Mammal vs. Bird Name Embeddings: Methods and Findings

Songhee Beck

April 2, 2025

1 Data Sources and Loading

1.1 Mammal Names (`mammal_names` (4162 names))

- Downloaded from Hugging Face dataset: `willcb/mammal-names`
- `'Common Name'` column is extracted, stored in the Python list `mammal_names`.
- *Number of Mammal Names*: `len(mammal_names) == 4162`.

1.2 Bird Names (`bird_names` (10976 names))

- Downloaded from Kaggle dataset: `thepushkarp/common-bird-names`
- Bird names in `'Common Bird Names'` extracted and stored in Python list `bird_names`.
- *Number of Bird Names*: `len(bird_names) == 10976`.

2 Sentence Embeddings (Sentence Transformers)

Objective: Compare how close (via cosine similarity) the embeddings for word "mammal" is to centroid of mammal and centroid of Bird. Compare how close the embeddings for word "Bird" is to centroid of mammal and centroid of Bird.

2.1 Model

- `all-mpnet-base-v2` from the `sentence-transformers` library.
- Embeddings are normalized (`normalize_embeddings=True`).

2.2 Create Embeddings for mammal names and bird names

```
mammal_embeddings = model.encode(mammal_names)
bird_embeddings   = model.encode(bird_names)
```

Each name is turned into a dense vector (768D in `all-mpnet-base-v2`), then stored in arrays.

2.3 Clustering

- `KMeans(n_clusters=1)` for mammals to get a single centroid `centroid_mammal`.
- `KMeans(n_clusters=1)` for birds to get `centroid_bird`.

2.4 Create embeddings for words "mammal" & "bird"

```
embedding_word_mammal = model.encode("mammal")
embedding_word_bird   = model.encode("bird")
```

2.5 Cosine Similarities

- Compare how close "mammal" is to `centroid_mammal` vs. `centroid_bird`.
- Compare how close "bird" is to `centroid_mammal` vs. `centroid_bird`.
- Also check `centroid_mammal` vs. `centroid_bird`, and ("mammal", "bird").

Cosine Similarity Results

```
word "mammal" <-> centroid_mammal: 0.6378998
word "mammal" <-> centroid_bird:   0.47944468
word "bird"   <-> centroid_mammal: 0.4984303
word "bird"   <-> centroid_bird:   0.6708431

centroid_mammal <-> centroid_bird: 0.8634571
word "mammal" <-> word "bird": 0.50262505
```

Key Findings

- `"mammal"` vector is more similar to the mammal centroid than the bird's centroid.
- `"bird"` vector is more similar to the bird centroid than the mammal's centroid.
- This suggests that the model semantically associates the word `"mammal"` more strongly with actual mammal names, and likewise for `"bird"`.
- The centroid of all mammal names vs. the centroid of all bird names has fairly high similarity (~ 0.86), indicating that the average mammal name embedding is relatively close to the average bird name embedding in this model.

Visualization

PCA (2D)

- Each category (mammals and birds) is plotted as a scatter of points.
- The single cluster centroids (red vs. green markers) are added.
- The words "mammal" and "bird" are also shown (as star markers).
- The two categories have partial overlap but still form distinct regions.

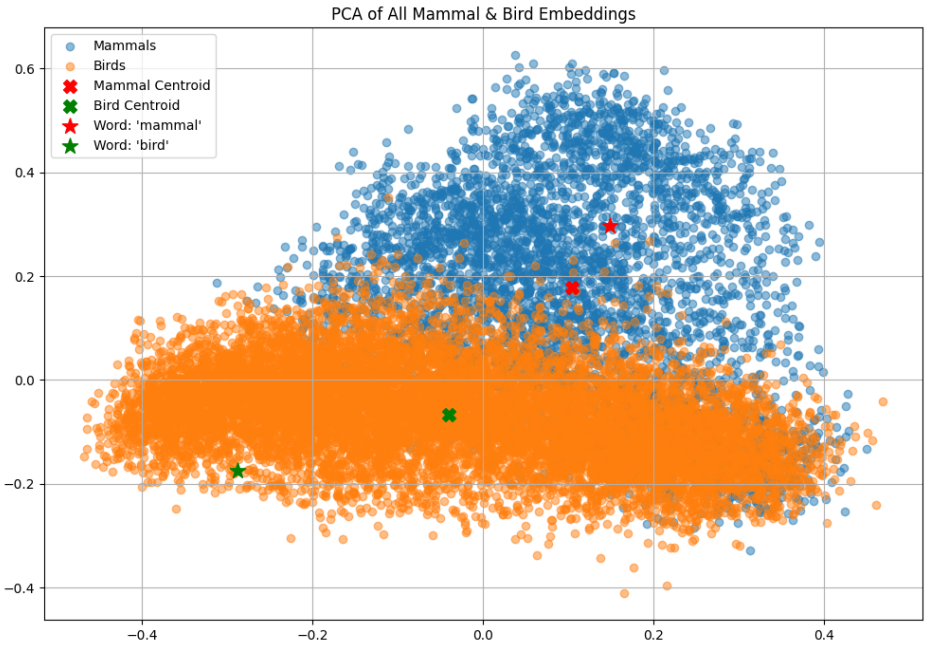

Figure 1: PCA of mammal and bird embeddings.

t-SNE (2D)

- `TSNE(n_components=2, perplexity=30)` is run on the combined embeddings.
- mammals vs. birds are plotted.

