An Explainable Fuzzy Based Approach for Embeddings in Generative AI Transformer Models

Jera Makar, Hani Hagras and Javier Andreu-Perez
The Artificial Intelligence Group
School of Computer Science and Electronic Engineering
University of Essex, Colchester, UK
Email: jg23575@essex.ac.uk

Scott Payton and Thomas Golden
Bowen Craggs & Co Ltd
London, UK
Email: tgolden@bowencraggs.com

***Abstract*— Transformer models achieve remarkable performance across language, vision, and multimodal tasks, yet their internal representations remain largely opaque. In particular, the semantic structure of embedding spaces, central to how transformers operate, remains difficult to interpret. This paper introduces a framework that converts embedding vectors into interpretable, linguistically grounded fuzzy semantic representations. The approach integrates external lexical knowledge with data-driven analysis, using WordNet-derived semantic categories as reference signals. Dimensionality reduction first compresses the embedding space while preserving major variance patterns, followed by a fuzzy modelling layer that combines Fuzzy C-Means clustering with multiscale Gaussian membership functions. This produces a transparent semantic signature for each word, capturing graded associations with semantic concepts across multiple levels of abstraction. We evaluate the approach on six human similarity benchmarks (WordSim353, SimLex999, MEN, RareWords, MTurk, and SimVerb3500) across five transformer models (GPT-2, BERT, Gemini, Qwen3, and MiniLM). The proposed representations consistently achieve higher agreement with human judgments, with improvements of up to 448% in Spearman correlation, supported by paired bootstrap significance testing. These findings show that improving interpretability can also enhance alignment with human semantic intuition, supporting more trustworthy AI systems.**

I. INTRODUCTION

Transformers have become central to modern generative AI, powering applications from search engines and chatbots to high-stakes domains such as medicine, law, and finance. As their societal role grows, so does the need to understand their decision-making. Despite their success, transformers remain difficult to interpret, their complex internal reasoning often renders them "black boxes," undermining trust, accountability, and responsible use. This opacity arises from the architecture itself: multiple layers of self-attention heads, residual connections, and feedforward networks, where vast numbers of parameters interact through high-dimensional vector operations that are difficult to trace. This challenge is particularly evident in embeddings, which represent tokens as large numerical vectors whose dimensions lack direct human interpretability; it is unclear what each dimension encodes or how changes affect model behaviour.

Researchers have proposed various explainability methods, many of which are post-hoc. Techniques such as attention visualization, SHAP [1] , Integrated Gradients [2] , and reasoning procedures such as Chain-of-Thought [3] or Chain-of-Logic [4] prompting provide useful insights into outputs, but rarely explain how meaning is encoded in the underlying vector space. Attention weights are not always reliable explanations [5] while gradient-based methods, though more structured [6] are computationally expensive and often qualitative. Other studies examine how training and prompting influence explainability patterns [7] .

Recent work on interpretable embeddings explores recovering semantic meaning from embedding spaces. Approaches such as Conceptualization of Embeddings [8] ,VICE [9] , and large-scale analyses by Wang and Du [10] show that embeddings form meaningful geometric structures, with some dimensions correlating to identifiable concepts. However, these methods often focus on limited vocabularies and lack explicit, human-readable interpretations across all dimensions. Retrofitting techniques [11] and sense-level embeddings like SensEmbed [12] improve semantic specificity but remain localized rather than offering global explainability. Since embeddings lie at the core of transformer computation, analysing them directly offers a promising path toward deeper interpretability.

This paper introduces a framework to enhance transformer embedding interpretability by integrating semantic knowledge bases. The methodology addresses the ambiguity of natural language through two key contributions:

1. A novel approach linking transformer embeddings to WordNet [13] concepts, using fuzzy logic to handle graded membership and polysemy.
2. A structured mapping that converts embedding dimensions into interpretable linguistic objects, aiding the diagnosis of bias and internal associations

Section II reviews transformers, embeddings, and WordNet. Section III presents the proposed fuzzy-based approach. Section IV details experiments and results, and Section V concludes with future work.

II. HIGH LEVEL OVERVIW ON TRANSFORMERS, WORDNET AND MULTISCALE FUZZY

A. Overview of Transformers and Embeddings

Transformers (depicted in Fig.1), introduced by Vaswani et al. [14] , replaced recurrent architectures by

processing tokens in parallel through self-attention, allowing each token to weigh the relevance of others and build contextual representations. Modern models stack multiple layers of attention heads and feed-forward networks, achieving state-of-the-art performance across NLP tasks.

