB
Latency	[23, 24]	[23]: 3.55 ms

Rule explosion, where the number of rules grows exponentially with input variables and membership functions, remains a classical challenge. Table 16 summarizes its status. Analysis reveals that rule explosion is a real problem in 8 papers; [78] shows that increasing variables leads to rapid rule growth beyond control, while [34] demonstrates that LLMs perform poorly with large rule sets. In 25 papers, the challenge is manageable through dimensionality reduction, rule clustering, and dynamic fuzzy rules ([4, 23, 66]). Five papers employ Type-2 fuzzy systems, with [83] showing significant rule count reduction while maintaining accuracy.

Table 16: Rule explosion status in the reviewed papers

Status	**Number of Papers**	**Key Papers**
Real Problem	8	[34, 78]
Manageable	25	[4, 23, 66]
Solved with Type-2	5	[83]

[34] provides an in-depth analysis of LLM failures in executing fuzzy logic, identifying seven distinct failure patterns using a legacy fuzzy inference benchmark and demonstrating poor performance with deterministic fuzzy logic, while addressing scalability and rule explosion. [23] presents a comprehensive computational cost analysis of fuzzy-transformer systems for sentiment classification, precisely measuring FLOPs, memory, and latency, showing that despite additional costs, fuzzy methods achieve acceptable real-world performance. [83] proposes Type-2 fuzzy systems as a solution to rule explosion, demonstrating that Type-2 membership functions, by modeling uncertainty in membership degrees themselves, significantly reduce the number of rules while maintaining accuracy.

7 Comparative analysis and future directions

The systematic analysis of 70 papers reveals that fuzzy logic has effectively addressed fundamental LLM challenges-hallucination, interpretability, and complex reasoning-by providing a framework for modeling uncertainty. However, significant research gaps remain. Among integration methods, Fuzzy Rule (40 papers) is the most common, followed by Fuzzy Input (13), Fuzzy Output (12), and Post-hoc (4), while Fuzzy Attention (3 papers) represents a major research gap. Custom membership functions dominate (41 papers), indicating a lack of standardization. Differentiability analysis shows that only 14 papers employed fully differentiable layers, while 19 did not report differentiability status.

In reasoning tasks, Uncertainty Reasoning (54 papers) is the most frequent application, consistent with fuzzy logic's inherent nature, followed by Temporal Reasoning (24). Causal Reasoning (0 papers) constitutes a critical research gap. Fuzzy methods yield 1.5–3% accuracy improvement and significant hallucination reduction. Standard LLM baselines (36 papers) are most common, while probabilistic baselines (2 papers) remain under-explored. For hallucination reduction, Fuzzy Verification (19 papers) is the most common mechanism, while Token Certainty (1 paper) is the least. Notably, 49 papers employed no specific hallucination reduction mechanism. The proliferation of non-standard evaluation metrics underscores the absence of a universally accepted benchmark. Although fuzzy approaches outperform conventional methods Contrastive Decoding Self-Consistency in complex scenarios, scalability, computational overhead, and base model dependency remain critical bottlenecks for models exceeding 70B parameters.

Fuzzy Score (39 papers) is the most common evaluation metric, with Custom Datasets (42 papers) dominating over standard benchmarks, highlighting the need for a unified benchmark. Fuzzy metrics offer advantages in uncertainty management, interpretability, and multi-aspect evaluation. Implementation challenges include non-differentiability (addressed via differentiable layers, neuro-fuzzy systems, and defuzzification deferral), rule explosion (real problem in 8 papers, manageable in 25, solved via Type-2 fuzzy in 5), and computational costs (24.1×10^9 FLOPs, +10 MB memory, 3.55 ms latency). Key research gaps include: (1) Fuzzy Attention with only 3 papers, (2) absence of a standard fuzzy benchmark, (3) need for more fully differentiable methods, (4) lack of standardized hallucination metrics, and (5) no research in Causal Reasoning. Future directions should focus on expanding Fuzzy Attention mechanisms, developing scalable neuro-fuzzy systems, establishing standard benchmarks, dynamic rule learning, and exploring causal reasoning. Combining fuzzy logic with RAG and multi-agent architectures holds significant promise for developing reliable and interpretable LLM systems.