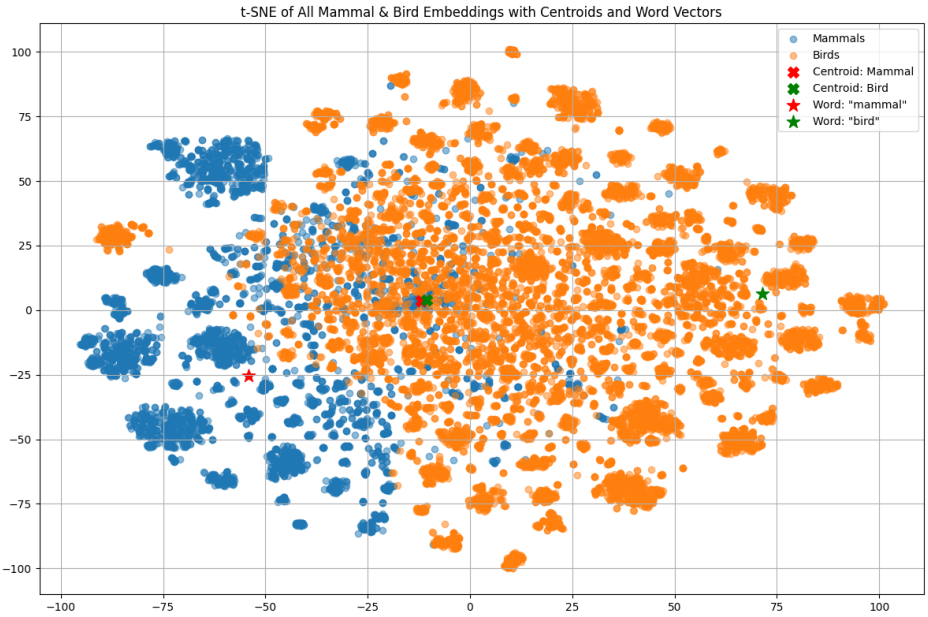

Figure 2: 2D t-SNE visualization of mammal and bird embeddings.

UMAP (3D)

- UMAP (`n_components=3`), applied to all mammal and bird embeddings
- Clusters were plotted in 3D using a very low opacity (`alpha=0.01`) to reduce visual clutter and highlight semantic anchors.

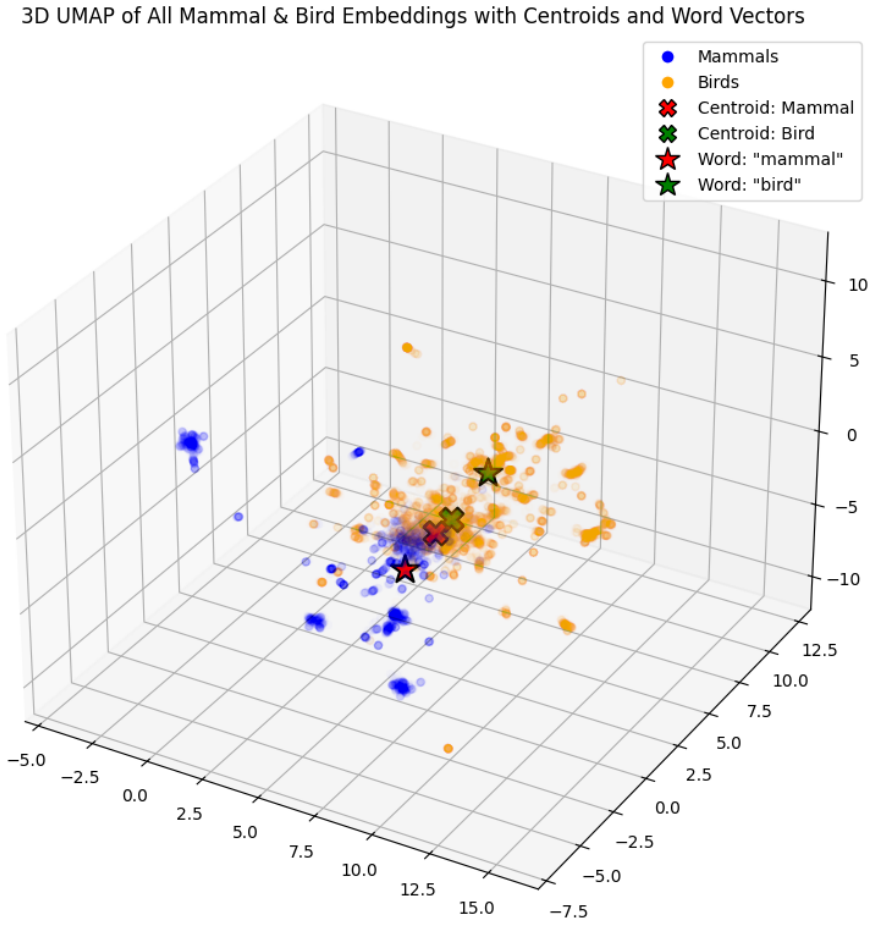

Figure 3: 3D UMAP of All Mammal and Bird Embeddings with Centroids and Word Vectors

Result

- The word embedding "mammal" clusters near mammal names, while "bird" clusters near bird names in PCA, t-SNE, and UMAP plots, indicating strong semantic coherence for each category.

3 Combined Clustering (K=2)

The mammal and bird embeddings are combined using `np.vstack`, and `KMeans(n_clusters=2)` is applied to the full set. The resulting clusters are then matched to their dominant categories based on majority labels.

- The cluster containing more mammal-labeled items is designated as the `more_mammal_cluster`, and the other is designated as the `more_bird_cluster`.

