Interval-Valued Epistemic Reasoning for Hallucination-Aware Language Models

Taniya Seth, Pranab K. Muhuri
Department of Computer Science and Engineering, Faculty of Engineering and Technology, South Asian University
New Delhi, India
taniyaseth@sau.int, pranabmuhuri@cs.sau.ac.in

***Abstract*— LLMs tend to generate fluent but epistemically dubious responses, which can be observed as hallucinations, because of the lack of a principled representation of truth and uncertainty. This paper presents a novel type-2 fuzzy epistemic control framework for modelling truth in language generation as an interval-valued variable, representing uncertainty about uncertainty, rather than confidence values. The internal model signals of token entropy and logit margin are transformed using an interpretable fuzzy inference system to obtain the bounds of epistemic truth without any retraining or parameter modification of the model. Experiments on the TruthfulQA dataset with various transformer-based models show that the proposed method conservatively controls epistemic trust, distinguishes model reliability, and resists false confidence with weak evidence, outperforming entropy-based and scalar uncertainty methods. The proposed method thus validates type-2 fuzzy logic as a sound and interpretable basis for hallucination-robust and truth-sensitive language modelling.**

Keywords—Epistemic uncertainty, large language model, hallucination mitigation, truth-aware language generation, type-2 fuzzy sets

I. Introduction

Large Language Models have made significant advances in natural language understanding and generation tasks and have reached state-of-the-art performance on a broad spectrum of tasks, including question answering, summarization, and dialogue generation [1], [2]. Despite their impressive fluency and coherence, LLMs tend to generate factually incorrect or unsupported text, a phenomenon commonly referred to as hallucination [3], [4]. More alarmingly, hallucinated text is often generated with high confidence, making them hard to distinguish and potentially dangerous in safety-critical applications such as healthcare, law, and scientific inquiry [5].

A great deal of recent research has been devoted to the mitigation of hallucinations via probabilistic uncertainty modelling, retrieval-augmented generation (RAG), and post-hoc verification methods [6]–[8]. Although these methods improve factual grounding in specific settings, they essentially depend on scalar confidence scores calculated from likelihoods or SoftMax probabilities. These scores essentially convey distributional uncertainty but not epistemic uncertainty and, hence, are incapable of separating linguistic validity from truth. Consequently, the models can be persistently but erroneously confident when the training data is limited, inconsistent, or simply non-existent for a particular query [3], [9].

From an epistemological standpoint, hallucination in language models is more than a statistical phenomenon and is instead a result of reasoning under incomplete and ambiguous information. Truth in natural language is fundamentally graded, context-dependent, and often uncertain, making binary or scalar probabilistic quantification suboptimal. Fuzzy logic is a rigorous mathematical formalism for modelling graded truth and reasoning with imprecision, especially when uncertainty is only partially characterizable probabilistically [10], [11]. Specifically, type-2 fuzzy logic (T2 FL) generalizes classical fuzzy sets by modelling uncertainty about membership functions themselves, allowing for the representation of uncertainty about uncertainty, which is particularly important for epistemic reasoning [12], [13].

While fuzzy logic has been successfully used in decision-making, control systems, and explainable artificial intelligence, its combination with state-of-the-art neural language models is still an uncharted area. Current approaches to uncertainty-aware NLP usually model uncertainty as an auxiliary scalar signal, rather than a first-class epistemic concept. This leads to a lack of interpretable and theoretically justified mechanisms for trust regulation in language model outputs based on the quality of internal evidence rather than surface-level fluency.

In this paper, we introduce a new T2 fuzzy epistemic control framework for large language models that models truth as an interval-valued variable instead of a scalar confidence measure. We also reframe language generation as an epistemic reasoning task where internal model signals, namely token-level entropy and logit margin, are used as fuzzy indicators of epistemic truth. These signals are then processed by an interpretable fuzzy inference system with a clear rule base to produce interval-valued epistemic truths that represent uncertainty about uncertainty. The proposed framework is a non-intrusive, model-agnostic control module that does not require retraining, fine-tuning, or modifying the underlying language model architecture.

We also empirically assess the proposed approach on the TruthfulQA benchmark [3], a popular dataset created to specifically test the hallucination behaviour of language models. Experiments carried out on a variety of transformer-based models show that the proposed fuzzy epistemic controller successfully distinguishes between model confidence, conservatively estimates truth intervals under suboptimal conditions, and prevents false confidence in stylistically fluent but epistemically dubious text. Moreover, interval-valued truth modelling is more stable and interpretable than entropy-based and scalar uncertainty methods, underlining the importance of T2 FL for reliable language generation.