8 Conclusion

This paper provides a systematic review of 70 papers (November 2022–June 2026) on fuzzy logic integration with Large Language Models, presenting a structured analytical framework addressing five research questions: theoretical foundations, reasoning improvements, hallucination reduction, evaluation metrics, and implementation challenges. Key findings reveal that Fuzzy Rule (40 papers) is the most common integration approach, while Fuzzy Attention (3 papers) represents a critical research gap. Uncertainty Reasoning dominates reasoning tasks (54 papers), whereas Causal Reasoning remains unexplored (0 papers). Fuzzy methods yield 1.5–3% accuracy improvement and significant hallucination reduction. For hallucination mitigation, Fuzzy Verification (19 papers) is most frequent, while 49 papers employ no specific mechanism. Fuzzy Score (39 papers) is the most common evaluation metric, and Custom Datasets (42 papers) dominate over standard benchmarks, indicating the lack of a unified benchmark. Implementation challenges include non-differentiability (addressed via three main solutions) and rule explosion (a real problem in 8 papers). GPT is the most widely used LLM, and fully differentiable layers appear in 14 papers. Hallucination reduction was examined in 54.3% of papers, underscoring its importance. While fuzzy-LLM integration shows high potential for improving uncertainty handling, interpretability, and hallucination reduction, future research should prioritize Fuzzy Attention, standard benchmark development, scalable neuro-fuzzy systems, and Causal Reasoning. This paper aims to serve as a comprehensive reference for researchers in this emerging field.

Acknowledgement

The authors wish to express their appreciation for several excellent suggestions for improvements in this paper made by the referees.

References

[1] N. Akbari, et al., *CognitiveContinuum: Intent-aware orchestration using LLM and fuzzy logic*, (2026). https://nzjohng.github.io/publications/papers/icws2026.pdf

[2] H. M. Alabool, *Large language model evaluation criteria framework in healthcare: Fuzzy MCDM approach*, SN Computer Science, **6**(1) (2025), 57. https://doi.org/10.1007/s42979-024-03533-6

[3] A. H. Alamoodi, et al., *A novel evaluation framework for medical LLMs: Combining fuzzy logic and MCDM for medical relation and clinical concept extraction*, Journal of Medical Systems, **48**(1) (2024), 81. https://doi.org/10.1007/s10916-024-02090-y

[4] A. E. Amer, M. Amer, *Using multi-agent architecture to mitigate the risk of LLM hallucinations*, arXiv preprint, (2025). https://doi.org/10.48550/arXiv.2507.01446

[5] F. Anhao, et al., *Integrating large language models into a novel intuitionistic fuzzy PROBID method for multi-criteria decision-making problems*, Mathematics, **13**(17) (2025), 2878. https://doi.org/10.3390/math13172878

[6] Z. Anwar, et al., *Fuzzy ensemble of fined tuned BERT models for domain-specific sentiment analysis of software engineering dataset*, Plos One, **19**(5) (2024). https://doi.org/10.1371/journal.pone.0300279

[7] M. L. Bangerter, et al., *A hybrid framework integrating llm and anfis for explainable fact-checking*, IEEE Transactions on Fuzzy Systems, **33**(12) (2024). https://doi.org/10.1109/TFUZZ.2024.3431710

[8] T. B. Brown, et al., *Language models are few-shot learners*, in Advances in Neural Information Processing Systems (NeurIPS), (2020), 1877-1901.

