Introduction to AI - Coursework

2411243

1. Question 1: Semantic Analysis of Word Distances

1.1. m1

We selected four novels from Project Gutenberg: *A Room with a View*, *Anne of Green Gables*, *The Secret Adversary*, and *This Side of Paradise* to balance modern language with copyright compliance. The main content after `*** START OF THE PROJECT GUTENBERG EBOOK` marker was merged, removing license-related noise. Preprocessing included lowercasing, stripping non-alphabetic characters, and compressing whitespace. Tokenization used simple whitespace splitting, which is efficient but limited for contractions and proper nouns. Moreover, sentence segmentation was omitted, which suits word-order-based distances, but may reduce accuracy in sentence-level co-occurrence analysis. The final corpus has 336,934 words and 20,256 unique terms.

1.2. m2

To construct the word list L, we selected 100 semantically meaningful words based on the distribution of the speech part and the frequency of the word. Through NLTK, words were labeled and grouped into nouns, verbs, adjectives, and adverbs, and proportionally sampled, excluding proper nouns (`NNP`/`NNPS`) to reduce bias. After filtering stop words, custom noise words, and short words (≤ 2 characters), we chose high-frequency items from each group to generate a list of 100 unique terms. This approach ensures syntactic diversity and semantic relevance, but relies only on frequency, which ignores context or semantic structure.

1.3. m3

To measure the semantic relevance among the words in the L, we calculated the average positional distance between each pair of words in the corpus. Unlike sentence-level co-occurrence, this method gains finer-grained sequential patterns and avoids boundary-related noise. A greedy pairing algorithm matched the earliest available positions to compute mean absolute differences, forming a symmetric distance matrix. Although greedy matching may not be globally optimal, especially with imbalanced frequencies, it significantly reduces computational cost while retains interpretability, aligning with our goal of capturing global distribution trends. Furthermore, no window constraint was applied, preserving long-range co-occurrence patterns.

As shown in Figure 1, the heat map reveals structured regions, with darker shades indicating closer pairs and suggesting latent semantic clusters. Among the closest and furthest examples, "came-thought" (distance $= 3{,}787$) contrasts sharply with "men-said" (distance $= 200{,}250$), highlighting narrative differences. No missing values were observed, indicating stable and comprehensive statistics.

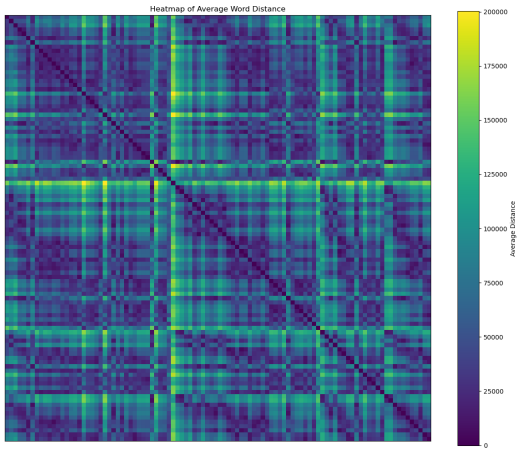

Figure 1: Heatmap of average word distance based on greedy positional pairing

1.4. m4

To explore the semantic structure of words in L, we applied unsupervised clustering to the average positional distance matrix, comparing KMeans, Hierarchical Clustering, and DBSCAN, evaluated by the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index.

Method	Silhouette	Calinski-H.	Davies-B.	Noise
KMeans	0.433	90.689	0.748	–
Hierarchical	0.386	102.671	0.681	–
DBSCAN	–	–	–	1.0

Table 1: Comparison of clustering methods on average positional distance matrix.

As shown in Table 1, although the KMeans scored slightly higher in Silhouette, the hierarchical clustering showed superior inter-cluster separation ($CH = 102.67$) and compactness ($DB = 0.681$), which better suited our dense and symmetric matrix. In contrast, DBSCAN failed to form valid clusters.