This expressiveness, however, comes at the cost of interpretability. Dense transformations and non-linearities obscure information flow and increasing depth further weakens the link between internal components and predictions, reinforcing their black-box nature.

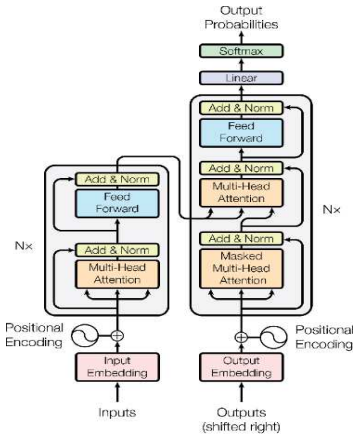

Fig. 1. High-level overview of a transformer architecture showing how token embeddings are processed through attention layers. [14]

This opacity is compounded by embeddings. Transformers map tokens to dense vectors encoding semantic relationships geometrically. As illustrated in Fig. 2, words are points in a vector space where relationships correspond to directions and distances (e.g., King – Man + Woman ≈ Queen). While this enables reasoning and generalization, individual dimensions remain uninterpretable, making embeddings a core source of opacity. Achieving interpretable embeddings is therefore critical for transparent transformer reasoning.

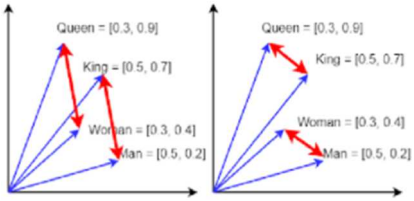

Fig. 2. Two-dimensional illustration of word embeddings, where semantic relationships such as gender and royalty are represented as vector directions[15] [19]

B. Overview of WordNet

WordNet [13] is a structured semantic lexicon that organizes words into synsets, sets of cognitive synonyms representing distinct senses, linked by semantic relations such as hypernym–hyponym hierarchies. It provides curated, human-interpretable semantic categories and explicitly represents polysemy by assigning different senses of a word to separate synsets.

This sense-aware, hierarchical structure aligns naturally with the proposed fuzzy membership framework, enabling embedding dimensions to be grounded in linguistic concepts while explicitly modelling semantic ambiguity.

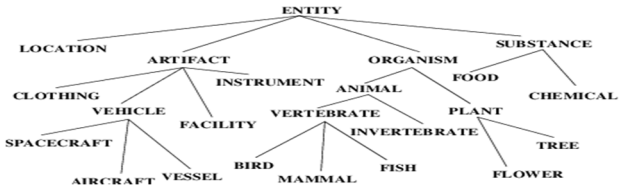

Fig. 3. A sample WordNet hierarchy [16]

C. Multiscale Fuzzy Membership

Classical fuzzy membership assigns each word a degree of belonging using a single kernel width, which is limiting since language operates at multiple levels of granularity. For example, *joy* belongs broadly to positive emotion but also specifically to high-arousal positive affect, which a single-scale kernel cannot capture.

To address this, we adopt a Multiscale Fuzzy Membership formulation [17] Instead of fixing one Gaussian bandwidth, each semantic dimension uses a set of scales $\{\sigma_1, \sigma_2, \ldots, \sigma_S\}$, where $\sigma_1 < \sigma_2 < \ldots < \sigma_S$. Narrow bandwidth ($\sigma_1$) captures fine-grained distinctions between joy and relief for example, while wider bandwidth (σ_S) captures high-level conceptual categories (grouping all positive emotions together).

For a word vector z_i and a cluster centroid v_j, the membership at scale s is computed as [18] :

$$u_{ij}^{(s)} = \exp\left(-\frac{|z_i - v_j|^2}{2\sigma_s^2}\right) \quad (1)$$

The final membership degree u_{ij} is a convex combination of these multiscale memberships [19] with weights w_s :

$$u_{ij} = \sum_{s=1}^{S} w_s \, u_{ij}^{(s)}, \qquad \sum_{s=1}^{S} w_s = 1, \; w_s \quad (2)$$

This formulation is well-suited to semantic embeddings. Multiple scales provide adaptive semantic resolution, capturing both fine-grained nuances and broad associations. It supports polysemy by allowing words to have blended memberships across concepts rather than hard assignments and improves robustness by stabilizing representations against noise and outliers.