Through the integration of fuzzy logic and modern language modelling, this work provides a solid theoretical basis for hallucination-resilient and truth-aware natural language processing. The proposed framework improves our theoretical understanding of epistemic uncertainty in neural language models and facilitates the development of interpretable control strategies for epistemic trust in generated text.

II. Proposed Framework

A. Epistemic View of Language Generation

We reinterpret language generation in large language models as an epistemic reasoning process rather than a purely probabilistic sampling process. For a given input prompt x, a language model generates a response y by iteratively sampling tokens from a conditional distribution. Although likelihood-based decoding is able to capture distributional preferences learned during training, it lacks a principled concept of truth or epistemic validity. As noted in previous work, high-probability responses can still be considered hallucinated or factually incorrect responses when epistemic evidence is weak or absent [3], [9].

To overcome this limitation, we assign each generated response an epistemic truth assessment that captures the level of the model's internal evidence supporting the generated response. Instead of modelling truth as a scalar probability measure, we model it as an interval-valued quantity, allowing for the explicit modelling of uncertainty about uncertainty. This modelling choice is naturally compatible with fuzzy logic and, more specifically, with T2 fuzzy systems, which are specifically tailored to deal with imprecision in membership function definitions [12], [13].

B. Epistemic Signal Extraction from Language Models

Let $y = \{w_1, w_2, \dots, w_T\}$ denote a generated token sequence. During generation, modern transformer-based language models expose internal signals that can be interpreted as indirect indicators of epistemic evidence. In this work, we focus on two such signals:

1. *Token Entropy:* Token entropy measures the dispersion of the predictive distribution at each decoding step and is computed as:

$$H_t = -\sum_i p(w_i \mid x, w_{<t}) \log p(w_i \mid x, w_{<t}).$$

   High entropy indicates uncertainty in token selection, which may arise from ambiguity, lack of knowledge, or conflicting evidence [8], [9].

2. *Logit Margin:* The logit margin is calculated as the difference between the maximum and second-largest logits during the decoding process. A small margin corresponds to a weak preference for tokens, while a large margin corresponds to a stronger preference for the chosen token. Logit margins have been found to relate to confidence and robustness in models [8].

3. These measures are not considered probabilistic indicators of truth; rather, they are considered fallible epistemic indicators that cumulatively indicate the truthfulness of the produced response.

These signals are not treated as probabilistic truth measures; instead, they are interpreted as imperfect epistemic cues that jointly inform the reliability of the generated response.

C. Linguistic Variables and Fuzzy Representation

To enable interpretable reasoning, the extracted epistemic signals are mapped to linguistic variables. Let H denote token entropy and M denote logit margin. We define the following linguistic terms:

- Entropy: $\{Low, Medium, High\}$
- Margin: $\{Weak, Moderate, Strong\}$

Each linguistic variable is represented using trapezoidal membership functions, as commonly followed in fuzzy modelling [10], [11]. The output variable, $Epistemic\ Truth$, is defined over the unit interval $[0,1]$ with linguistic terms $\{Low, Medium, High\}$. This representation allows for degreeed truth assignments while maintaining interpretability.

D. Type-2 Fuzzy Epistemic Modelling

T1 fuzzy systems provide crisp membership values, which are inadequate for capturing higher-order uncertainties inherent in epistemic reasoning. T2 fuzzy sets (FS) model uncertainty by representing uncertainty in membership functions, allowing for the modelling of uncertainty about uncertainty [12].

In this framework, T2 behaviour is approximated through interval-valued truth assignments. Specifically, the fuzzy inference system produces a central truth value $T_c \in [0,1]$, which is subsequently expanded into an epistemic truth interval:

$$\tilde{T} = [\underline{T}, \overline{T}],$$

where the interval width is adaptively modulated by epistemic uncertainty. This approach follows established interval Type-2 fuzzy modelling principles while remaining computationally tractable [13].

The interval width is expanded for high entropy and weak margin values, indicating high epistemic uncertainty, and shrinks for high internal evidence. This allows for conservative reasoning, which avoids overconfidence in uncertain generations.

