Context-Aware Automatic Coding of Category Fluency Using LLMs

Pranav Tadepalli, Gilbert Mao, Ivy He, Yeonsoo Chang, Sashank Varma
Georgia Institute of Technology

Abstract

Category fluency tasks, which involve retrieval from semantic memory, help reveal the clustered structure of human semantic memory. However, traditional coding schemes for this task rely on fixed groupings that do not capture humans' varied semantic clusters and retrieval strategies. We evaluate whether the attention patterns of Large Language Models (LLMs) can help generate tailored coding schemes that better reveal the structure of semantic memory within and across people. Using the Hills et al. (2012) data, we prefill the LLM with each human-generated sequence to derive a per-sequence coding scheme. These attention-derived groups (from middle model layers) show inter-item response time increases at subcategory boundaries – a key prediction of optimal foraging theory – that are up to 30% greater than when using the standard, fixed Hills et al. (2012) scheme. We also evaluate the cognitive plausibility of LLM-generated sequences, finding that they are broadly consistent with the predictions of optimal foraging theory.

Keywords: Category Fluency; Large Language Models; Attention Weights; Optimal Foraging Theory

Introduction

Category fluency tasks have been used to study the organization of semantic memory in cognitive science, neuropsychology, and neuroscience studies. In this task, participants are given a category (e.g., Animals) and asked to generate exemplars over a fixed period of time. Troyer et al. (1997) documented a pattern of semantic clustering and controlled switching, where participants produce items from one subcategory (e.g., Pets) in a burst before switching to another (e.g., Farm Animals). This pattern has been foundational for understanding the organization of semantic memory.

To quantify the characteristics of a participant's clustering behavior (e.g., cluster size), the sequence of exemplars they produce must be coded into subcategories. Traditional approaches use fixed coding schemes assumed to hold for all individuals; these take the form of static taxonomies, static word embeddings, or static semantic graphs. Such approaches cannot explain individual differences in semantic memory (i.e., one person's subcategories may not be another's) nor can they account for the dynamics of sequence generation, where clusters might emerge in a context-specific way. This limits how well they can capture fluent retrieval from semantic memory (Kumar et al., 2025).

Large Language Models (LLMs) potentially enable a new approach to coding exemplars, and their (inspectable) internal mechanisms offer a new source of ideas about theoretical mechanisms (Vaswani et al., 2017). Studies showing that LLMs capture aspects of human semantic memory and of human language production have led to recent work investigating the potential of LLMs as cognitive models of category fluency (Heineman et al., 2024; Agarwal et al., 2025; Qiu et al., 2025; Zarrieß et al., 2025). The current study builds on this work, exploring the use of LLMs to automatically code the exemplar sequences produced by humans. It also further evaluates the potential of LLM generations as models of human category fluency. Specifically, the current study addresses three research questions: (1) Can LLM attention weights be used to derive the semantic subcategories of an individual, leading to a better understanding of retrieval from semantic memory than fixed coding schemes? (2) Do humans and LLMs show alignment with the MVT (Marginal Value Theorem) at LLM-derived and Hills et al. (2012) switch boundaries? (3) Do LLM-generated sequences share the same statistical structure (size of clusters and switch frequency) as human sequences?

Related Works

Theories and Models of Category Fluency

Clustering and Switching Troyer et al. (1997) documented a pattern in people's sequences of (1) clusters of exemplars of a given subcategory and (2) switches between subcategories. They proposed that people alternate between two phases during semantic retrieval. In the switching phase, people use the category cue (e.g., Animals) to search semantic memory for a new subcategory (e.g., Pets). In the clustering phase, people use the subcategory as a cue to semantic memory, retrieving exemplars (e.g., dog) until an exhaustion criterion is reached. At this point, they return to the switching phase to identify a new subcategory cue (e.g., Farm Animals).

Importantly, Troyer et al. (1997) provided a scheme for coding exemplar sequences, to be able to identify subcategory clusters and thus switches between clusters. This scheme has found widespread use in subsequent studies.

Optimal Foraging Hills et al. (2012) introduced the optimal foraging model of category fluency, drawing an analogy between humans 'foraging' their semantic memories for information and animals foraging their environment for food. Focusing on the switching phase, it proposes that exemplar retrieval from a 'patch' of semantic memory (e.g., the subcat-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

egory Pets) does not continue until that patch is completely exhausted, for this would be suboptimal in terms of total yield given total cost (i.e., time). Rather, it continues until the estimated cost of continuing in the current patch is greater than the estimated cost of switching (i.e., traveling to) a new patch. This switch rule is given by the Marginal Value Theorem (MVT) (Charnov, 1976). Hills et al. (2012) showed that response times increase at switches (e.g., between generating the last exemplar of the Pets subcategory and the first exemplar of the Farm Animals subcategory) compared to generations within a cluster.

