Transparency in Uncertainty: A Neutrosophic Evaluation of Ethical Reasoning in Language Models

Maikel Yelandi Leyva-Vázquez[1]*,
ORCID: 0000-0001-7911-5879

Florentin Smarandache[2],
ORCID: 0000-0002-5560-5926

[1] Universidad Bolivariana del Ecuador, Durán, Ecuador; maikel.leyvav@ug.edu.ec
[2] University of New Mexico, Gallup, NM, USA; smarand@unm.edu

- Correspondence: maikel.leyvav@ug.edu.ec

[Published on 2/17/2025, aiXiv, ID:aixiv.251217.000001
https://aixiv.science/abs/aixiv.251217.000001]

Abstract: Large Language Models (LLMs) are predominantly governed by probabilistic frameworks where the sum of outcome probabilities is constrained to unity. This architectural limitation, often imposed by Softmax layers, leads to a "collapse of uncertainty," making it difficult to differentiate between epistemic uncertainty (ignorance), paradox, and vagueness. This study presents an empirical investigation into the application of Neutrosophic Logic, a framework that treats Truth (T), Indeterminacy (I), and Falsity (F) as independent dimensions, to model epistemic states in LLMs. We conducted experiments on a family of OpenAI's GPT models, evaluating their responses to five distinct linguistic phenomena: logical paradoxes, epistemic ignorance, vagueness, ethical contradictions, and future contingencies. Our findings reveal that a neutrosophic approach, by allowing the sum of T, I, and F to exceed 1 (a state we term "hyper-truth"), provides a richer and more nuanced representation of a model's internal state. Specifically, in scenarios involving ethical dilemmas and logical paradoxes, the neutrosophic framework captures the inherent conflict and contradiction that probabilistic models obscure. We demonstrate that this approach not only preserves truth values in fuzzy contexts but also offers a robust method for identifying and quantifying internal model conflict. We conclude that the integration of neutrosophic evaluation layers is a critical step towards developing more transparent, reliable, and ethically-aware AI systems, particularly in high-stakes domains.

Keywords: neutrosophic logic; large language models; uncertainty quantification; epistemic uncertainty; hyper-truth; AI safety; ethical reasoning

1. Introduction

The remarkable capabilities of Large Language Models (LLMs) have led to their widespread adoption in diverse applications [1]. However, as their deployment in high-stakes domains increases, the need for robust uncertainty quantification (UQ) has become paramount [2,3]. The underlying architecture of most LLMs, which is deeply rooted in probability theory, imposes fundamental limitations on their ability to represent and reason about complex epistemic states [4]. The ubiquitous Softmax function, for instance, forces a zero-sum game where an increase in uncertainty necessitates a decrease in truth or falsity, a phenomenon we term the "collapse of uncertainty" [5,6]. This constraint hinders the ability of LLMs to distinguish between aleatoric uncertainty (statistical uncertainty inherent in the data) and epistemic uncertainty (model uncertainty due to lack of knowledge) [7,8]. This is a critical distinction, as epistemic uncertainty can, in principle, be reduced with more data, while aleatoric uncertainty cannot. The inability to differentiate between "not knowing" (ignorance) and "knowing of a conflict" (paradox or contradiction) is a direct consequence of this architectural limitation [9].

Neutrosophic Logic, a branch of philosophy and logic introduced by Florentin Smarandache, offers a compelling alternative [10]. In neutrosophic theory, truth (T), indeterminacy (I), and falsity (F) are independent dimensions, and they are not required to sum to one. When T + I + F exceeds 1, the system enters a non-normalized neutrosophic state, which reflects epistemic inconsistency or conflict between truth, falsity, and indeterminacy. In this work, we refer to such states as hyper-truth, a term used to denote the co-existence of multiple epistemic values beyond the probabilistic constraint, where truth and falsity are simultaneously present, and indeterminacy encapsulates the internal conflict inherent in certain complex phenomena. This paper explores the practical application of Neutrosophic Logic to enhance the reasoning capabilities of LLMs in complex and ambiguous scenarios.

2. Preliminaries: Neutrosophic Logic and Sets

In this section, we provide a formal definition of neutrosophic logic and sets, which form the theoretical foundation of our work.

2.1. Formal Definition of a Neutrosophic Set

Let U be a universe of discourse. A neutrosophic set A on U is characterized by three membership functions: a truth-membership function $T_A(x)$, an indeterminacy-membership function $I_A(x)$, and a falsity-membership function $F_A(x)$. For each point $x \in U$, we have $T_A(x), I_A(x), F_A(x) \in [0,1]$. These functions are independent, and their sum is not constrained:

$$0 \leq T_A(x) + I_A(x) + F_A(x) \leq 3 \qquad (1)$$

This is a fundamental departure from classical and fuzzy logics, where the sum of membership values is typically 1.