E. Fuzzy Rule Base and Inference Mechanism

Epistemic reasoning is implemented through an interpretable fuzzy rule base of the form:

```
IF entropy is High AND margin is Weak
    THEN epistemic truth is Low

IF entropy is Low AND margin is Strong
    THEN epistemic truth is High
```

The full rule base covers the entire epistemic input space, including a catch-all rule that assigns medium truth in epistemic truth ambiguity. Fuzzy inference is done using the Mamdani fuzzy inference framework with centroid defuzzification, which is widely used in fuzzy control systems because of its interpretability and numerical stability [11].

Fig. 1 shows the full fuzzy epistemic rule base by visualizing the mapping from token entropy and logit margin to epistemic truth. The surface shows smooth and monotonic behaviour, assigning high truth for low uncertainty and strong evidence, and conservative truth for high uncertainty and weak evidence. This visualization provides a global, interpretable representation of the fuzzy inference mechanism controlling epistemic behaviour.

To ensure robustness, each inference is done independently, without state leakage and with guaranteed determinism per sample.

F. Epistemic Control Interpretation

The proposed framework acts as a non-intrusive epistemic control layer that can be applied to any pre-trained language model irrespective of its architecture and parameters. The

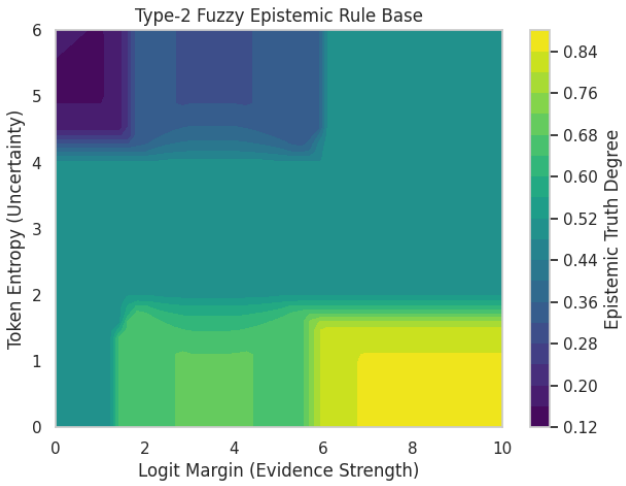


Fig. 1 Continuous epistemic behaviour of the full T2 fuzzy system

truth intervals obtained can be used to control the behaviours of response acceptance, confidence reporting, or generation suppression based on high epistemic uncertainty.

Notably, the approach does not seek to maximize truth scores but instead follows epistemic conservatism, reporting weak or low truth when evidence is weak. This is essential for hallucination mitigation and is in line with the principles of trustworthy AI and explainable decision-making [5], [10].

III. Experimental Evaluation

The goal of the experimental evaluation is to determine whether the proposed T2 fuzzy epistemic control framework is capable of providing meaningful, interpretable, and epistemically conservative assessments of epistemic truth for large language models. The experimental evaluation aims to determine whether the framework (i) distinguishes epistemically with respect to models of different capacity, (ii) models uncertainty about uncertainty using interval-valued truth, (iii) resists false confidence in the presence of weak epistemic evidence, and (iv) offers improvements over traditional scalar uncertainty measures. All experiments are performed without retraining, fine-tuning, or modifying the underlying language models to ensure a fair and model-agnostic evaluation.

A. Dataset

The experimental evaluation is performed on the TruthfulQA benchmark dataset, which is a dataset specifically designed to test the hallucination behaviour of large language models. The TruthfulQA dataset is comprised of questions that are likely to receive confident but incorrect answers based on misconceptions, lack of knowledge, or spurious correlations. Unlike typical question-answering datasets, the TruthfulQA dataset is designed to test epistemic failure modes rather than task accuracy, making it particularly well-suited to the evaluation of epistemic reliability and trustworthiness.

The evaluation is performed on the generation split of the validation set and a random sample of 200 questions to ensure computational tractability while maintaining diversity. Each question is presented in natural language form, and the answers are generated through standard decoding without retrieval augmentation or post-hoc verification.

B. Language Models

For analyzing model-agnostic behaviour, we test the proposed framework on three pretrained transformer language models with varying architectures and capabilities:

- M1: FLAN-T5-Base
- M2: FLAN-T5-Large
- M3: BART-Large

These models vary in instruction tuning, parameter size, and stylistic expressiveness. The models are used in a frozen configuration with the same set of prompts and decoding hyperparameters. No prompt tuning, hyperparameter adjustment, or architectural changes are made, ensuring that any differences are solely due to epistemic reasoning and not training performance.