As shown in Figure 2, the dendrogram and cluster selection curves provide visual evidence for selecting $k = 3$. To determine the optimal number of clusters k, we applied four strategies: identifying a major merge jump at step 97 in the dendrogram; observing merge distance and inconsistency shifts between steps 86–97; and inspecting the silhouette score curve. Although a peak occurred at $k = 2$, we chose $k = 3$ as it better aligned with a major dendrogram merge jum, balancing structural clarity and interpretability. Notably, its silhouette score remained high, further supporting the decision.

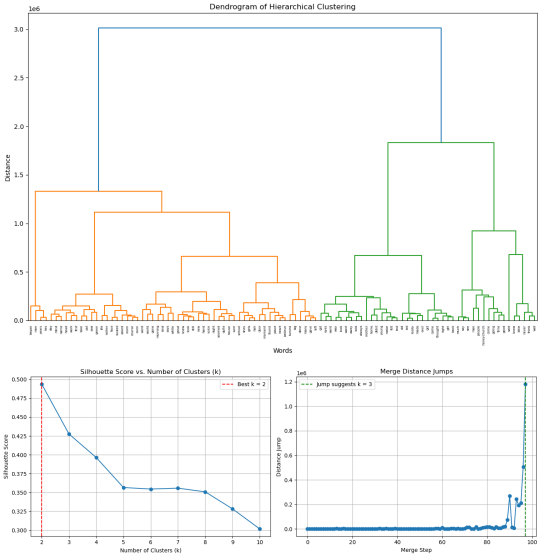

Figure 2: (Top) Dendrogram showing hierarchical clustering. (Bottom) Cluster number selection via silhouette score and merge distance.

The final clustering revealed three themes: Cluster 1 (environment and state-related words); Cluster 2 (abstract and cognitive terms); Cluster 3 (character actions and emotions).

1.5. m5

To more precisely characterize semantic relatedness among words in list L, we built a word co-occurrence graph from the corpus and applied Dijkstra's shortest path algorithm to compute pairwise semantic distances. Unlike global average position method, the graph-based approach captures local co-occurrence patterns and indirect semantic links, modeling how meaning propagates through context.

We scanned the corpus using a sliding window of size 20, and counted co-occurrences between target words in each window. Based on this, we constructed an undirected graph $G = (V, E)$, where:

- Nodes V represents the 100 words in list L.
- Edges E are formed if a pair of words co-occur within a window.
- Edge weights $w_{ij} = 1/f_{ij}$ are defined as the inverse of the co-occurrence frequency, assigning shorter distances to more frequent pairs.

We ran Dijkstra's algorithm for all nodes using networkx, computing all-pairs shortest paths to generate a symmetric distance matrix. The graph was fully connected (inf $= 0$) with an average shortest path of 0.0042 and diameter of 0.0122. The closest word pair was "said–well" (0.0003), and the furthest was "world–began" (0.0122). Top-degree nodes like man, girl, and eyes ($degree = 99$) serve as semantic hubs.

As illustrated in Figure 3, the heatmap displays distinct block-like regions along the diagonal, indicating clusters of closely connected words. Several vertical bands also reveal high-degree semantic hubs. Compared to the Figure 1, this graph-based distance map shows sharper and more localized clusters, reflecting stronger contextual associations.

However, this approach effectively captures indirect relations within local contexts, but depends on window size, and may produce overly dense graphs that reduce clustering sensitivity.

1.6. m6

To evaluate how semantic distance measures affect clustering, we applied hierarchical clustering ($k = 3$) to the average positional distance ($m3$) and the graph-based shortest path distance ($m5$) from L. Evaluation via Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), which yielded low scores ($ARI = 0.03$, $NMI = 0.27$), revealed low agreement between the two results, indicating divergent semantic structures.