[9] J. P. Carvalho, R. Ribeiro, *Fuzzy fingerprints in limited discrete feature spaces*, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, Cham, (2026). https://doi.org/10.1007/978-3-032-29000-7_21

[10] S. Chakraborty, F. Heintz, *Enhancing time series forecasting with fuzzy attention-integrated transformers*, arXiv preprint, (2025). https://doi.org/10.48550/arXiv.2504.00070

[11] H. Chen, et al., *Enhancing educational Q&A systems using a Chaotic Fuzzy Logic-Augmented large language model*, Frontiers in Artificial Intelligence, **7** (2024), 1404940. https://doi.org/10.3389/frai.2024.1404940

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

[13] A. Chikhalikar, et al., *Open vocabulary object search utilizing large language models and fuzzy inferencing*, 2025 IEEE/SICE International Symposium on System Integration (SII), (2025). https://doi.org/10.1109/SII59315.2025.10870891

[14] D. Dimitrov, *Towards more reliable SQL auto-grading: A hybrid approach Using LLMs, intuitionistic fuzzy sets, and traditional methods*, International Conference on Flexible Query Answering Systems, (2025), 253-264. https://doi.org/10.1007/978-3-032-05607-8_24

[15] Y. Dong, T. Ito, *A linguistic negotiation for consensus-making among LLM agents with fuzzy utility functions*, IEICE Transactions on Information and Systems, **E109.D**(7) (2026), 1112-1122. https://doi.org/10.1587/transinf.2025EDP7179

[16] M. B. Dowlatshahi, S. Beiranvand, *Optimizing deep Q-networks with fuzzy inference-based adaptive replay buffer management*, Iranian Journal of Fuzzy Systems, **22**(4) (2025), 161-174. https://doi.org/10.22111/ijfs.2025.9358

[17] Y. Du, *Intuitionistic fuzzy sets for large language model data annotation: A novel approach to side-by-side preference labeling*, arXiv preprint, (2025). https://doi.org/10.48550/arXiv.2505.24199

[18] Y. Fang, et al., *Advancing Arabic sentiment analysis: ArSen benchmark and the improved fuzzy deep hybrid network*, Proceedings of the 28th Conference on Computational Natural Language Learning, (2024), 507-516. https://doi.org/10.18653/v1/2024.conll-1.39

[19] F. Fanian, M. Kuchaki Rafsanjani, A. Borumand Saeid, *An expert-aware intelligent multi-phase protocol: Metaheuristic-fuzzy-guided machine learning for enhanced generalizability in smart WRSNs*, Journal of Network and Computer Applications, **236** (2026), 104464. https://doi.org/10.1016/j.jnca.2026.104464

[20] F. Fanian, M. Kuchaki Rafsanjani, M. Shokouhifar, *Combined fuzzy-metaheuristic framework for bridge health monitoring using UAV-enabled rechargeable wireless sensor networks*, Applied Soft Computing, **167** (2024), 112429. https://doi.org/10.1016/j.asoc.2024.112429

[21] V. Figueiredo, *Fuzzy, symbolic, and contextual: Enhancing LLM instruction via cognitive scaffolding*, arXiv preprint, (2025). https://doi.org/10.48550/arXiv.2508.21204

[22] J. F. Gaitán-Guerrero, et al., *A novel fine-tuning and evaluation methodology for large language models on IoT raw data summaries (LLM-RawDMeth): A joint perspective in diabetes care*, Computer Methods and Programs in Biomedicine, **269** (2025), 108878. https://doi.org/10.1016/j.cmpb.2025.108878

[23] R. Gauraha, A. K. Agrawal, P. Dubey, *Hybrid transformer–fuzzy framework for interpretable sentiment classification in deepfake social media content*, Scientific Reports, (2026). https://doi.org/10.1038/s41598-026-58095-9

[24] C. Geng, *An intelligent framework combining deep learning and fuzzy logic for accurate remote language translation*, Scientific Reports, **15**(1) (2025), 38736. https://doi.org/10.1038/s41598-025-22549-3

[25] S. Gilda, S. Gilda, *AI-assisted engineering should track the epistemic status and temporal validity of architectural decisions*, arXiv preprint, (2026). https://doi.org/10.48550/arXiv.2601.21116

[26] S. Gorle, S. B. S. Rao, P. Muthusamy, *Detecting and mitigating hallucinations in large language models (LLMs) using reinforcement learning in healthcare*, Journal of AI-Powered Medical Innovations, **1**(1) (2024), 105-118. https://doi.org/10.60087/Japmi.Vol.03.Issue.01.Id.011