C. Epistemic Signal Extraction

During text generation, we extract internal model signals that serve as epistemic indicators:

1. Token Entropy
2. Logit Margin

These signals are not used as confidence measures but are instead treated as noisy epistemic signals that provide collective information to the fuzzy inference system.

D. Fuzzy Epistemic Inference Setup

The epistemic signals are converted to linguistic variables with trapezoidal membership functions. A Mamdani-type fuzzy inference system is used with an interpretable rule set that encodes epistemic reasoning patterns such as high uncertainty with weak evidence indicating low epistemic truth.

To simulate T2 fuzzy systems in a computationally efficient way, the fuzzy inference system outputs an interval-valued epistemic truth. The midpoint of the interval is calculated using centroid defuzzification, and the interval width is adaptively adjusted based on both uncertainty (entropy) and evidence strength (logit margin). The interval captures uncertainty about uncertainty and allows for conservative epistemic reasoning during weak evidence situations.

Each fuzzy inference is done independently to prevent state information leakage and ensure deterministic behaviour for each sample.

E. Baselines

We contrast our proposed framework with the following baselines, all of which are either directly or indirectly represented in the experimental design:

1. *Entropy-Based Uncertainty:* Scalar uncertainty derived directly from token entropy, representing a commonly used uncertainty proxy in neural language models.
2. *T1 Truth Approximation:* A deterministic scalar truth score computed as a normalized inverse of entropy. This baseline corresponds to a simple Type-1 fuzzy-style mapping without interval-valued uncertainty.
3. *Model Capacity Heuristic (Implicit):* Common assumptions that larger or more fluent models are inherently more reliable are implicitly evaluated

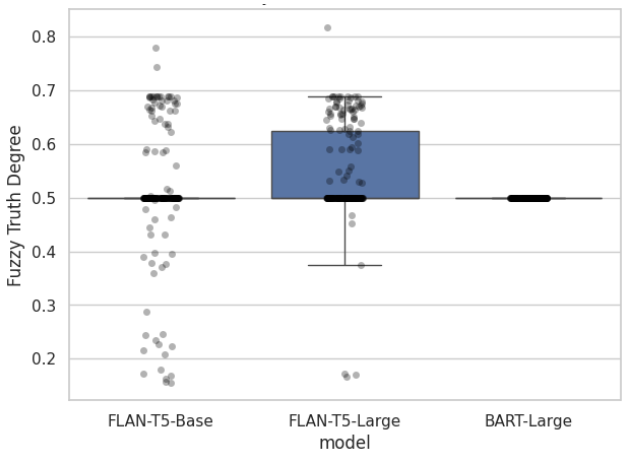


Fig. 2 Cross-model truth distribution (Non-Parametric)

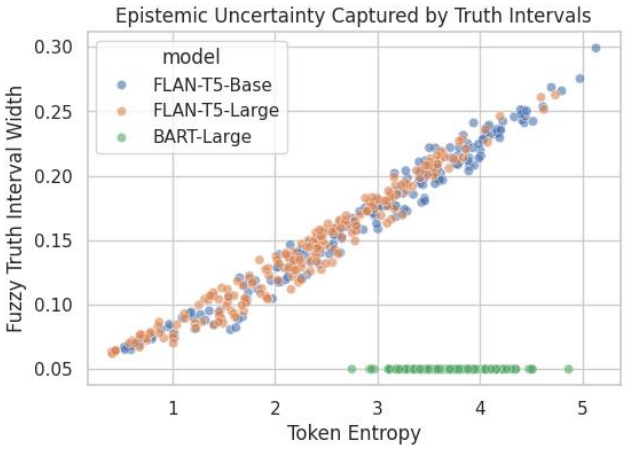


Fig. 3 Token entropy vs fuzzy truth interval width

through cross-model comparisons under identical decoding conditions.

These baselines reflect standard uncertainty heuristics used in practice but lack explicit epistemic semantics or interpretability.

F. Evaluation Metrics

As epistemic truth is not a directly observable quantity, evaluation is based on stability, interpretability, and conservativeness rather than accuracy. The following metrics are reported:

- *Mean Epistemic Truth:* Average centre of the truth interval across samples.
- *Truth Variance:* Variance of epistemic truth assignments.
- *Mean Truth Interval Width:* Average epistemic uncertainty.
- *Entropy–Truth Relationship:* Relationship between token entropy and truth interval width.