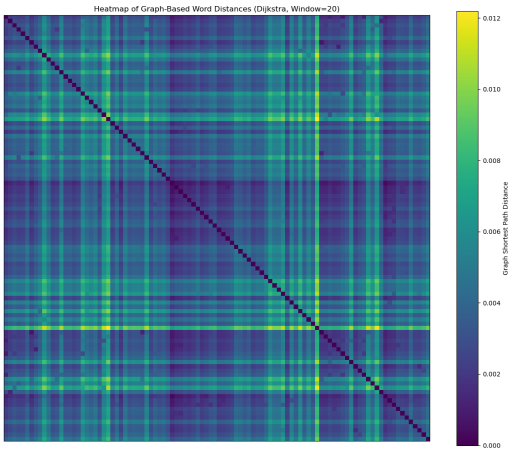

Figure 3: Heatmap of graph-based word distances computed using Dijkstra's algorithm with a co-occurrence window of 20.

As shown in Figure 4, the overlap between $m3$ and $m5$ clusters is low. The confusion matrix shows that most words from $m3$'s Clusters 2 and 3 were merged into Cluster 1 in $m5$, while $m3$'s Cluster 1 was dispersed across multiple $m5$ clusters. For example, terms like eyes, came, and hand—originally in $m3$ Cluster 2—shifted to $m5$ Cluster 1. These are typically high-frequency, interactional, or descriptive words, and their frequent co-occurrence makes them tightly linked in the graph-based representation.

This contrast highlights the differing semantic perspectives each method captures. Positional distance, aligned with narrative flow, groups words based on global order proximity, resulting in balanced but sometimes semantically diffuse clusters. In contrast, the graph-based method emphasizes local co-occurrence, form-

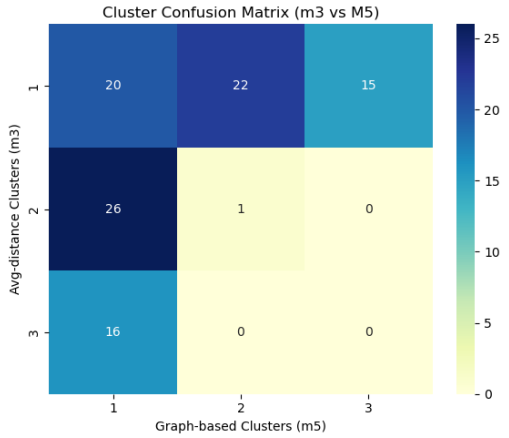

Figure 4: Confusion matrix comparing average-position-based clusters ($m3$) and graph-based clusters ($m5$). The diagonal indicates agreement; off-diagonal values reflect divergence.

ing dense semantic hubs that may over-merge functionally distinct terms. Each approach reflects a valid but partial view of word relationships; a hybrid strategy may offer a more robust and nuanced semantic segmentation in the future.

2. Question 2: Model Performance Compare on Nonlinear Classification

To evaluate model performance on nonlinear classification tasks, we generated synthetic 2D datasets via uniform sampling, with no artificial noise or vertical shifts. The decision boundary followed $y = ax^2 + x$, with points above and below the curve labeled as Class A and Class B, respectively. We varied curvature $a \in \{0.0, 0.5, 1.0, 2.0\}$ to control non-linearity and applied dynamic y-axis scaling to fully visualize each boundary.

We first applied logistic regression (LR) with standardized features and 5-fold cross-validation. LR performed well for $a = 0.0$, but both accuracy and F1-score declined as curvature increased. Misclassifications were concentrated near the curved boundary, particularly at extrema, with a noticeable rise in false negatives. Figure 5 visualizes the model's decision boundaries and error regions as curvature increases. While classification was nearly perfect for $a = 0.0$, performance visibly deteriorates for higher curvature levels, especially along the curved sections. As shown in Table 2, false positives and false negatives increase significantly, and F1-score gradually drops from 0.998 to 0.853.