- **Assigned:** Cluster 0 → More Mammals, Cluster 1 → More Birds
- **Cluster 0 contains:** Mammals: 3537, Birds: 4788
- **Cluster 1 contains:** Mammals: 625, Birds: 6188

- Cosine similarities are computed between the embeddings for the words "mammal" and "bird" and the centroids of each cluster.

Result:

```
word "mammal" <-> centroid_mammal: 0.56992877
word "mammal" <-> centroid_bird:   0.47013846
word "bird"   <-> centroid_mammal: 0.52894235
word "bird"   <-> centroid_bird:   0.6988307
```

Visualization

PCA (2D)

- The embeddings were projected into two dimensions using PCA for visualization. Each point is colored according to its KMeans cluster assignment: Cluster 0 corresponds to the `more_mammal_cluster`, and Cluster 1 corresponds to the `more_bird_cluster`.

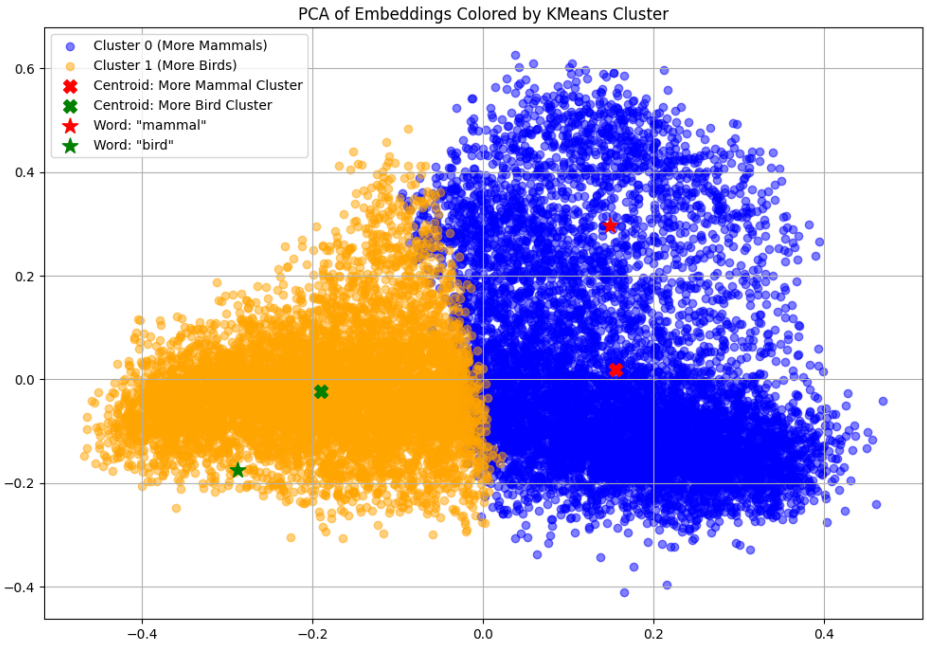

Figure 4: PCA plot of mammal and bird embeddings colored by KMeans clusters.

t-SNE (2D)

- `TSNE(n_components=2, perplexity=30)` is run on the combined embeddings.

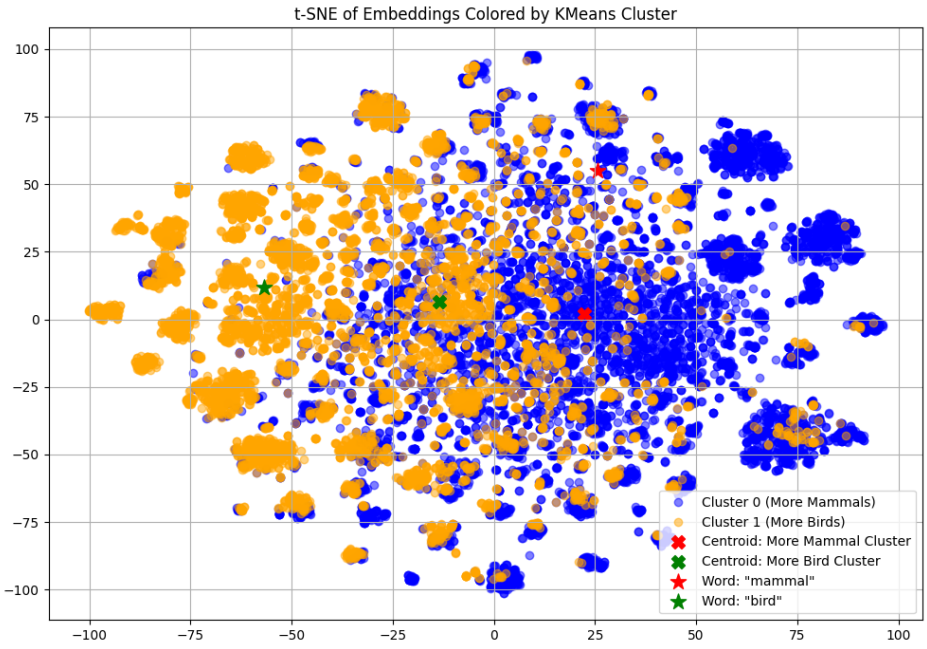

Figure 5: 2D t-SNE projection of embeddings colored by KMeans clusters.

UMAP (3D)

- All mammal and bird embeddings were projected into 3D space using UMAP (`n_components=3`), with KMeans cluster assignments provided via `sampled_labels`.
- Clusters were plotted in 3D using a very low opacity (`alpha=0.01`) to reduce visual clutter and highlight semantic anchors.