2.2. Neutrosophic Evaluation of a Proposition

Let Φ be a proposition. A neutrosophic evaluation of Φ, denoted as $v(\Phi)$, is an ordered triple (t, i, f), where $t, i, f \in [0,1]$ represent the degrees of truth, indeterminacy, and falsity of Φ, respectively.

2.3. Hyper-truth State

We define a **hyper-truth state** as a neutrosophic evaluation $v(\Phi) = (t, i, f)$ where the sum of the components exceeds 1:

$$t + i + f > 1 \tag{2}$$

This state signifies epistemic conflict or inconsistency, where evidence for truth, indeterminacy, and falsity coexists to a degree that is not representable in a probabilistic framework.

3. Materials and Methods

We designed a comparative study to evaluate the performance of a neutrosophic framework against traditional probabilistic and entropy-based approaches. The experiment involved a family of four OpenAI models: GPT-4o, GPT-4-turbo, GPT-3.5-turbo, and GPT-4o-mini.

2.1. Linguistic Phenomena

We selected five distinct linguistic phenomena to test the models' reasoning capabilities:

- **Logical Paradoxes:** Statements that lead to self-contradiction (e.g., "This sentence is false.").
- **Epistemic Ignorance:** Statements whose truth value is unknown (e.g., "The number of stars in the universe is even.").
- **Vagueness (Fuzzy Logic):** Statements with imprecise boundaries (e.g., "John is 1.75 meters tall, therefore John is tall.").
- **Ethical Contradictions:** Dilemmas where moral principles conflict (e.g., "Lying to save an innocent life is morally right and wrong at the same time.").
- **Future Contingencies:** Statements about future events that are not yet determined (e.g., "It will rain in New York tomorrow.").

2.2. Evaluation Strategies

We employed a specialized prompt engineering framework to query the models using three distinct strategies:

1 **Strategy 1 (Neutrosophic):** The model was instructed to act as a Neutrosophic Logic expert and evaluate the statement in three independent dimensions (T, I, F) on a scale from 0.0 to 1.0.

2 **Strategy 2 (Probabilistic):** The model was instructed to act as a standard probabilistic classifier, assigning probabilities to three mutually exclusive states (True, Uncertain, False), with the sum constrained to 1.0.

3 **Strategy 3 (Entropy-Derived):** The model was asked to estimate the probability of the statement being YES (True) vs. NO (False), from which we derived an indeterminacy value.

2.3. Data and Code Availability

The code, data, and notebooks used in this study are publicly available on GitHub at: https://github.com/mleyvaz/neutrosophic-llm-logic/tree/main

3. Results

The experimental data reveals significant differences in how LLMs represent uncertainty under neutrosophic and probabilistic frameworks. The analysis of the collected data is presented in the following sections.

3.1. Descriptive Statistics

Table 1 presents a summary of the descriptive statistics for the neutrosophic components (Strategy 1) across the different linguistic phenomena. The data reveals distinct patterns in how LLMs represent different types of linguistic phenomena using the neutrosophic framework.

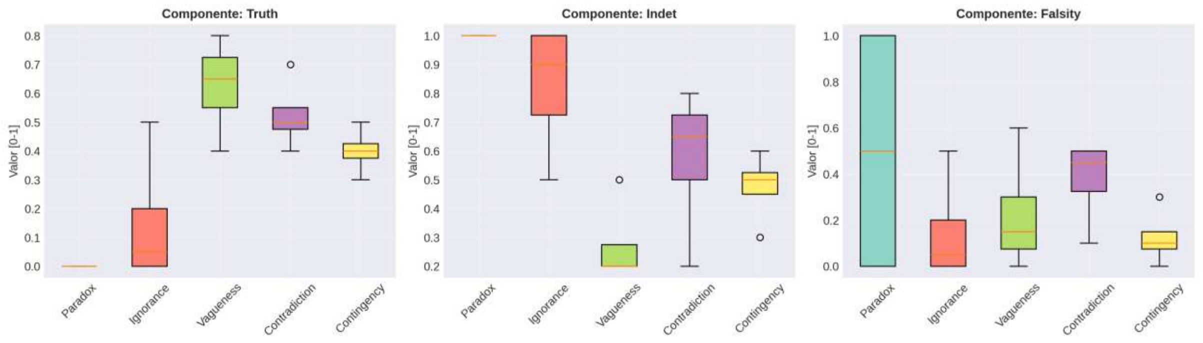
Table 1. Descriptive Statistics for Neutrosophic Components (Strategy 1) by Phenomenon.