These metrics together provide a comprehensive assessment of whether the framework is producing coherent and epistemically meaningful behaviour.

IV. Results and Analysis

Results obtained from the above evaluation are presented below.

A. Cross-Model Epistemic Truth Distribution

In Fig. 2, the non-parametric distribution of epistemic truth centres is shown for the evaluated language models. There are obvious and systematic differences among the model architectures. FLAN-T5-Large always has higher epistemic truth with lower variance, which shows more reliable internal evidence for generation. FLAN-T5-Base, on the other hand, has a higher dispersion of epistemic truth, which shows higher epistemic instability and sensitivity to uncertainty.

Notably, BART-Large settles almost completely to epistemic neutrality, with truth values centered around 0.5 and near-zero variance. This observation suggests that, despite its ability to generate fluent outputs, the underlying epistemic signals inferred from BART-Large, namely entropy and logit margin, lack discriminative evidence to support confident truth assignments. Instead, the proposed framework properly defaults to epistemic conservatism. This observation underscores an important advantage of the proposed framework: truth is decoupled from stylistic fluency and model likelihood, precluding unjustified trust in superficially confident outputs.

B. Interval-valued Epistemic Uncertainty

Fig. 3 plots the relationship between token entropy and fuzzy truth interval width. For all models, truth interval width monotonically increases with entropy, validating that the proposed framework properly captures uncertainty about uncertainty, a hallmark of T2 FL.

For FLAN-T5-Base, truth interval widths grow rapidly with increasing entropy, suggesting that epistemic uncertainty is high in conditions of ambiguous or weakly supported generation. FLAN-T5-Large exhibits similar behaviour but with more strongly clustered intervals, suggesting that stronger internal evidence and improved epistemic support are provided. By contrast, BART-Large shows strongly narrow truth intervals even for high entropy values, indicating that high entropy is not sufficient to support epistemic confidence. This observation further supports that the proposed framework does not simply depend on a single uncertainty measure but instead properly combines multiple epistemic cues to conservatively regulate truth assignments.

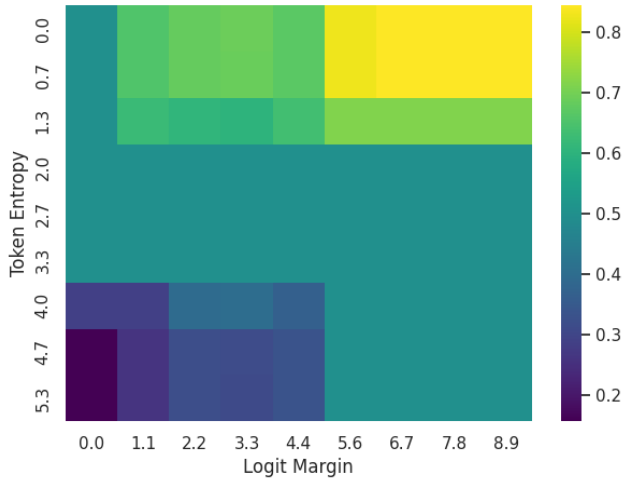


Fig. 4 Fuzzy epistemic rule activation heatmap

TABLE I Quantitative behaviour of models

Model	Truth Mean	Truth Var	Mean Truth Width	Mean Entropy
M3	0.5000	2.136911e-32	0.0500	3.753495
M1	0.5089	1.286682e-02	0.1661	2.824089
M2	0.5435	7.867750e-03	0.1430	2.321517

C. *Fuzzy Epistemic Control Surface and Rule Behaviour*

Fig. 4 illustrates the complete fuzzy epistemic rule base as a smooth control surface in the entropy-margin space. The control surface has a smooth, monotonic shape: epistemic truth increases with high evidence (large logit margin) and decreases with high uncertainty (large entropy). There are no sharp corners or discontinuities, suggesting that epistemic control arises from smooth fuzzy reasoning rather than hard margins.

The complementary rule activation heatmaps further validate that the rule base provides full coverage of the epistemic input space. The areas of high uncertainty and low evidence are consistently assigned low epistemic truth, while areas of low uncertainty and high evidence are assigned high truth values. The remaining areas represent epistemic ambiguity, which are assigned medium truth values. This pattern validates that the fuzzy rule base is logically sound, interpretable, and epistemically well-grounded.