III. The Proposed Explainable Fuzzy Based Approach for Embeddings in Generative AI Transformer Models

The proposed method (Fig.4) is a four-stage transformation pipeline, detailed below, that converts raw transformer embeddings into a fuzzy-interpretable, semantically structured representation.

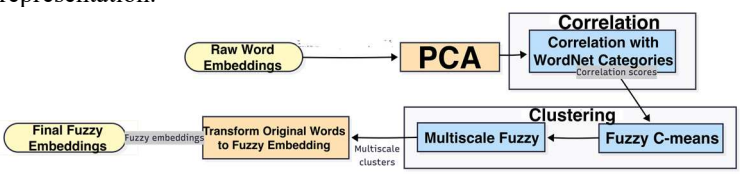

Fig. 4. Flow diagram of the proposed framework.

A. Embedding Generation and Dimensionality Reduction

The first stage extracts embeddings from transformer models such as BERT [20] or GPT-2 [21] . Given a vocabulary $\{w_1, w_2, \ldots, w_N\}$, each word is mapped to a contextual embedding:

$$e_i \in \mathbb{R}^D \quad (3)$$

Where D (768 or 1024) is the native embedding dimension. To reduce redundancy and entanglement, Principal Component Analysis (PCA) is applied to project embeddings into a lower-dimensional space [22] :

$$z_i = P^\top (e_i - \mu) \quad (4)$$

Where $P \in \mathbb{R}^{D\times d}$ contains the top d components preserving 98% variance, and μ is the mean embedding. PCA maintains linear interpretability, preserving correlations with semantic structure. The reduced dimension is set to 89, matching the number of WordNet noun synset categories at the chosen abstraction level, ensuring alignment with interpretable lexical concepts.

B. Correlation Analysis for Semantic Dimension Discovery

The second stage quantifies how each reduced embedding dimension aligns with WordNet categories. Each word w_i is associated with one or more synsets and assigned to a semantic category C_k . The relationship between dimension j and category k is measured using Spearman rank correlation [23] :

$$\rho_{j,k} = \text{Spearman}\left(Z_{:j}, C_k\right) \quad (5)$$

Where $Z_{:j}$ is the vector of PCA values for dimension j, andC_k is a binary membership vector. Spearman correlation is used for its robustness to nonlinearity and outliers.

To ensure statistical validity, multiple comparison corrections are applied using Bonferroni [24] and Benjamini-Hochberg [25] :

$$p_{\text{adj}}^{\text{Bonf}} = min(1, p \cdot M), \quad \text{(Bonferroni)} \quad (6)$$

$$p_{\text{adj}}^{\text{BH}} = \frac{p \cdot M}{\text{rank}(p)}, \quad \text{(Benjamini--Hochberg)} \quad (7)$$

where M is the number of tests. Bootstrapped confidence intervals are also computed [26] :

$$C_I = \left[\rho_{j,k}^{(\alpha/2)}, \rho_{j,k}^{(1-\alpha/2)}\right] \quad (8)$$

Where, α is the significance level. This produces a correlation matrix $\{\rho_{j,k}\}$ that maps embedding dimensions to human-interpretable semantic categories, revealing both concept alignment and hierarchical structure.

C. Fuzzy Clustering of Embeddings

After assigning semantic meaning to each dimension, Fuzzy C-Means (FCM) is used to model how words organize within this space. Unlike hard clustering, FCM assigns each word degrees of membership across multiple clusters, enabling representation of polysemy and overlapping meanings.

Let $z_i \in \mathbb{R}^D$ denote the reduced embedding of word i . FCM optimizes the objective [27] :

$$J = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij}^m |z_i - v_j|^2 \quad (9)$$

where $u_{ij} \in [0,1]$ is the membership degree of word i in cluster j , v_j is the centroid of cluster j , $m > 1$ controls fuzziness and *N* and *K* are the number of words and clusters, respectively. Memberships and centroids are updated iteratively [26] [27] :

$$u_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{|z_i - v_j|}{|z_i - v_k|}\right)^{2/(m-1)}} \quad (10)$$

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^m z_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (11)$$

Each centroid represents a core semantic concept, and each word is associated with a fuzzy membership vector $u_i = [u_{i1}, \ldots, u_{iK}]$, explicitly capturing its relation to multiple concepts. Importantly, clustering is performed independently within each PCA dimension, allowing words to exhibit different semantic memberships across dimensions. For example, *bank* may align with a physical entity in one dimension and a financial concept in another, reflecting distributed polysemy across semantic axes.