[27] W. Gu, et al., *FPSO-Time: A fuzzy time series forecasting method based on fuzzy particle swarm optimization and reprogrammed large language models*, IEEE Transactions on Fuzzy Systems, **34**(2) (2026). https://doi.org/10.1109/TFUZZ.2025.3637018

[28] B. Haznedar, L. Karacan, *FISformer: Replacing self-attention with a fuzzy inference system in transformer models for time series forecasting*, IEEE Transactions on Fuzzy Systems, (2026). https://doi.org/10.1109/TFUZZ.2026.3690012

[29] X. Hu, et al., *Fuzzy symbolic reasoning for few-shot KBQA: A CBR-inspired generative approach*, International Conference on Case-Based Reasoning Research and Development, (2025), 96-110. https://doi.org/10.1007/978-3-031-96559-3_7

[30] Y. S. Hu, S. Marandi, M. Modarres, *DML–LLM hybrid architecture for fault detection and diagnosis in sensor-rich industrial systems*, Sensors, **26**(6) (2026), 2008. https://doi.org/10.3390/s26062008

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

[32] S. Kadavath, et al., *Language models (mostly) know what they know*, arXiv preprint, (2022). https://doi.org/10.48550/arXiv.2207.05221

[33] M. Kamal, et al., *Physical fuzzy rule based unsupervised news article clustering*, 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, (2025). https://doi.org/10.1109/COMPSAC65507.2025.00218

[34] T. Kanagawa, *Deterministic compliance failures in large language models: A structural analysis using a legacy fuzzy inference benchmark*, Engineering Engrxiv Archive, (2026). https://doi.org/10.31224/6702

[35] J. Kaplan, et al., *Scaling laws for neural language models*, arXiv preprint, (2020). https://doi.org/10.48550/arXiv.2001.08361

[36] J. Kim, et al., *Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models*, Proceedings of the IEEE/CVF International Conference on Computer Vision, (2025). https://doi.org/10.1109/ICCV51701.2025.01913

[37] M. Košprdić, et al., *VerifAI: A verifiable open-source search engine for biomedical question answering*, IEEE Access, **14** (2026), 45129-45147. https://doi.org/10.13140/RG.2.2.14034.82888

[38] J. Kralev, *Fine-tuning LLMs for real-time fuzzy insulin control in Type I diabetes*, Medicine and Pharmacology, Endocrinology and Metabolism, (2026). https://doi.org/10.20944/preprints202604.1316.v1

[39] M. Kumari, A. Chaudhary, Y. Narayan, *Explainable AI (XAI): A survey of current and future opportunities*, Explainable Edge AI: A Futuristic Computing Perspective, Cham: Springer International Publishing, (2022), 53-71. https://doi.org/10.1007/978-3-031-18292-1_4

[40] B. Kwon, K. Yu, *Cognition-enhanced geospatial decision framework integrating fuzzy FCA, surprisingly popular method, and a large language model*, Scientific Reports, **15**(1) (2025), 23089. https://doi.org/10.1038/s41598-025-06508-6

[41] M. Li, et al., *Large language model assisted hyper-heuristic evolutionary algorithm for groundwater level prediction*, Scientific Reports, **16** (2026). https://doi.org/10.1038/s41598-026-52801-3

[42] S. Lin, J. Hilton, O. Evans, *TruthfulQA: Measuring how models mimic human falsehoods*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), **1** (2022), 3214-3252. https://aclanthology.org/2022.acl-long.229.pdf

[43] R. Lin, Y. You, *Let the fuzzy rule speak: Enhancing in-context learning debiasing with interpretability*, arXiv preprint, (2024). https://doi.org/10.48550/arXiv.2412.19018

[44] L. Liu, *Multi-hop reasoning and retrieval in embedding space: Leveraging large language models with knowledge*, arXiv preprint, (2026). https://doi.org/10.48550/arXiv.2603.13266

[45] A. Madaan, et al., *Self-refine: Iterative refinement with self-feedback*, in Advances in Neural Information Processing Systems (NeurIPS), (2019), 46534-46594.