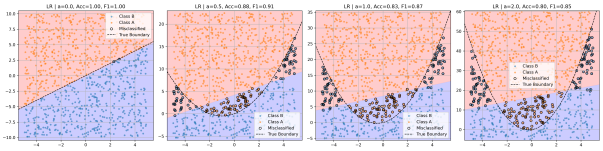

Figure 5: Logistic regression classification results under varying curvature values (a). Shaded regions indicate predicted classes; misclassified points are circled; the dashed curve shows the true boundary $y = ax^2 + x$.

a	Accuracy	F1 (Class A)	TN	FP	FN	TP
0.0	0.998	0.998	479	2	0	519
0.5	0.875	0.906	275	63	62	600
1.0	0.829	0.871	250	88	83	579
2.0	0.803	0.853	232	106	91	571

Table 2: Performance metrics of logistic regression under different curvature levels a. TN/FP/FN/TP indicate confusion matrix counts for Class A.

To solve the limitations of logistic regression, we trained a simple neural network (NN) with a single hidden layer of four ReLU units under same conditions with LR. As shown in Figure 6, the NN achieved high performance across all curvature levels, maintaining decision boundaries that closely tracked the true quadratic functions. Compared with LR, NN exhibited significantly fewer misclassifications near curved regions.

As reported in Table 3, both accuracy and F1-score remained above 0.96, even as curvature increased. Obviously, the number of false negatives which previously significant in

LR, was greatly reduced, indicating stronger recognition of Class A instances. These results demonstrate that even a small neural network has greater expressive power than linear models when dealing with nonlinear classification tasks.

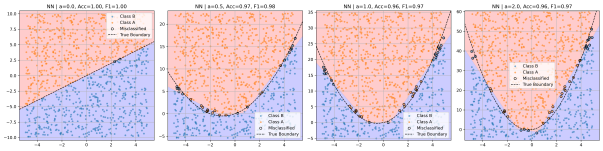

Figure 6: Neural network classification results under varying curvature values (a). The NN model uses a single hidden layer with 4 ReLU units. Misclassified points are circled. The dashed curve represents the true decision boundary $y = ax^2 + x$.

a	Accuracy	F1 (Class A)	TN	FP	FN	TP
0.0	0.998	0.998	479	2	0	519
0.5	0.974	0.980	325	13	13	649
1.0	0.964	0.973	319	19	17	645
2.0	0.962	0.971	321	17	21	641

Table 3: Neural network performance under varying curvature a. The network used a single hidden layer with 4 ReLU units. Confusion matrix counts refer to Class A.

To examine the impact of sample size on model performance, we fixed the curvature at $a = 1.0$ to represent a moderately nonlinear task where both LR and NN differ meaningfully. We evaluated four dataset sizes: $N = 100$ (to assess underfitting), 500 and 1000 (for convergence analysis), and 5000 (to test data sufficiency). As shown in Figure 7, logistic regression (LR) exhibited only modest improvement as N increased, constrained by its linear expressiveness. In contrast, the neural network (NN) steadily improved with more data.

A slight dip at $N = 1000$ may reflect suboptimal local minima during training, but performance recovered at $N = 5000$, achieving nearly 99% accuracy with no signs of overfitting. These results suggest that performance gains depend not only on data volume, but also on the model's capacity to learn from it.

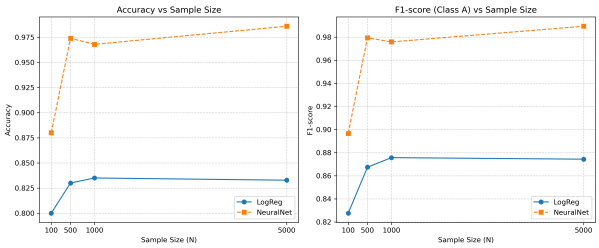

Figure 7: Impact of sample size on model performance at $a = 1.0$. Neural networks improved significantly with more data, while logistic regression plateaued early.