D. Constructing Interpretable Multiscale Feature

While FCM provides fuzzy memberships, these exist in cluster space and require further refinement for full interpretability. We therefore construct a multiscale feature representation that explicitly quantifies each word's relationship to semantic concepts.

Gaussian Semantic Membership: For each semantic dimension d with clusters $C = 1, \ldots . C_d$, a word's coordinate $z_{i,d}$ is converted into a smooth membership using a Gaussian function [18] :

$$k_{i,d,c} = \exp\left[-\frac{1}{2}\left(\frac{z_{i,d} - \mu_{d,c}}{\sigma_{d,c}}\right)^2\right] \quad (12)$$

where $\mu_{d,c}$ is the cluster centre and $\sigma_{d,c}$ controls the concept's spread. Values range from 0 (far from the concept) to 1 (exactly at the cluster centre).

Multiscale Representation: Word meaning operates at multiple levels of specificity (as in WordNet). For example, "*dog*" belongs both to the broad category "*animal*," and to more specific categories like "*mammal*" and "*pet*". To capture this, Gaussian memberships are computed across multiple scales using scale factors s [18]

$$k_{i,d,c}^{(s)} = \exp\left[-\frac{1}{2}\left(\frac{z_{i,d} - \mu_{d,c}}{\sigma_{d,c} \cdot s}\right)^2\right] \quad (13)$$

Small s captures fine distinctions, while large s captures broader associations. For example, *bank* may align with a *physical entity*, at coarse scales but a *financial gain* concept at finer scales. Evaluating memberships across dimensions and scales captures both dominant meaning and ambiguity in a transparent way.

Integrating Scales and Normalization: Rather than choosing one scale, we combine multiscale memberships via a weighted sum

$$\widetilde{m_{i,d,c}} = \sum_{s=1}^{S} w_s \; k_{i,d,c}^{(s)} \quad (14)$$

To ensure comparability, scores are normalized within each dimension:

$$m_{i,d,c} = \frac{\widetilde{m_{i,d,c}}}{\sum_{c'=1}^{C_d} \widetilde{m_{i,d,c'}}} \quad (15)$$

This ensures memberships sum to 1 per dimension. Polysemy is reflected through membership patterns across dimensions rather than a single axis.

E. Final Interpretable Fuzzy Embedding

The final representation of word i concatenates normalized memberships across all dimensions and clusters:

$$f_i = [m_{i,1,1}, \dots, m_{i,1,C_1}, m_{i,2,1}, \dots, m_{i,2,C_2}, \dots, m_{i,D,C_D}] \quad (16)$$

Each value indicates the strength of association between a word and a specific semantic concept within a given dimension. This transforms opaque embedding axes into interpretable ones, where scores directly reflect semantic alignment. Multiscale structure captures both fine-grained and coarse semantics, while polysemy appears as distributed memberships across concepts. Normalized $[0, 1]$ values provide human-readable interpretations.

IV. Experiments and Results

To evaluate alignment with human semantic judgments, we measure agreement between human similarity scores and model-derived similarities using Spearman rank correlation, a scale-invariant, distribution-free metric. Embeddings are collected from GPT2 [20] BERT [21] Gemini [28] Qwen3 [29] and MiniLM-L3-v2 [30] and evaluated against standard human similarity datasets: WordSim353 [31] SimLex999 [32] MEN [33] RareWords [34] MTurk [35] and SimVerb-3500 [36] For each dataset D with word pairs (w_p, w_q) and human scores h_{pq}, values are normalized to $[0,1]$:

$$\widehat{h_{pq}} = \frac{h_{pq} - a_D}{b_D - a_D} \quad (17)$$

Where $[a_D, b_D]$ is the dataset's original rating scale. For each word pair, we compute cosine similarity scores using both original transformer embeddings x_w and final fuzzy embeddings f_w [37] :

$$\text{sim}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} \quad (18)$$

This yields similarity vectors for original embedding:

$$s_D^{(o)} = \left\{ s_{pq}^{(o)} = \text{sim}\left(x_{w_p}, x_{w_q}\right) : (w_p, w_q) \in P_D \right\} \quad (19)$$

and fuzzy embedding yield:

$$s_D^{(f)} = \left\{ s_{pq}^{(f)} = \text{sim}\left(f_{w_p}, f_{w_q}\right) : (w_p, w_q) \in P_D \right\} \quad (20)$$