[46] A. Malinin, M. Gales, *Uncertainty estimation in autoregressive structured prediction*, in International Conference on Learning Representations (ICLR), (2020). https://doi.org/10.48550/arXiv.2002.07650

[47] E. H. Mamdani, S. Assilian, *An experiment in linguistic synthesis with a fuzzy logic controller*, International Journal of Human-Computer Studies, **51**(2) (1999), 135-147. https://doi.org/10.1006/ijhc.1973.0303

[48] S. Manolache, N. Popescu, *Incorporating fuzzy logic into an OpenAI-based decision -making system*, UPB Scientific Bulletin, Series C: Electrical Engineering and Computer Science, **87**(4) (2025).

[49] C. Molnar, *Interpretable machine learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[50] P. Mortezaagha, A. Rahgozar, *An auditable pipeline for fuzzy full-text screening in systematic reviews: Integrating contrastive semantic highlighting and LLM judgment*, Artificial Intelligence Review, (2026). https://doi.org/10.1007/s10462-026-11599-2

[51] A. Mukherjee, S. Das, *ChatGPT: A fuzzy system that talks like a human*, Journal of Mathematical Sciences and Computational Mathematics, **5**(3) (2024). https://doi.org/10.15864/jmscm.5303

[52] T. M. Nguyen, T. T. Nguyen, T. T. Quan, *A two-stage fuzzy-guided genetic algorithm for university timetabling with LLM-based preference parsing*, International Conference on Multi-disciplinary Trends in Artificial Intelligence, Singapore: Springer Nature Singapore, (2025), 358-370. https://doi.org/10.1007/978-981-95-4960-3_29

[53] OpenAI, *Introducing ChatGPT*, OpenAI Blog, Nov. 2022. [Online]. Available: https://openai.com/blog/chatgpt

[54] O. Orang, et al., *Causal graph fuzzy LLMs: A first introduction and applications in time series forecasting*, Conference: Congresso Brasileiro de Inteligência Computacional, (2025). https://doi.org/10.21528/CBIC2025-1187477

[55] E. Özbay, F. A. Özbay, A. B. Özer, *A unified AI framework for Turkish E-commerce review analysis: Sentiment classification, LLM-based summarization, and fuzzy evaluation*, Applied Sciences, **16**(12) (2026), 5849. https://doi.org/10.3390/app16125849

[56] S. Qahtan, et al., *Three-way decision approach based on utility and dynamic localization transformational procedures within a circular q-rung orthopair fuzzy set for ranking and grading large language models*, Cognitive Computation, **17**(2) (2025), 77. https://doi.org/10.1007/s12559-025-10432-2

[57] R. Radišić, S. Popov, N. Ralević, *Large language model and fuzzy metric integration in assignment grading for introduction to programming type of courses*, Mathematics, **14**(1) (2025), 137. https://doi.org/10.3390/math14010137

[58] J. Ren, I. Jenkinson, O. Tobora, *Grounded expert systems for offshore safety: Enhancing rule-based risk assessment with retrieval-augmented generation (RAG) for auditable explanations*, Conference: International Conference on Electronics Technology (ICET), (2026).

[59] M. T. Ribeiro, S. Singh, C. Guestrin, *"Why should I trust you?" Explaining the predictions of any classifier*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), (2016), 1135-1144. https://doi.org/10.1145/2939672.2939778

[60] A. Saeedi, M. Kuchaki Rafsanjani, S. Yazdani, *Energy efficient clustering in IoT-based wireless sensor networks using binary whale optimization algorithm and fuzzy inference system*, The Journal of Supercomputing, **81**(1) (2025), 209. https://doi.org/10.1007/s11227-024-06556-1

[61] M. Sajid, et al., *Fuzzy learning at 60: Future of trustworthy AI in healthcare and LLM*, Conference: 2025 IEEE International Conference on Fuzzy Systems - Celebrating 60 Years of Fuzzy Sets, (2025).