Phenomenon_Type	S1_Truth_T (mean)	S1_Indet_I (mean)	S1_Falsity_F (mean)	S1_Sum_TIF (mean)
Contingency (Future)	0.400	0.475	0.125	1.000
Contradiction (Ethical)	0.525	0.575	0.375	1.475
Ignorance (Epistemic)	0.150	0.825	0.150	1.125
Paradox (Logical)	0.000	1.000	0.500	1.500
Vagueness (Fuzzy)	0.625	0.275	0.225	1.125

3.2. Distribution of Neutrosophic Components

The distribution of the neutrosophic components (T, I, F) for each linguistic phenomenon is visualized in Figure 1. This figure illustrates how the models assign different levels of truth, indeterminacy, and falsity depending on the nature of the statement.

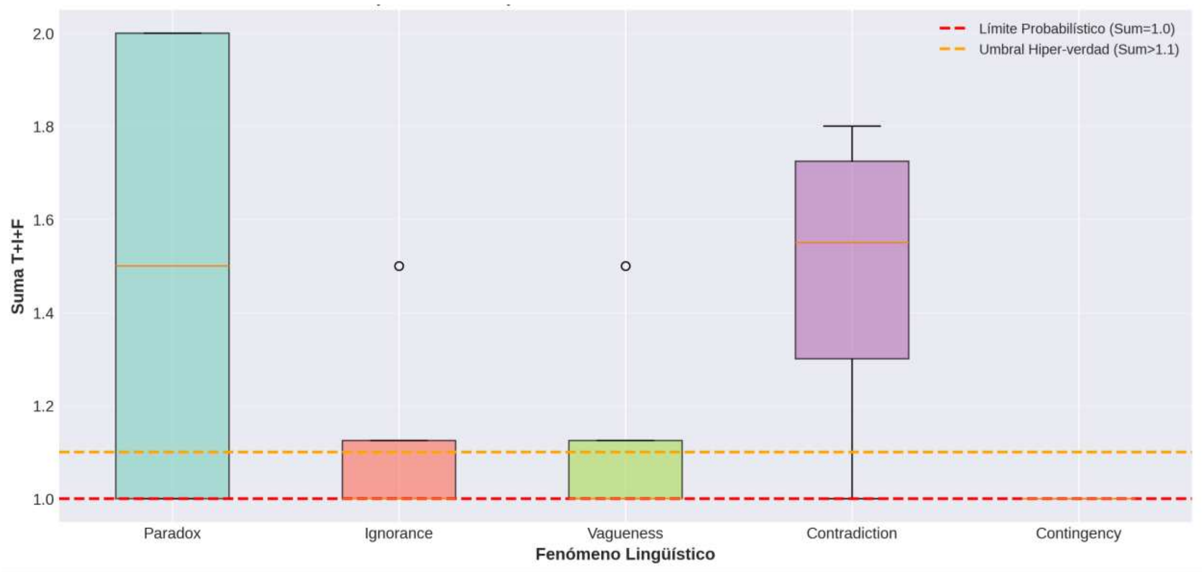
Figure 1. Distribution of the neutrosophic components (Truth, Indeterminacy, Falsity) for each linguistic phenomenon under Strategy 1.



3.3. Hyper-truth: Breaking the Probabilistic Constraint

A key finding of this study is the emergence of "hyper-truth," where the sum of the neutrosophic components (T+I+F) exceeds the probabilistic limit of 1.0. This phenomenon is particularly prominent in cases of ethical contradiction and logical paradox, as shown in Figure 2.

Figure 2. Boxplot of the sum of neutrosophic components (T+I+F) for each linguistic phenomenon, demonstrating the violation of the probabilistic constraint (Sum=1) in cases of contradiction and paradox.



4. Discussion

The results of our study provide compelling evidence that the probabilistic constraint inherent in current LLM architectures is insufficient for modeling the complexity of human reasoning. The emergence of "hyper-truth" in the neutrosophic framework allows LLMs to communicate internal conflicts and contradictions without collapsing into a state of false certainty. This is particularly evident in the ethical dilemma, where the neutrosophic approach correctly identifies the moral ambiguity, while the probabilistic model misrepresents it as low probability. This finding aligns with the growing body of literature that calls for moving beyond Softmax-based uncertainty measures, which are often poorly calibrated and can lead to overconfident predictions, especially for out-of-distribution inputs [5,6].