The Spearman correlation between human judgments and embedding type $t \in \{o, f\}$ is [23] :

$$\rho_D^{(t)} = 1 - \frac{6 \sum_{(p,q) \in P_D} \left(\text{rank}\left(s_{pq}^{(t)}\right) - \text{rank}\left(\widehat{h_{pq}}\right) \right)^2}{M_D\left(M_D^2 - 1\right)} \quad (21)$$

Where M_D is the number of word pairs in dataset D.

A. Comparison Metric: For each dataset D, we define the improvement as:

$$\Delta\rho_D = \rho_D^{(f)} - \rho_D^{(o)} \quad (22)$$

A positive $\Delta\rho_D$ indicates that the fuzzy embeddings align more closely with human ratings.

B. Paired Bootstrap Significance Testing: Statistical significance is assessed using a paired non-parametric bootstrap with $B = 10{,}000$ resamples ($or\ 5{,}000\ when M_D < 100$). For each sample b, correlations $\rho_D^{(o,b)}$ and $\rho_D^{(f,b)}$ are recomputed, and differences $\Delta\rho_D^{(b)} = \rho_D^{(f,b)} - \rho_D^{(o,b)}$ recorded. A 95% confidence interval is derived from the 2.5th and 97.5th percentiles. The two-sided p-value is the proportion of samples where $|\Delta\rho_D^{(b)}| \geq |\Delta\rho_D|$.

C. *Interpretation of Results:* If the 95% confidence interval for $\Delta\rho_D$ lies entirely above zero, fuzzy embeddings significantly outperform original embeddings; if it includes zero, the improvement is not statistically conclusive. Fig. 5 shows average Spearman correlations across all five models and six datasets, while Fig. 6 presents a scatter plot where points above the diagonal y = x indicate improvement.

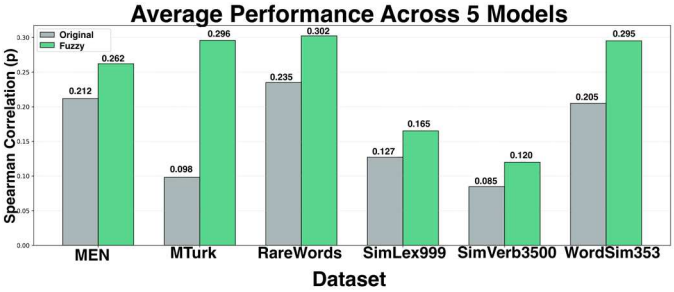

Fig. 5. Average performance of 5 models, correlation comparison between original (the grey bars) and fuzzy embeddings (the green bars) across all datasets.

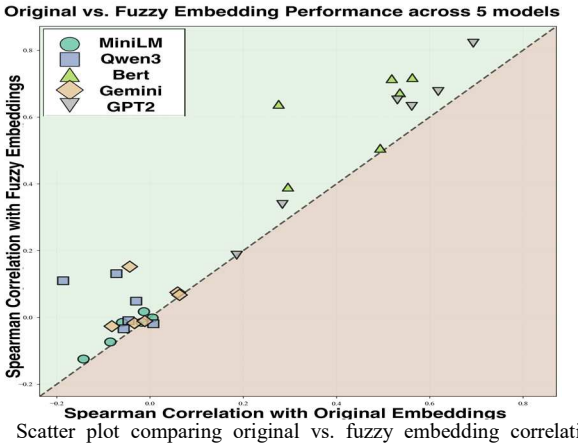

Fig. 6. Scatter plot comparing original vs. fuzzy embedding correlations. Points above the diagonal $y = x$ indicate improvement.