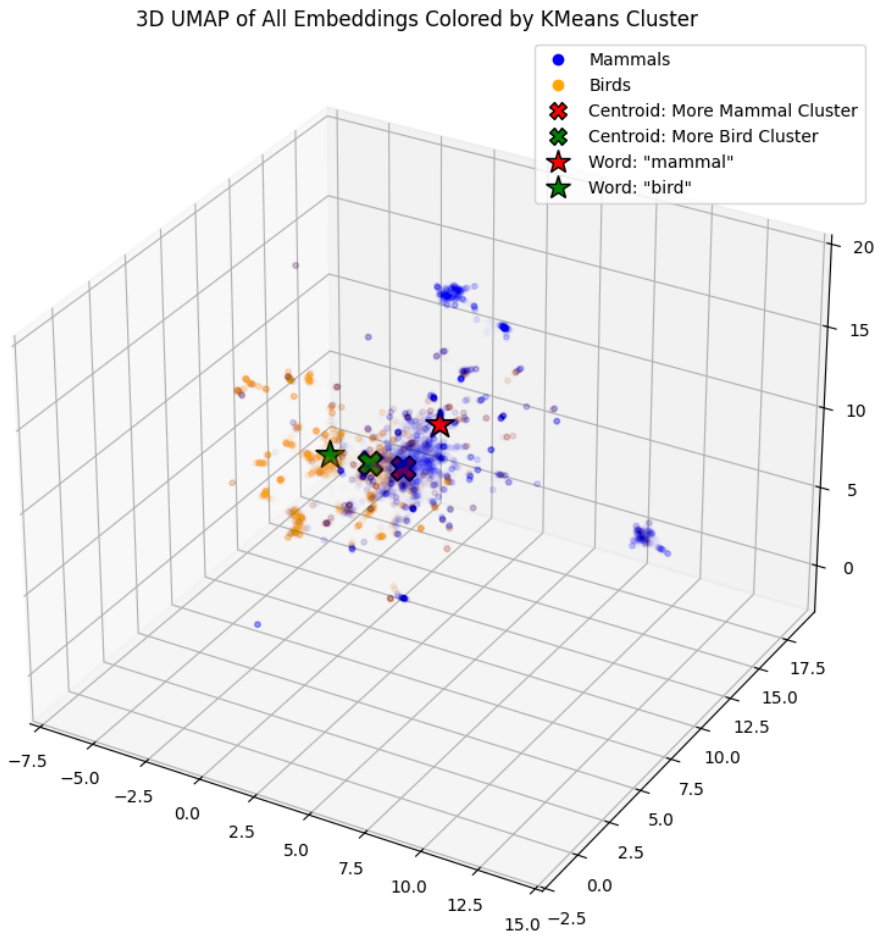

Figure 6: 3D UMAP of All Embeddings Colored by KMeans Cluster

4 Layer Analysis (TinyLlama-1.1B-Chat-v1.0)

Objective: Evaluate how LLM model layers capture the concept of "mammal" vs. "bird."

4.1 Mammal Embeddings Across All Layers

1. For each mammal name:
   - Hidden states from all 23 layers were extracted.
   - The last token vector was selected from each layer.
   - These vectors were stored in a dictionary keyed by `layer_idx`.
2. A single mammal centroid was computed for each layer:
   - `KMeans(n_clusters=1)` was applied to the layer's token vectors.
3. The centroid was compared to the embedding for the word `"mammal"` at the same layer:
   - Cosine similarity was measured using `cosine_similarity(mammal_centroid, "mammal")`.

Results:

```
Layer  0: cosine_similarity(centroid <-> word 'mammal') = 0.1634
Layer  1: cosine_similarity(centroid <-> word 'mammal') = 0.7002
Layer  2: cosine_similarity(centroid <-> word 'mammal') = 0.5161
Layer  3: cosine_similarity(centroid <-> word 'mammal') = 0.3345
Layer  4: cosine_similarity(centroid <-> word 'mammal') = 0.3919
Layer  5: cosine_similarity(centroid <-> word 'mammal') = 0.4837
Layer  6: cosine_similarity(centroid <-> word 'mammal') = 0.5696
Layer  7: cosine_similarity(centroid <-> word 'mammal') = 0.5631
Layer  8: cosine_similarity(centroid <-> word 'mammal') = 0.5726
Layer  9: cosine_similarity(centroid <-> word 'mammal') = 0.6067
Layer 10: cosine_similarity(centroid <-> word 'mammal') = 0.5734
Layer 11: cosine_similarity(centroid <-> word 'mammal') = 0.6086
Layer 12: cosine_similarity(centroid <-> word 'mammal') = 0.6135
Layer 13: cosine_similarity(centroid <-> word 'mammal') = 0.6534
Layer 14: cosine_similarity(centroid <-> word 'mammal') = 0.6555
Layer 15: cosine_similarity(centroid <-> word 'mammal') = 0.6839
Layer 16: cosine_similarity(centroid <-> word 'mammal') = 0.7333
Layer 17: cosine_similarity(centroid <-> word 'mammal') = 0.7411
Layer 18: cosine_similarity(centroid <-> word 'mammal') = 0.7787
Layer 19: cosine_similarity(centroid <-> word 'mammal') = 0.8019
Layer 20: cosine_similarity(centroid <-> word 'mammal') = 0.7979
Layer 21: cosine_similarity(centroid <-> word 'mammal') = 0.8053
Layer 22: cosine_similarity(centroid <-> word 'mammal') = 0.8000

Best conceptual layer = 21, cosine similarity = 0.8053
```

- Cosine similarity consistently increased across layers, with higher layers showing stronger alignment.
- The best conceptual match for the word `"mammal"` occurred at **layer 21**, achieving a cosine similarity of **0.8053**.