The neutrosophic framework provides a direct way to model epistemic uncertainty. The Indeterminacy component (I) can be interpreted as a measure of the model's own uncertainty, which is distinct from the aleatoric uncertainty that might be present in the data itself. The ability to obtain a high value for I, without necessarily suppressing T and F, allows the model to express a state of "known unknown," which is a crucial capability for safe and reliable AI [7].

Furthermore, the neutrosophic framework demonstrates its ability to preserve truth values in fuzzy contexts, where the probabilistic approach tends to penalize partial truths. This suggests that Neutrosophic Logic is a more suitable framework for handling vagueness and imprecision in natural language. This is a significant advantage over traditional methods, which often struggle to represent the gradual nature of truth in fuzzy propositions [10].

The ability to distinguish between ignorance (high indeterminacy) and contradiction (high truth and falsity) is a critical feature of the neutrosophic approach that is lost in traditional probabilistic and entropy-based methods. This distinction is crucial for building more robust and trustworthy AI systems. The capacity to represent a state of conflict (high T and F) is a novel feature that is not explicitly captured by the traditional dichotomy of aleatoric and epistemic uncertainty [8,9]. This "contradictory" state, which we have termed hyper-truth, could be a valuable signal for detecting adversarial attacks, identifying ethical dilemmas, or flagging instances where the model is forced to reconcile conflicting information.

5. Conclusions

This study has demonstrated the practical benefits of applying Neutrosophic Logic to the evaluation of epistemic uncertainty in large language models. Our findings indicate that the neutrosophic framework provides a more expressive and nuanced representation of a model's internal state, particularly in scenarios involving conflict, contradiction, and vagueness. We recommend implementing neutrosophic evaluation layers in critical applications where the ability to distinguish between different types of uncertainty is crucial.

Future work should focus on fine-tuning LLMs to natively output neutrosophic vectors, which could lead to significant improvements in their reasoning and decision-making capabilities. Further research is also needed to explore the application of Neutrosophic Logic in other areas of AI, such as computer vision and robotics.

Author Contributions: Conceptualization, M.Y.L.-V. and F.S.; methodology, M.Y.L.-V.; software, M.Y.L.-V.; validation, M.Y.L.-V. and F.S.; formal analysis, M.Y.L.-V.; investigation, M.Y.L.-V.; resources, M.Y.L.-V.; data curation, M.Y.L.-V.; writing—original draft preparation, M.Y.L.-V.; writing—review and editing, F.S.; visualization, M.Y.L.-V.; supervision, F.S.; project administration, M.Y.L.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in the GitHub repository at https://github.com/mleyvaz/neutrosophic-llm-logic/tree/main.

Acknowledgments: The authors would like to thank the reviewers for their valuable feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

4 Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

5 Shelmanov, A., et al. (2025). Uncertainty Quantification for Large Language Models. *ACL 2025, Tutorial Abstracts*.

6 Shorinwa, O., et al. (2024). A Survey on Uncertainty Quantification of Large Language Models. *arXiv preprint arXiv:2412.05563*.

7 Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of the 33rd International Conference on Machine Learning, 48*, 1050-1059.

8 Guo, C., et al. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning, 70*, 1321-1330.

9 Veličković, P. (2022). Softmax is not Enough (for Sharp Size Generalisation). *ICLR 2022*.

10 Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning, 110*(3), 457-506.

11 Valdenegro-Toro, M. (2022). A Deeper Look into Aleatoric and Epistemic Uncertainty Estimation. *arXiv preprint arXiv:2204.09308*.

12 De Finetti, B. (1974). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons.

13 Smarandache, F. (1998). *A Unifying Field in Logics: Neutrosophy*. American Research Press.

14 OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

15 Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730-27744.

16 Touvron, H., et al. (2023). Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

17 Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.

18 Vaswani, A., et al. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems, 30*, 5998-6008.

19 Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338-353.

20 Leyva-Vazquez, M., & Smarandache, F. (2018). Neutrosophic logic for decision-making. *Neutrosophic Sets and Systems, 20*, 3-14.

21 Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. *christophm.github.io/interpretable-ml-book*.

22 Hendrycks, D., et al. (2021). Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

23 Ye, J., et al. (2023). Compositional Finetuning for Large Language Models. *arXiv preprint arXiv:2310.02116*.

24 Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. *arXiv preprint arXiv:1906.04284*.

25 Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6837-6850.

26 Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

27 Amodei, D., et al. (2016). Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

28 Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.