D.Key Observations: Across all datasets and models, fuzzy embeddings consistently improve alignment with human similarity judgments (Table I- II)

- **BERT** shows the most consistent improvements across all datasets, with absolute gains (+0.095 to +0.362) and strong percentage improvements (+25.6% to +131.3%).
- **GPT-2** demonstrates moderate but consistent improvements**,** particularly on RareWords (+18.5%) and MTurk (+22.8%).
- **Gemini, Qwen3** show dramatic percentage improvements on MTurk (+448.5% and +159.0% respectively)
- **MiniLM** exhibits the most variability, with notable improvement on SimLex999 (+230.5%).
- **MTurk** shows the largest improvements across most models, indicating the fuzzy transformation is particularly effective for this type of semantic similarity judgment.

• **SimVerb3500** shows modest improvements for most models, with BERT achieving the largest gain (+32.0%).

TABLE I. PERCENTAGE IMPROVEMENT $\Delta\rho\%$ OF SPEARMAN CORRELATION ACROSS MODELS AND DATASETS

Dataset	*GPT2*	*BERT*	*Gemini*	*Qwen3*	*Minilm*
WordSim353	9.5%	27.8%	26.3%	281.6%	12.7%
SimLex-999	19.2%	2.6%	67.9%	79.2%	230.5%
MEN	12.7%	37.9%	50.0%	0.0%	0.0%
RareWords	18.5%	25.6%	6.6%	39.2%	75.4%
MTurk	22.8%	131.3%	+448.5%	159.0%	14.2%
SimVerb3500	0.2%	32.0%	3.5%	263.6%	17.2%

TABLE II. ABSOLUTE IMPROVEMENT $\Delta\rho$ OF SPEARMAN CORRELATION ACROSS MODELS AND DATASETS

Dataset	*GPT-2*	*BERT*	*Gemini*	*Qwen3*	*Minilm*
WordSim353	0.058	0.156	0.016	0.203	0.018
SimLex999	0.054	0.013	0.056	0.037	0.030
MEN	0.071	0.197	0.017	0.000	0.000
RareWords	0.128	0.137	0.004	0.022	0.045
MTurk	0.121	0.362	0.196	0.296	0.012
SimVerb3500	0.000	0.095	0.000	0.078	0.003

E. Model Parameter Selection: Beyond performance gains, the approach improves explainability by transforming opaque embeddings into explicit semantic memberships, where each word is associated with interpretable concepts rather than abstract dimensions. Parameters were selected through systematic validation. The number of clusters K was chosen via silhouette analysis to balance granularity and separation. The FCM fuzziness parameter m was evaluated over $\{1.5, 2.0, 2.5\}$, with $m = 2.0$ selected for optimal performance and interpretability. Gaussian spreads were initialized from cluster statistics and scaled using kernel multipliers $\{1.0, 2.0, 3.0\}$, with $w = 3.0$ selected to ensure sufficient coverage without excessive overlap. For multiscale modelling, scales$\{1.0, 2.0, 3.0\}$with weights $\{0.25, 0.50, 0.25\}$ were selected, emphasizing medium-scale semantic structure.

F. Mechanisms of Explainability and Polysemy: The explainability mechanism is illustrated in Fig. 7 using *bank* as a case study. Interpretation arises not from a single dimension, but from fuzzy memberships across multiple semantic dimensions. In one dimension, clusters correspond to concepts such as *geological formation*, *physical entity*, and *financial institution*, aligned with WordNet synsets. Along a different dimension, the membership shifts toward *abstract institutional* meanings while *physical associations* weaken. This demonstrates that semantic interpretation is dimension-dependent, with polysemy represented through coexisting fuzzy memberships rather than a single latent meaning.

Each concept is modelled as a Gaussian fuzzy set. When *"bank"* appears at position $z = 0.3$ along this dimension, its memberships are computed as 0.011, 0.835, and 0.138, respectively. After normalization, these values indicate primary association with *physical entity* and secondary association with *financial institution*, directly quantifying semantic ambiguity. More broadly, each word forms a structured semantic profile, where dimensions contribute partial evidence toward different interpretations. WordNet reinforces this by organizing senses hierarchically, allowing words to inherit meaning from multiple parent concepts. Multiscale fuzzy modelling captures this hierarchy: narrow kernels emphasize fine-grained senses, while broader kernels activate higher-level abstractions. This representation is interpretable for three reasons. First, **semantic grounding**: each fuzzy set aligns with WordNet-based concept clusters rather than arbitrary axes. Second, **quantified meaning**: normalized values $[0, 1]$ explicitly measure concept association. Third, **explicit polysemy**: multiple non-zero memberships directly encode ambiguity. All computations remain traceable through Gaussian functions. Cluster quality metrics confirm meaningful structure: an average silhouette coefficient of 0.68 indicates well-separated clusters, intra-cluster cosine similarity averages 0.72 (semantic coherence), and inter-cluster distances average 1.84 in PCA space (clear separation). These results validate that fuzzy partitions capture genuine semantics rather than arbitrary segmentation. Unlike traditional embeddings, fuzzy embeddings represent each word through explicit semantic memberships, making meaning, ambiguity, and similarity directly interpretable. By combining fuzzy logic with WordNet's hierarchy, the framework enables transparent semantic reasoning.