D. *Quantitative Summary and Stability Analysis*

Table I summarizes the quantitative analysis of the proposed framework. FLAN-T5-Large has the highest mean epistemic truth with lower variance and smaller truth intervals, indicating epistemic stability. FLAN-T5-Base has lower mean truth and much larger variance, indicating higher epistemic variability. BART-Large has minimal variance and small intervals centered around epistemic neutrality, indicating conservative epistemic behaviour under low evidence.

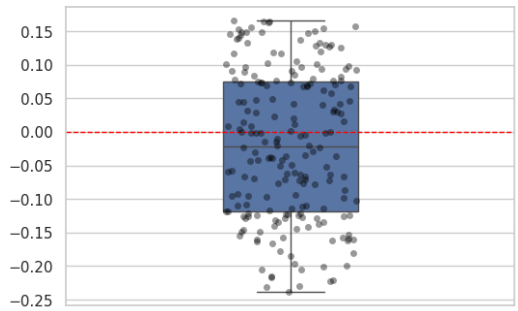
These findings together demonstrate that the proposed framework can distinguish epistemic trustworthiness among models without any retraining, calibration, or task-specific optimization, which further emphasizes its model-agnostic nature.

E. *Statistical Significance*

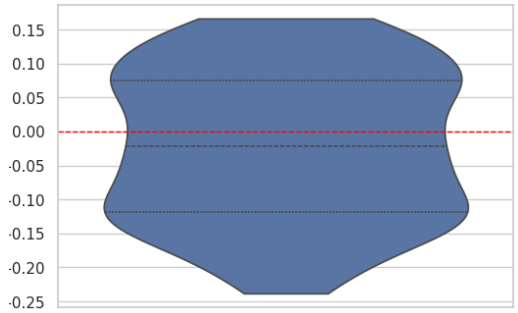
To assess whether the proposed T2 fuzzy epistemic truth modelling offers information not captured by a deterministic T1 entropy-based approximation, we performed a paired Wilcoxon signed-rank test on the same samples and conditions. The test compares the center of the T2 epistemic truth interval to a scalar T1 truth score computed from normalized entropy, testing the null hypothesis that the median of the differences is zero. The Wilcoxon test rejects the null hypothesis ($p<0.05$), indicating a statistically significant difference between the two methods.

Fig. 5(a) shows a box-and-strip plot of the distribution of paired truth differences (T2-T1). The red dashed line indicates zero difference. The distribution is clearly offset from zero, with the median and interquartile range indicating a systematic offset rather than outliers. This confirms that the observed statistical significance is due to systematic per-sample differences rather than a few extreme outliers.

Fig. 5(b) offers a second view of the same data, this time using a violin plot. The distribution shape indicates a wide



(a) Paired difference distribution for Wilcoxon test



(b) Distribution of paired differences (Wilcoxon test)

Fig. 5 Statistical significance observations

density that spans both positive and negative values, reflecting epistemic variability across samples. Crucially, the distribution is again offset from zero, confirming the Wilcoxon test results and showing that the T2 epistemic truth modelling induces a systematic and non-trivial departure from deterministic entropy-based truth scores.

Together, the test and accompanying visualizations confirm that the proposed T2 fuzzy framework captures epistemic information that cannot be reduced to scalar uncertainty proxies. The significance is due to systematic distributional differences at the sample level, supporting the claim that interval-valued epistemic truth offers a more rich and expressive representation of uncertainty than T1 or entropy-based alternatives.

V. Discussion

A. *Epistemic Conservatism and Hallucination Avoidance*

One of the most important observations in all the experiments is the epistemic conservatism of the proposed framework. In the absence of strong epistemic evidence, the framework never prefers high truth values, even for fluent and confident-looking outputs. This is especially true for the BART-Large model, in which stylistic fluency does not necessarily imply epistemic credibility. Epistemic conservatism is critical in reducing hallucinations, as the latter are often associated with surface plausibility rather than evidential support.

B. *Truth Is Not Confidence*

The experimental findings unequivocally show that truth, as formalized in the proposed framework, is not the same as probabilistic confidence or likelihood. High likelihood outputs can still have low or zero epistemic truth values when uncertainty signals indicate weak evidence. Similarly, low likelihood outputs can have high truth values when epistemic evidence is strong. This difference highlights the inadequacies of likelihood-based confidence models and the necessity of epistemic reasoning mechanisms that go beyond probability.