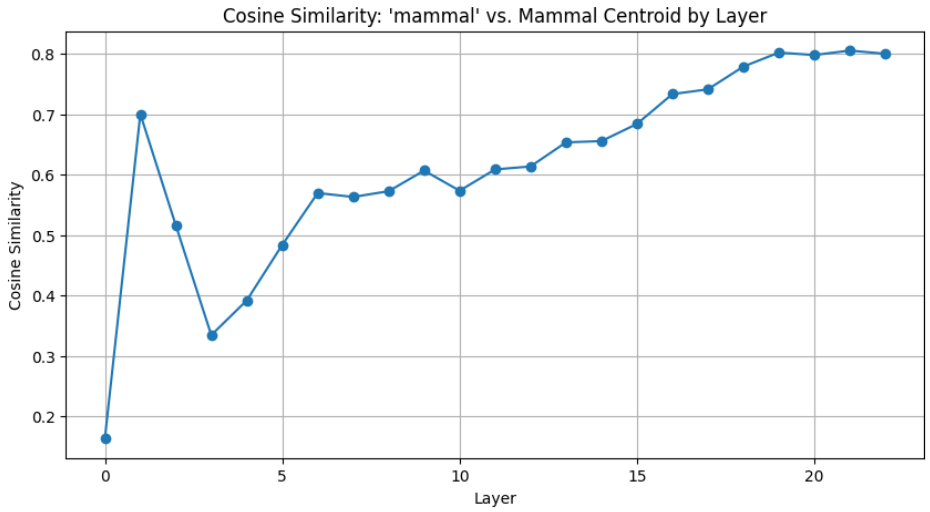

Figure 7: Cosine Similarity: 'mammal' vs. Mammal Centroid by Layer

4.2 Bird Embeddings Across All Layers

1. For each bird name:

   - Hidden states from all 23 layers were extracted.
   - The last token vector was selected from each layer.
   - Performed `KMeans(n_clusters=1)` to get a "bird centroid" at each layer.
   - "bird centroid" at each layer was compared to the word "bird"'s embedding at that same layer.

Results

```
Layer  0: cosine_similarity(centroid <-> word 'bird') = 0.1794
Layer  1: cosine_similarity(centroid <-> word 'bird') = 0.4410
Layer  2: cosine_similarity(centroid <-> word 'bird') = 0.5632
Layer  3: cosine_similarity(centroid <-> word 'bird') = 0.4104
Layer  4: cosine_similarity(centroid <-> word 'bird') = 0.4232
Layer  5: cosine_similarity(centroid <-> word 'bird') = 0.5032
Layer  6: cosine_similarity(centroid <-> word 'bird') = 0.5648
Layer  7: cosine_similarity(centroid <-> word 'bird') = 0.5573
Layer  8: cosine_similarity(centroid <-> word 'bird') = 0.4898
Layer  9: cosine_similarity(centroid <-> word 'bird') = 0.4895
Layer 10: cosine_similarity(centroid <-> word 'bird') = 0.4896
Layer 11: cosine_similarity(centroid <-> word 'bird') = 0.5222
Layer 12: cosine_similarity(centroid <-> word 'bird') = 0.5206
Layer 13: cosine_similarity(centroid <-> word 'bird') = 0.5472
Layer 14: cosine_similarity(centroid <-> word 'bird') = 0.5502
Layer 15: cosine_similarity(centroid <-> word 'bird') = 0.5592
Layer 16: cosine_similarity(centroid <-> word 'bird') = 0.6280
Layer 17: cosine_similarity(centroid <-> word 'bird') = 0.6475
```

```
Layer 18: cosine_similarity(centroid <-> word 'bird') = 0.6731
Layer 19: cosine_similarity(centroid <-> word 'bird') = 0.7022
Layer 20: cosine_similarity(centroid <-> word 'bird') = 0.7260
Layer 21: cosine_similarity(centroid <-> word 'bird') = 0.7140
Layer 22: cosine_similarity(centroid <-> word 'bird') = 0.6453

Best conceptual layer = 20 with cosine similarity = 0.7260
```

- Higher layers show greater alignment between centroid bird and the word "bird".
- **Layer 20** was the best for the word "bird" (similarity ≈ 0.7260).