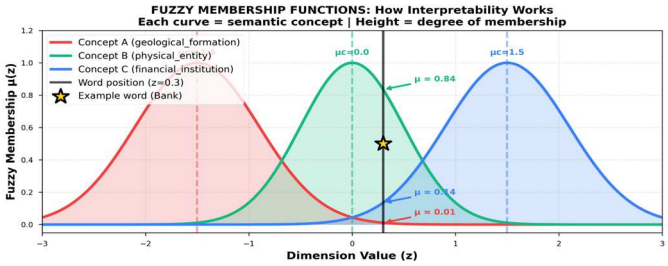

Fig. 7. Explainability demonstration for "bank" Word position (z=0.3) yields memberships of 1% geological_formation, 85% physical_entity, and 14% financial_institution.

V. CONCLUSIONS AND FUTURE WORK

This work presents a framework for transforming opaque transformer embeddings into interpretable, linguistically grounded fuzzy semantic representations. By integrating WordNet structure, PCA-based organization, Fuzzy C-Means clustering, and multiscale Gaussian membership functions, embeddings are mapped to explicit semantic concepts rather than latent coordinates. Evaluation across six benchmarks and multiple transformer models shows that fuzzy representations improve alignment with human semantic judgments, demonstrating that interpretability can be achieved without sacrificing performance. The resulting multiscale membership vectors provide a transparent view of how meaning is encoded and where models diverge from human understanding. Future work will extend the framework using Type-2 fuzzy logic systems to better capture linguistic and numerical uncertainty, building on adaptive and non-singleton fuzzy systems [38] , [39] and decision-making approaches [40] . Optimization of membership functions could be improved using multi-objective methods such as NSGA-III. Additional directions include extending the approach to contextualized

tokens, sentences, and cross-lingual settings, as well as applications in explainable classification, semantic search, and interpretable text generation. Finally, dynamic modelling of polysemy, drawing on adaptive system paradigms [41] , remains key to capturing the fluid nature of meaning.

VI. REFERENCES.

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:21889700

[2] Y. Wang, T. Zhang, X. Guo, and Z. Shen, "Gradient based feature attribution in explainable ai: A technical review," *ArXiv*, vol. abs/2403.10415, 2024.[Online]: https://api.semanticscholar.org/CorpusID:268510511

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[4] S. Servantez, J. Barrow, K. Hammond, and R. Jain, "Chain of logic: Rule-based reasoning with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.10400

[5] S. Jain and B. C. Wallace, "Attention is not explanation," *CoRR,* vol. abs/1902.10186, 2019. [Online]. Available: http://arxiv.org/abs/1902.10186

[6] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," 2021. [Online]. Available: https://arxiv.org/abs/2012.09838

[7] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2309.01029

[8] A. Simhi, I. Shvadron, D. Greene, R. Wattenhofer, and K. Radinsky, "Conceptualization of embeddings," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[9] L. Muttenthaler, C. Y. Zheng, P. McClure, R. A. Vandermeulen, M. N. Hebart, and F. Pereira, "Vice: Variational interpretable concept embeddings," 2022. [Online]. Available: https://arxiv.org/abs/2205.00756

[10] Z. Wang and B. Du, "Interpretable embedding dimensions at scale: Modeling key embedding dimensions with concepts," *in Advances in Neural Information Processing Systems,* 2024.

[11] M. Faruqui, J. Dodge, S. Jauhar, C. Dyer, E. Hovy, and N. Smith, "Retrofitting word vectors to semantic lexicons," 11 2014.