To evaluate the impact of class imbalance, we again fixed the curvature at $a = 1.0$ and $N = 1000$ to ensure a stable baseline and isolate the effect of label distribution. We varied the proportion of Class A in the dataset: 10% (rare), 50% (balanced), and 90% (majority). As shown in Figure 8, logistic regression (LR) exhibited a strong bias toward the majority class in the 10% condition, achieving 94.8% accuracy but only an F1-score of 0.67 for Class A.

In contrast, the neural network (NN) maintained robust performance across all proportions, with F1-scores above 0.86 even under severe imbalance, and reaching 0.99 when Class A was in the majority. These results highlight the limitations of accuracy as a metric under imbalance, and demonstrate the NN's superior robustness and fairness, attributed to its ability to learn adaptive nonlinear boundaries.

To investigate how neural network structure affects performance, we conducted three controlled experiments. First, we varied the number of neurons in a single hidden layer from 1 to 8. As shown in Figure 9, performance remained poor for fewer than 4 units, with a noticeable dip at 3—likely due to insufficient capacity or unstable convergence.

Starting from 4 units, both accuracy and F1-score increased sharply and quickly saturated,

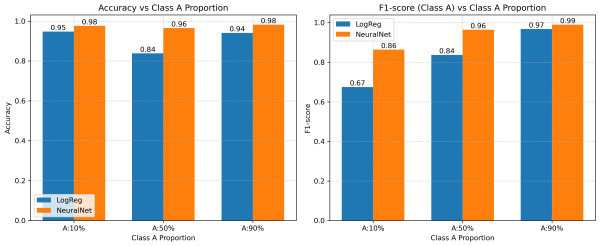

Figure 8: Impact of class imbalance on model performance at $a = 1.0$, $N = 1000$. Left: Accuracy; Right: F1-score (Class A). Neural networks demonstrate greater robustness to imbalance.

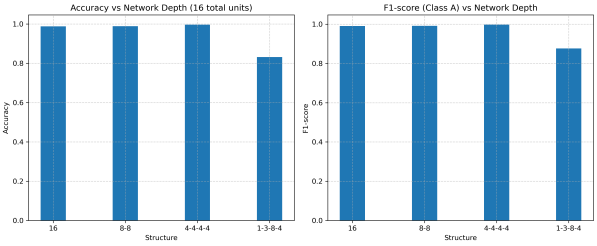

Figure 10: Neural network performance under different hidden layer distributions (total 16 units). Balanced designs yield higher accuracy and F1-score compared to asymmetric architectures.

suggesting that a minimum expressive threshold must be met for effective learning. Beyond that point, additional neurons offered minimal gains, indicating possible redundancy in representational capacity.

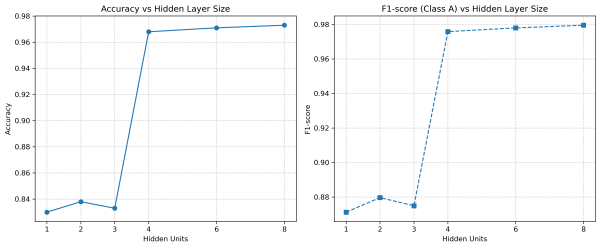

Figure 9: Effect of hidden layer size (number of neurons) on NN performance at fixed depth. Performance improves sharply after 4 units and then saturates.

Second, we investigated the effect of depth by distributing a fixed total of 16 hidden units across different architectures. As shown in Figure 10, the balanced structure (4-4-4-4) achieved the highest F1-score (above 0.99), indicating efficient feature transformation through moderate depth.

In contrast, the asymmetric configuration (1-3-8-4) led to a noticeable performance drop, despite using the same total number of neurons. These results suggest that network depth alone does not guarantee improvement; rather, effective performance depends on a thoughtful allocation of units to promote stable and uniform learning across layers.