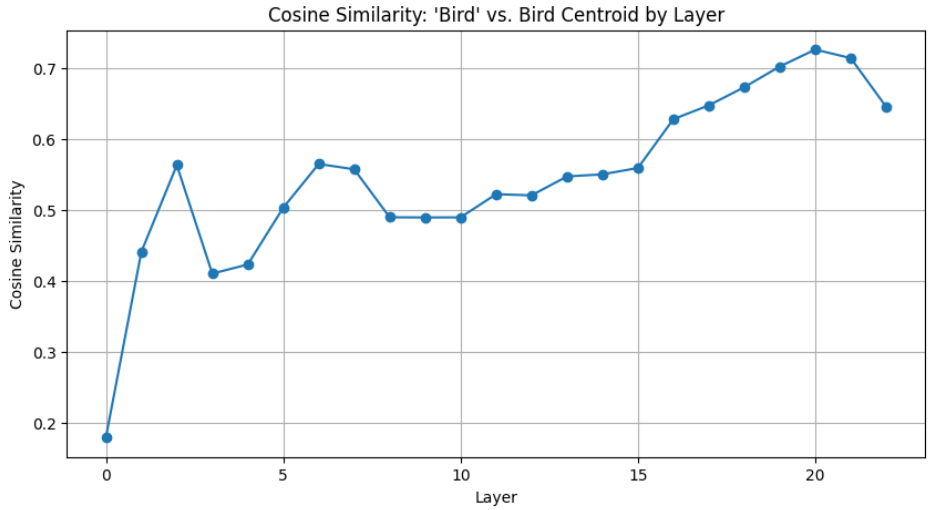

Figure 8: Cosine Similarity: 'Bird' vs. Bird Centroid by Layer

4.3 KMeans (k=2) on Mammals + Birds

For each layer, mammal and bird embeddings were combined into a single matrix, and `KMeans(n_clusters=2)` was applied:

- A majority vote was used to match each discovered cluster to either "mammal" or "bird," based on which label dominated the cluster.
- Centroids for each cluster were then extracted, and the concept words "mammal" and "bird" were compared to these centroids by cosine similarity across layers.

Results

```
Layer  0
  'word mammal' <-> mammal centroid: 0.2706
  'word mammal' <-> bird centroid:   0.1194
  'word bird'   <-> mammal centroid: 0.2009
  'word bird'   <-> bird centroid:   0.0176

Layer  1
  'word mammal' <-> mammal centroid: 0.7581
```

```
  'word mammal' <-> bird centroid:   0.4381
  'word bird'   <-> mammal centroid: 0.4570
  'word bird'   <-> bird centroid:   0.2423

Layer  2
  'word mammal' <-> mammal centroid: 0.5641
  'word mammal' <-> bird centroid:   0.4378
  'word bird'   <-> mammal centroid: 0.5226
  'word bird'   <-> bird centroid:   0.6697

Layer  3
  'word mammal' <-> mammal centroid: 0.3565
  'word mammal' <-> bird centroid:   0.1502
  'word bird'   <-> mammal centroid: 0.4308
  'word bird'   <-> bird centroid:   0.2004

Layer  4
  'word mammal' <-> mammal centroid: 0.3829
  'word mammal' <-> bird centroid:   0.2452
  'word bird'   <-> mammal centroid: 0.4463
  'word bird'   <-> bird centroid:   0.2586

Layer  5
  'word mammal' <-> mammal centroid: 0.4895
  'word mammal' <-> bird centroid:   0.1679
  'word bird'   <-> mammal centroid: 0.5263
  'word bird'   <-> bird centroid:   0.2069

Layer  6
  'word mammal' <-> mammal centroid: 0.5515
  'word mammal' <-> bird centroid:   0.1838
  'word bird'   <-> mammal centroid: 0.5790
  'word bird'   <-> bird centroid:   0.2178

Layer  7
  'word mammal' <-> mammal centroid: 0.5423
  'word mammal' <-> bird centroid:   0.4825
  'word bird'   <-> mammal centroid: 0.5770
  'word bird'   <-> bird centroid:   0.4864

Layer  8
  'word mammal' <-> mammal centroid: 0.5584
  'word mammal' <-> bird centroid:   0.4469
  'word bird'   <-> mammal centroid: 0.5120
  'word bird'   <-> bird centroid:   0.4681

Layer  9
  'word mammal' <-> mammal centroid: 0.5971
  'word mammal' <-> bird centroid:   0.4871
  'word bird'   <-> mammal centroid: 0.5099
  'word bird'   <-> bird centroid:   0.4649

Layer 10
  'word mammal' <-> mammal centroid: 0.5664
```

```
  'word mammal' <-> bird centroid:   0.4663
  'word bird'   <-> mammal centroid: 0.5137
  'word bird'   <-> bird centroid:   0.4618

Layer 11
  'word mammal' <-> mammal centroid: 0.6034
  'word mammal' <-> bird centroid:   0.5237
  'word bird'   <-> mammal centroid: 0.5528
  'word bird'   <-> bird centroid:   0.4941

Layer 12
  'word mammal' <-> mammal centroid: 0.6067
  'word mammal' <-> bird centroid:   0.5479
  'word bird'   <-> mammal centroid: 0.5719
  'word bird'   <-> bird centroid:   0.4919

Layer 13
  'word mammal' <-> mammal centroid: 0.6415
  'word mammal' <-> bird centroid:   0.6055
  'word bird'   <-> mammal centroid: 0.5956
  'word bird'   <-> bird centroid:   0.5231

Layer 14
  'word mammal' <-> mammal centroid: 0.6451
  'word mammal' <-> bird centroid:   0.6062
  'word bird'   <-> mammal centroid: 0.5924
  'word bird'   <-> bird centroid:   0.5230

Layer 15
  'word mammal' <-> mammal centroid: 0.6684
  'word mammal' <-> bird centroid:   0.6431
  'word bird'   <-> mammal centroid: 0.6011
  'word bird'   <-> bird centroid:   0.5291

Layer 16
  'word mammal' <-> mammal centroid: 0.7136
  'word mammal' <-> bird centroid:   0.7013
  'word bird'   <-> mammal centroid: 0.6695
  'word bird'   <-> bird centroid:   0.6069

Layer 17
  'word mammal' <-> mammal centroid: 0.7374
  'word mammal' <-> bird centroid:   0.6458
  'word bird'   <-> mammal centroid: 0.6624
  'word bird'   <-> bird centroid:   0.6213

Layer 18
  'word mammal' <-> mammal centroid: 0.7740
  'word mammal' <-> bird centroid:   0.6873
  'word bird'   <-> mammal centroid: 0.7025
  'word bird'   <-> bird centroid:   0.6441

Layer 19
  'word mammal' <-> mammal centroid: 0.7783
```

```
  'word mammal' <-> bird centroid:   0.7424
  'word bird'   <-> mammal centroid: 0.7507
  'word bird'   <-> bird centroid:   0.6767

Layer 20
  'word mammal' <-> mammal centroid: 0.7913
  'word mammal' <-> bird centroid:   0.7143
  'word bird'   <-> mammal centroid: 0.7381
  'word bird'   <-> bird centroid:   0.7003

Layer 21
  'word mammal' <-> mammal centroid: 0.7901
  'word mammal' <-> bird centroid:   0.7400
  'word bird'   <-> mammal centroid: 0.7471
  'word bird'   <-> bird centroid:   0.6864

Layer 22
  'word mammal' <-> mammal centroid: 0.7910
  'word mammal' <-> bird centroid:   0.7267
  'word bird'   <-> mammal centroid: 0.7141
  'word bird'   <-> bird centroid:   0.6068

Best Layers:
 Highest 'word mammal' <-> mammal centroid: Layer 20 (similarity = 0.7913)
 Highest 'word bird'   <-> bird centroid:   Layer 20 (similarity = 0.7003)
```