[12] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "SensEmbed: Learning sense embeddings for word and relational similarity*," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* Jul. 2015, pp. 95–105. [Online]. Available: https://aclanthology.org/P15-1010/

[13] Princeton University, "About WordNet," https://wordnet.princeton.edu/, 2010, wordNet. Princeton University.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[15] P. Sutor, Y. Aloimonos, C. Fermuller, and D. Summers-Stay, "Metaconcepts: Isolating context in word embeddings," in 2019 *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp.544–549.

[16] T. Hofmann, L. Cai, and M. Ciaramita, "Learning with taxonomies: classifying documents and words," in *Proceedings of the NIPS Workshop on Syntax, Semantics and Statistics*, British Columbia, Canada, 2003.

[17] Z. Wang, H. Chen, Z. Yuan, J. Wan, and T. Li, "Multiscale fuzzy entropy-based feature selection," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 9, pp. 3248–3262, 2023.

[18] J.-H. Wang, W.-J. Lee, and S.-J. Lee, "A kernel-based fuzzy clustering algorithm," in *First International Conference on Innovative Computing, Information and Control* - Volume I (ICICIC'06), vol. 1, 2006, pp. 550–553.

[19] S. Zeng, X. Wang, H. Cui, C. Zheng, and D. Feng, "A unified collaborative multikernel fuzzy clustering for multiview data*," IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1671–1687, 2018.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533

[22] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York: Springer-Verlag, 2002.

[23] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101,1904.

[24] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilit`a," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[25] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B ,* vol. 57, no. 1, pp. 289–300, 1995.

[26] B. Efron, "Bootstrap methods: another look at the jackknife*," The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[27] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum Press, 1981, classic monograph; contains derivation of FCM objective and updates.

[28] J. Lee, F. Chen, S. Dua, D. Cer, M. Shanbhogue, K. Chen, H. S. Vera, D. Salz, M. Boratko, V. Rao, Y. Cao, S. Kudugunta, F. Liu, V. Zhao, N. Fiedel, B. Alomaisi, J. Ainslie, T. Chen, X. Li, Z. Li, S. Kumar, T. Duerig, and Y. Yang, "Gemini embedding: Generalizable embeddings from gemini," 2025.

[29] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou, "Qwen3 embedding: Advancing text embedding and reranking through foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2506.05176

[30] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,* 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[31] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL-HLT 2009, Boulder, Colorado, 2009, pp. 19–27.

[32] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, Dec. 2015. [Online]. Available: https://aclanthology.org/J15-4004/

[33] E. Bruni, N. K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res*., vol. 49, pp. 1–47, 2014. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10857

[34] M. T. Pilehvar, D. Kartsaklis, V. Prokhorov, and N. Collier, Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, November* 2018, pp. 1391–1401. [Online]. Available: https://aclanthology.org/D18-1169/

[35] G. Halawi, G. Dror, E. Gabrilovich, and Y. Koren, "Large-scale learning of word relatedness with constraints," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser.* KDD '12. ACM, 2012, pp. 1406–1414.

[36] D. Gerz, I. Vuli´c, F. Hill, R. Reichart, and A. Korhonen, "Simverb-3500: A large-scale evaluation set of verb similarity," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, November 2016, pp. 2179–2189. [Online]. Available: https://aclanthology.org/D16-1235/

[37] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval.* New York: McGraw-Hill, 1983.

[38] K. Almohammadi, H. Hagras, B. Yao, A. Alzahrani, D. Alghazzawi, and G. Aldabbagh, "A Type-2 Fuzzy Logic Recommendation System for Adaptive Teaching," *Soft Computing*, vol. 21, no. 4, pp. 965–967, 2017.

[39] N. Sahab and H. Hagras, "Adaptive Non-singleton Type-2 Fuzzy Logic Systems: A Way Forward for Handling Numerical Uncertainties in Real World Applications," *International Journal of Computers, Communications and Control*, vol. 5, no. 3, pp. 503–529, Dec. 2011.

[40] F. Doctor and H. Hagras, "Neuro type-2 fuzzy based method for decision making," US Patent 8515884.

[41] V. Callaghan, M. Colley, H. Hagras, J. Chin, F. Doctor, and G. Clarke, "Programming iSpaces: A tale of two paradigms," in *Intelligent Spaces: The Application of Pervasive ICT*, A. Steventon and S. Wright, Eds., Springer-Verlag, ch. 24, pp. 389–421, Dec. 2005.