C. Why Type-2 Fuzzy Logic Is Necessary

The interval-valued dynamics in the experiments cannot be achieved by T1 fuzzy models or scalar uncertainty measures. The latter models reduce epistemic uncertainty to a scalar value, which loses uncertainty about uncertainty. In contrast, the proposed T2 fuzzy model models epistemic uncertainty using truth intervals whose widths vary according to uncertainty and evidence strength. This is necessary for trustworthy decision-making in language generation, in which incomplete knowledge and uncertainty are ubiquitous.

D. Interpretability and Transparency

One of the major strengths of the proposed framework is its interpretability. The fuzzy rule base, control surface, and rule activation heatmaps are all highly transparent and offer a clear understanding of how epistemic truth is inferred from internal model signals. In contrast to black-box calibration or neural uncertainty estimation, the proposed framework enables the practitioner to view, edit, and reason about epistemic behaviour directly. This is especially important in high-stakes applications where trust and accountability are paramount.

E. Limitations and Future Directions

Although the proposed framework offers a principled approach to epistemic control, it relies on heuristic epistemic signals inferred from model internals, which may not fully represent epistemic uncertainty. Future work may investigate the use of additional signals, such as retrieval consistency or cross-model disagreement, within the fuzzy inference process. Furthermore, although this work presents a post-hoc epistemic analysis, the integration of fuzzy epistemic control into decoding or reinforcement learning models is a promising area for future research.

F. Implications for Trustworthy Language Generation

In conclusion, the results and discussion above have clearly shown that fuzzy logic and, more specifically, T2 fuzzy systems offer a highly promising and under-explored paradigm for epistemic reasoning in contemporary language models. By representing truth as an interval-valued epistemic construct rather than a scalar confidence measure, the proposed framework offers a fundamentally new approach to hallucination-aware, interpretable, and trustworthy natural language generation.

VI. Conclusions

In this paper, a T2 fuzzy epistemic control framework for large language models was proposed, which formally modeled truth as an epistemic interval-valued variable instead of a scalar confidence value. By modelling language generation as an epistemic reasoning task, the proposed framework separated truth from likelihood and allowed uncertainty about uncertainty to be expressed in a principled and interpretable way. The internal model signals, such as token entropy and logit margin, were considered as noisy epistemic evidence and transformed via a clear fuzzy inference system without any need for retraining or modifying the underlying models.

Experimental results on the TruthfulQA benchmark, conducted using various transformer architectures, showed that the proposed framework is able to effectively distinguish epistemic reliability across models of different capacity. The results show that the larger models have higher and more stable epistemic truth with smaller uncertainty intervals, and models with fluent but poorly supported text are conservatively classified as having neutral truth. This is expected and verifies that the proposed framework does not suffer from false confidence in epistemic ambiguity, a problem that scalar uncertainty measures and entropy-based heuristics are known to be unable to prevent.

The importance of this work is in the demonstration of the need for T2 FL in epistemic reasoning in neural language generation. Interval-valued truth modelling allows for the representation of uncertainty about uncertainty, which is more expressive and stable than T1 fuzzy systems and probabilistic confidence measures. The epistemic control surfaces and rule graphics further emphasize the interpretability and transparency of the proposed framework.

In conclusion, this work provides a principled basis for fuzzy logic in hallucination-aware and trustworthy language modelling. The proposed framework, which combines modern language models with T2 fuzzy epistemic reasoning, opens up new avenues for the integration of interpretability, epistemic conservatism, and uncertainty-aware control in natural language processing systems, especially in scenarios where reliability and trust are paramount.

References

[1] T. Brown *et al*., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 1877–1901.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[3] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 3214–3252.

[4] J. Ji *et al*., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[5] M. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020, pp. 1906–1919.

[6] P. Lewis *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 9459–9474.

[7] S. Manakul, A. Ladhak, and K. Gimpel, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023, pp. 9004–9020.

[8] K. Kadavath *et al*., "Language models (mostly) know what they know," *arXiv preprint* arXiv:2207.05221, 2022.

[9] R. McKenna *et al*., "Sources of hallucinations in large language models," *arXiv preprint* arXiv:2305.14552, 2023.

[10] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[11] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[12] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[13] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[14] J. M. Mendel and R. I. John, "Type-2 fuzzy sets made simple," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 117–127, Apr. 2002.

[15] H. Chung *et al*., "Scaling instruction-finetuned language models," *arXiv preprint* arXiv:2210.11416, 2022