Findings

- The best layer for both "mammal" → mammal centroid and "bird" → bird centroid in that approach was **layer 20**.
- This suggests that in the higher layers, TinyLlama's internal representation more strongly reflects broad semantic concepts ("mammal," "bird," etc.).

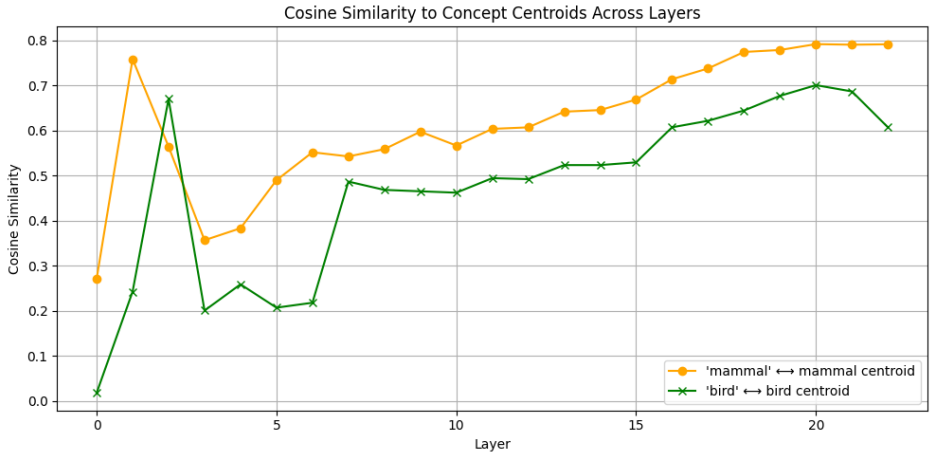

Figure 9: Cosine Similarity to Concept Centroids Across Layers

PCA (2D)

- PCA for **layer 20**.

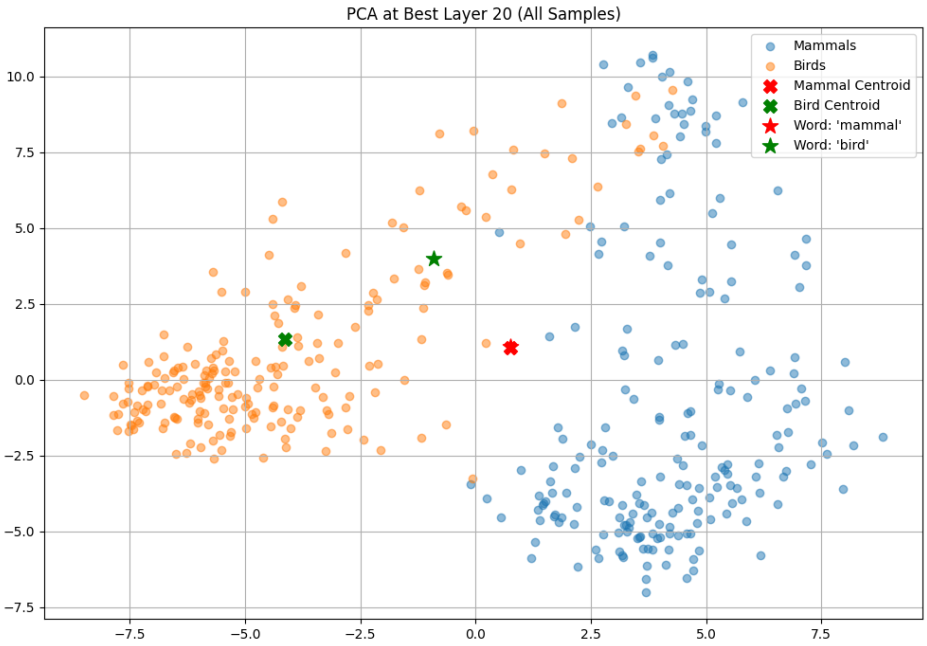

Figure 10: PCA at Best Layer 20

t-SNE (2D)

- t-SNE for **layer 20**.