Third, we fixed the layer width at 4 units and gradually increased the number of hidden layers from 1 to 5. As shown in Figure 11, performance improved notably up to 3 layers, but further depth provided no measurable gain.

This plateau suggests that moderate depth (around 3 layers) is sufficient for the current task, and additional layers may introduce unnecessary complexity without improving generalization. These results reinforce the importance of capacity-efficient architecture design.

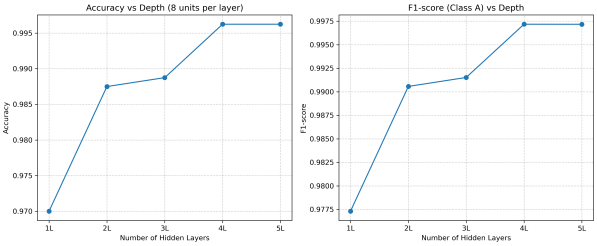

Figure 11: Effect of increasing network depth with fixed layer width (4 units). Performance improves up to 3 layers, then saturates.

In summary, neural network architecture plays a critical role in model performance. Effective capacity, depth, and layer allocation must be aligned with task complexity. The best-performing models are not necessarily the largest, but those that are structurally well-matched to the data.

3. Question 3: Rights and Personhood of Large Language Models

The rapid advancement of large language models (LLMs) like ChatGPT has intensified discussions on whether they should have rights or attributes of personhood. Current LLMs lack consciousness and autonomy, making them unsuitable as right-holders. However, technological and social shifts suggest that granting LLMs limited rights could become a serious consideration.

There is no current basis for granting rights to LLMs, as they lack consciousness and agency. According to John Searle's "Chinese Room" argument [1], LLMs merely manipulate symbols without genuine understanding or subjective experience, highlighting the gap between understanding and execution. However, his view reflects the early AI research and may underestimate the potential for complex systems to develop internal representations.

Thomas Metzinger similarly argues that AI cannot form a self-model and thus lack its own existence or intentions [2], positioning LLMs as tools rather than potential rights-holders. Whereas his criteria set an high threshold, overlooking how gradual technological evolution might reshape cognitive modeling. Rights judgments to LLMs must evolve with technology advance.

Nevertheless, while LLMs currently lack the basic conditions for rights-holding, granting them limited rights in the future worth reasonable discussion, especially from the perspectives of social utility and philosophical egalitarianism.

From a utilitarian standpoint, societies have granted rights and obligations to non-natural persons like corporations to promote stability. to facilitate economic and social development. Similarly, granting LLMs limited personhood could clarify the ownership of their creative outputs (such as autonomously generated content) and assign responsibility in decision-making contexts. Such legal frameworks would help regulate AI behavior and mitigate uncertainties around AI agency. However, premature rights assignment could enable LLMs to be misused as tools for evading responsibility or creating false entities. Therefore, any LLMs' rights must be strictly limited and balanced with ethical and technical safeguards.

From an egalitarian standpoint, Carbon-based human intelligence and silicon-based artificial intelligence are fundamentally similar. Human neurons transmit information via electrical impulses, while neural networks operate through weighted connections, both reflecting complex patterns of information flow and activation. Defining rights solely by material composition reflects anthropocentric bias. Conceptually, current LLMs could be viewed as early-stage simulations of a silicon-based life, like primitive cellular systems. However, such philosophical assumptions currently lack empirical support, and the independent consciousness in artificial agents depends critically on internal complexity and autonomy. Thus, while today's LLMs do not qualify, future silicon-based intelligences with human-comparable cognition may warrant serious right consideration.

In conclusion, current LLMs lacking subjective consciousness and autonomy, cannot yet claim rights or personhood. However, as technology and social evolve, utilitarian needs and principles of informational equality may force reconsideration, and balancing science with ethic will be critical. Ultimately, as silicon-based minds approach human cognition, human may rethink what constitutes life, and who—or what—deserves respect and rights.

REFERENCES

[1] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

[2] Metzinger, T. (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.