Random Walks in Semantic Graphs Abbott et al. (2015) suggested that optimal foraging models may not be necessary to explain the apparent 'switching' behavior observed in humans. That is, there may not be a switch phase at all. Rather, they considered semantic memory as a graph where each vertex is a word and the length of the edge connecting two words represent the time cost of traveling from one to the other. Such a network would naturally have clusters, or neighborhoods of closely connected words corresponding to subcategories. A random walk over this network, weighted by the number of words connected to the current word and the time cost of each edge, would produce the same behavior as optimal foraging. This model proved to be consistent with the MVT, predicting longer times at purported 'cluster transitions' than within clusters. Notably, when Heineman et al. (2024) used random walks on a semantic graph derived from a language model, it did not produce human-like sequences.

Shortcomings of Current (Fixed) Coding Schemes

Hills et al. (2015) found evidence that clusters are *dynamically* created by associative connections from one word to the next, and thus cannot be described by a static coding scheme, i.e., a fixed set of subcategories as in Troyer et al. (1997). However, Kumar et al. (2025) found that cognitively plausible automated coding methods utilizing semantic word embeddings (Lundin et al., 2023; Sung et al., 2013) fare much worse than the standard Troyer coding at explaining a participant's self-reported groupings. At the same time, a static Troyer-like coding scheme only explained about 28% of the variance in participant-designated clusters. By contrast, when human raters judged consecutive word pairs, the resulting clusters explained 55% of the variance. This suggests that there are subtle differences between individuals and that semantic retrieval is context-dependent. These findings show that fixed coding schemes (and models that rely on them) cannot capture the dynamic, context-dependent, and associative grouping strategies that emerge during task performance.

LLMs as Cognitive Models of Category Fluency

Semantic Graphs of LLM Generations Prior studies have constructed semantic graphs from LLM-generated fluency sequences, where each word is a node and edges connect consecutively generated words, weighted by semantic similarity. These graphs are then compared to human-generated sequences to assess whether LLMs capture similar structural properties. Wang et al. (2025) found that compared to humans, LLMs generate sparse, long graphs without dense clusters (larger average shortest path lengths and smaller clustering coefficients). They also found that LLMs show less variability than humans. They tend to repeat the same high-frequency words across simulated LLM 'participants' prompted to take on different personas or demographics. This results in lower lexical diversity and fewer idiosyncratic responses (i.e., words produced by only one participant).

Attention and Logits of LLMs as Predictors A key question for our study is whether LLM internal representations (i.e., attention patterns) can capture the context-dependence of exemplar relationships that static coding schemes cannot. Zarrieß et al. (2025) examined how well linear regression models combining various combinations of internal LLM metrics, human Inter-item Response Times (IRTs), and exemplar embedding similarity predict subcategory switches. They used a variety of LLM metrics corresponding to different cognitive features during token generation (attention entropy, attention JS-divergence, surprisal, and logit entropy). For example, surprisal measures how unexpected each word is given the preceding context—the negative log probability the model assigns to that word. Logit entropy is the Shannon entropy of the output token distribution, mirroring uncertainty in humans. Attention entropy captures how distributed the model's attention is across prior tokens: high entropy means attention is diffuse (possibly indicating uncertainty), while low entropy means attention is concentrated on specific prior words. They found that an increase in attention-based metrics such as attention entropy is predictive of cluster switch boundaries, consistent with the MVT's prediction that switches are cognitively costly and mirroring elevated human response times. Importantly, when using regression models to predict whether a given exemplar position is a switch or cluster continuation, attention-based metrics provided significant predictive power beyond what static semantic embeddings offer. They find that attention-based metrics capture much of the same predictive information as human IRTs. This suggests that attention patterns encode something about a human's dynamic, context-dependent retrieval process that fixed word similarity measures and other LLM metrics do not capture.