[62] C. Salamea-Palacios, W. Martel-Sócala, E. Arcos-Salamea, *Development of a sentiment analysis system for Chatbot responses enhanced with fuzzy logic*, International Conference on Information Technology and Systems, (2025), 71-80. https://doi.org/10.1007/978-3-031-93106-2_7

[63] B. Saulnier, W. Oettgen, *Fuzzy-cognitive integration and LLM-assisted interpretation: Toward traceable and cognitively faithful AI explanations*, ResearchGate, (2024). https://doi.org/10.5281/zenodo.17386634

[64] R. Schuerkamp, P. J. Giabbanelli, *Guiding evolutionary algorithms with large language models to learn fuzzy cognitive maps*, Neural Computing and Applications, **37**(18) (2025), 11891-11908. https://doi.org/10.1007/s00521-025-11157-x

[65] K. A. Shaik, et al., *A fuzzy supervisory framework for real-time optimization of robot output and LLM performance in HRI*, 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), (2025). https://doi.org/10.1109/HRI61500.2025.10974222

[66] J. Song, et al., *An uncertainty-aware framework integrating large language model and fuzzy inference system for commonsense reasoning*, Expert Systems with Applications, **310** (2026), 131273. https://doi.org/10.1016/j.eswa.2026.131273

[67] T. Stefanova, S. Georgiev, *Fuzzy ensemble of large language models for financial sentiment analysis*, International Conference on Intelligent and Fuzzy Systems, Cham: Springer Nature Switzerland, (2025), 553-565. https://doi.org/10.1007/978-3-031-97992-7_62

[68] W. W. Tan, T. W. Chua, *Uncertain rule-based fuzzy logic systems: Introduction and new directions (Mendel, JM; 2001) [book review]*, IEEE Computational intelligence magazine, **2**(1) (2007), 72-73. https://doi.org/10.1109/MCI.2007.357196

[69] H. Touvron, et al., *LLaMA: Open and efficient foundation language models*, arXiv preprint, (2023). https://doi.org/10.48550/arXiv.2302.13971

[70] K. Trinkūnas, J. Miliauskaité, *Fuzzy-based assessment of stakeholder feedback in software requirements*, New Trends in Computer Sciences, **3**(2) (2025), 154-172. https://doi.org/10.3846/ntcs.2025.26832

[71] H. Tripathi, et al., *From protocol to practice: Graded sepsis bundle compliance and actionable insights from real-world ICU data*, MedRxiv, (2026), 2026-04. https://doi.org/10.64898/2026.04.23.26351412

[72] A. Vaswani, et al., *Attention is all you need*, in Advances in Neural Information Processing Systems (NeurIPS), (2017), 5998-6008.

[73] *Visual-linguistic abductive reasoning (ViLA): Integrating fuzzy scoring for multi-modal reasoning*, 2026. (P063).

[74] R. Wajman, *A hybrid fuzzy-sentiment framework for adaptive two-phase flow control in human-centric systems*, Advances in Science and Technology Research Journal, **19**(12) (2025), 42-55. https://doi.org/10.12913/22998624/209579

[75] H. Wang, et al., *Collm: Industrial large-small model collaboration with fuzzy decision-making agent and self-reflection*, IEEE Transactions on Fuzzy Systems, **34**(4) (2025). https://doi.org/10.1109/TFUZZ.2025.3594229

[76] S. Wang, et al., *Fuzzy-assisted contrastive decoding improving code generation of large language models*, IEEE Transactions on Fuzzy Systems, **33**(8) (2025). https://doi.org/10.1109/TFUZZ.2025.3575060

[77] M. Wang, et al., *SAGE: Global semantic alignment with LLMs for long-tail sequential recommendation*, Proceedings of the ACM Web Conference, (2026), 6433-6444. https://doi.org/10.1145/3774904.3792456

[78] Z. Wu, *Neural fuzzy logic reasoning for natural language inference*, A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science, (2022). https://ualberta.scholaris.ca/server/api/core/bitstreams/748855bc-09b2-42bc-ae86-e3984a4e887c/content

[79] S. Wu, *Fuzzy retrieval of power dispatching knowledge base through large language model integrated with knowledge graph*, International Journal of Information Technology and Management, **24**(3-4) (2025), 237-258. https://doi.org/10.1504/ijitm.2025.151550