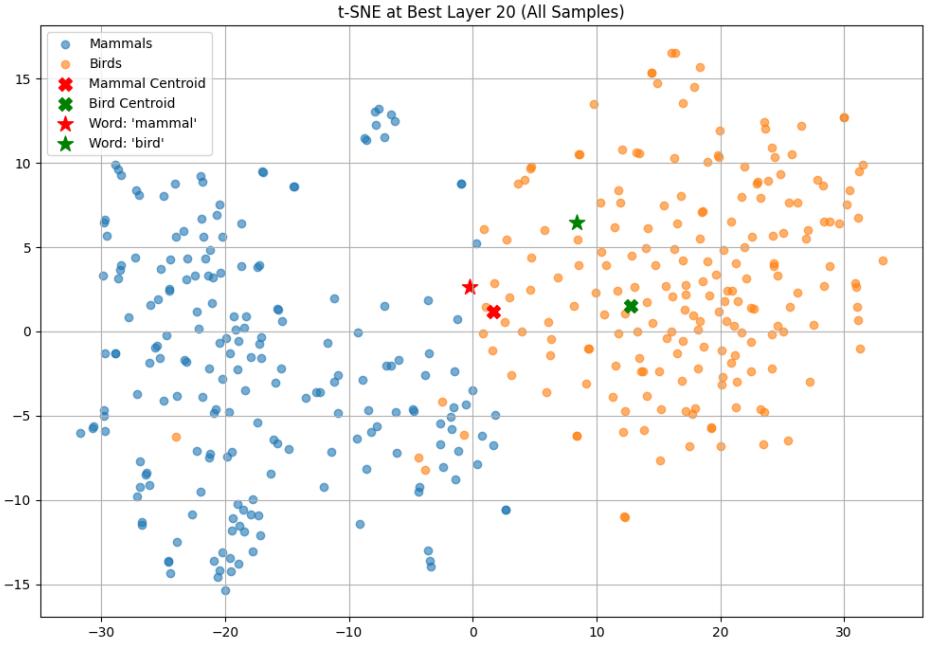

Figure 11: t-SNE at Best Layer 20

UMAP (3D)

- UMAP for **layer 20**.

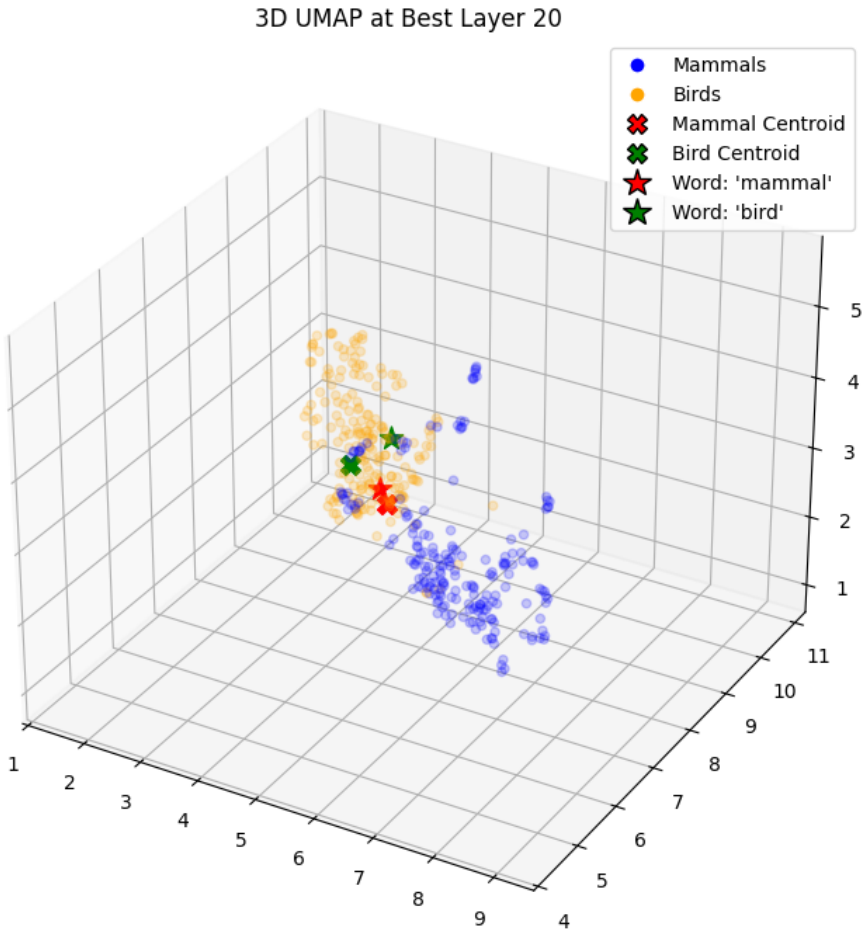

Figure 12: 3D UMAP at Best Layer 20

5 Final Conclusions

- **Sentence Transformers (all-mpnet-base-v2)** effectively separate mammal vs. bird names.
  - The word "mammal" is closest to the mammal centroid and the word "bird" is closest to the bird centroid.
- **TinyLlama Layer**:
  - Lower layers are more syntactic or local in representation while higher layers capture more semantic distinctions.
  - For the word "mammal," the top conceptual alignment peaks around **layer 21**.
  - For "bird," the top conceptual alignment is around **layer 20**.

  - When using KMeans(k=2) on combined mammal and bird embeddings, the best layer for semantic alignment is **layer 20**.

- **Mammal–Bird Overlap**:
  - Even though mammals and birds are conceptually distinct, their average embeddings are close and partially overlap.