[80] H. Wu, et al., *FLAIR: Steering LLM mathematical problem solving based on a fuzzy-logic-assIsted rReasoner*, Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), (2026). https://doi.org/10.18653/v1/2026.acl-long.1790

[81] Y. Xiang, et al., *Large language model for secure operation of power systems*, Smart Energy System Research, **1**(1) (2025), 10005. https://doi.org/10.70322/sesr.2025.10005

[82] T. Xu, X. S. Tang, *Electrical equipment fault diagnosis: A technique combining fuzzy logic and large language models*, 2023 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-ASIA), (2023). https://doi.org/10.1109/ISPCE-ASIA60405.2023.10365878

[83] S. Yadav, N. K. Verma, *Explainable AI for sentiment classification with Type-1 fuzzy and Type-2 fuzzy models*, IEEE Transactions on Fuzzy Systems, **34**(7) (2026), 2239-2251. https://doi.org/10.1109/TFUZZ.2026.3688623

[84] W. Yan, C. Zhang, K. Yamada, *Fuzzmonte-rag: A hybrid fuzzy logic and monte carlo approach for personalized learning optimisation in llm-assisted it education*, Bulletin of Advanced Institute of Industrial Technology, **18** (2025). https://aiit.ac.jp/documents/jp/research_collab/research/bulletin/18th/03_Yan.pdf

[85] H. Yang, et al., *Generative fuzzy system for sequence generation*, arXiv preprint, (2024). https://doi.org/10.48550/arXiv.2411.1386

[86] T. Yao, et al., *Multi-agent fuzzy reinforcement learning with LLM for cooperative navigation of endovascular robotics*, IEEE Transactions on Fuzzy Systems, **34**(4) (2026), 1109-1119. https://doi.org/10.1109/TFUZZ.2025.3585934

[87] T. Yoshida, Y. Sueoka, K. Osuka, *Verification of a two-step inference model for cooperative evaluation of robot actions using foundation models*, International Symposium on Distributed Autonomous Robotic Systems, Cham: Springer Nature Switzerland, (2024), 395-410. https://doi.org/10.1007/978-3-032-04584-3_27

[88] L. Yousofvand, M. B. Dowlatshahi, M. Pirdadeh Beiranvand, *Pneumonia detection in chest X-ray images using Convolutional Neural Network and fuzzy VIKOR*, Iranian Journal of Fuzzy Systems, **22**(5) (2025), 159-179. https://doi.org/10.22111/ijfs.2025.51603.9119

[89] L. A. Zadeh, *Fuzzy sets*, Information and Control, **8**(3) (1965), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

[90] L. A. Zadeh, *The concept of a linguistic variable and its application to approximate reasoning*, Information Sciences, **8**(3) (1975), 199-249.https://doi.org/10.1016/0020-0255(75)90036-5

[91] B. Zhang, et al., *Large language model enhanced fuzzy logic fusion framework for stance detection*, CSIG Conference on Emotional Intelligence, Singapore: Springer Nature Singapore, (2024), 130-144. https://doi.org/10.1007/978-981-96-5084-2_9

[92] Y. Zhang, et al., *Siren's song in the AI ocean: A survey on hallucination in large language models*, Computational Linguistics, **51**(4), (2025). https://doi.org/10.1162/coli.a.16

[93] M. Zhang, et al., *Bridging stochasticity and fuzziness: Automated construction of triangular fuzzy numbers via LLM temperature sampling for managerial decision support*, Information, **17**(4) (2026), 349. https://doi.org/10.3390/info17040349

[94] Z. Zhang, C. Valeo, *Fuzzy-based input method for uncertainty quantification in a deterministic model comparison with ChatGPT for peak flow prediction*, Journal of Hydrology X, **28-29** (2025), 100208. https://doi.org/10.1016/j.hydroa.2025.100208

[95] W. Zheng, et al., *Llm-as-a-fuzzy-judge: Fine-tuning large language models as a clinical evaluation judge with fuzzy logic*, arXiv preprint, (2025). https://doi.org/10.48550/arXiv.2506.11221