Heineman et al. (2024) used LLMs to address the debate between optimal foraging (Hills et al., 2012) and random walk (Abbott et al., 2015) models. They also went beyond these cognitive science models to generate exemplar sequences and evaluate their correspondence to human-generated sequences. Following Troyer et al. (1997) and Hills et al. (2012), they used a cluster-and-switch framework where subcategory cues are represented as a hidden state. They found that LLM-generated sequences are not human-like with respect to their clusters and switches, and so used them for generation only within a cluster, while programmatically generating the subcategory cues. With this support, LLMs can generate more human-like sequences than network-based methods, with higher BLEU

scores and comparable cluster sizes. (Note that more recent LLMs can produce human-like sequences without the need for such support (Wang et al., 2025; Qiu et al., 2025).) They also found an increase in attention entropy at switch boundaries, consistent with human response times and the predictions of the MVT; see also Zarrieß et al. (2025).

Method

Dataset

We used the Hills et al. (2012) Animals category fluency dataset containing sequences from 141 undergraduates (5,079 total exemplars, 358 unique species, sequence length range: 14-57 exemplars). The Hills et al. (2012) taxonomy served as our baseline fixed coding scheme.

Model Setup

For all experiments, we use GPT-OSS-20B (Agarwal et al., 2025) with the following prompt to prevent "pauses" in animal listing for non-task generation behaviors, along with a special command to disable reasoning:

```
"You are performing a category fluency task.
You have 3 minutes  to name as many animals
as possible. List them in a continuous
comma-separated format. Go as fast as you can
and keep naming animals without pausing. Start
now:<|end|><|start|>assistant<|channel|>analysis
<|message|><|end|><|start|>assistant<|channel|>
final<|message|>" .
```

For direct evaluation of the Hills et al. (2012) data, when modeling each participant, we prefill all exemplars in their sequence in a comma-separated list after this prompt. For the generation task we run 11 separate generations, each limited to 57 animals (the maximum observed among the Hills et al. (2012) participants), and use these sampling parameters to avoid repetition (we increased repetition penalty until the model did not repeat animals; temperature was maximized while maintaining model coherence to ensure diverse outputs since lower temperatures resulted in similar sequences per generation): temperature=0.9, repetition_penalty=1.2

We extract attention weights averaged across all heads from each token to prior tokens for each layer. For each animal exemplar, we group the tokens in the animal and use the mean of these tokens' attention to prior animals to compute the attention from each animal to prior animals. We use these weights to create an *affinity matrix* containing the affinity between each animal as measured by the square root of the attention weight from the latter animal to the prior. The square root dampens extreme values while preserving relative ordering. We create affinity matrices both per-layer and averaged across all layers. We also extract logits for surprisal calculation.

Creating LLM-derived Coding Schemes

Research question (1) asks whether the attention weights of an LLM while processing human exemplar sequences contain

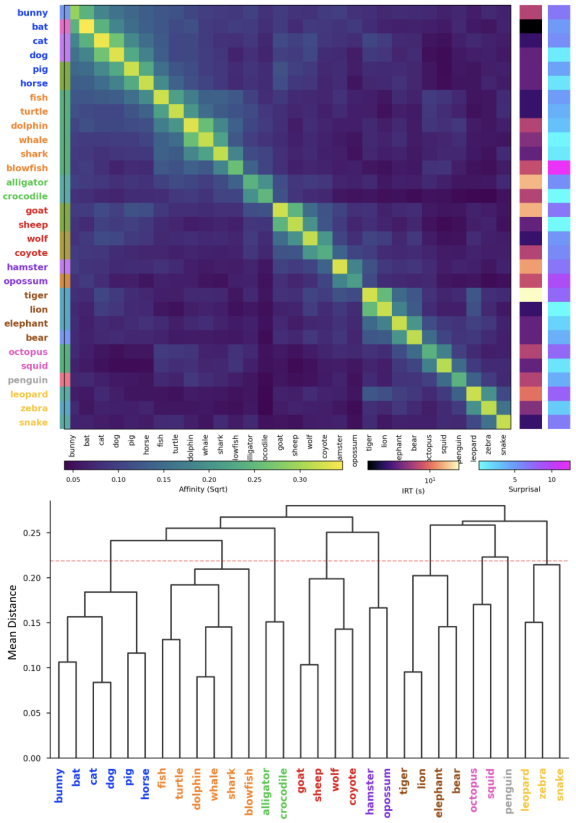

Figure 1: Affinity matrix and dendrogram for a participant. Animals in original order and colored with PB scheme. Left highlight-bar colored with Hills scheme.

information about the organization of semantic memory and can be the basis of individualized coding schemes. To evaluate this, we first derived the latent semantic structure of the LLM while processing human sequences. While processing each sequence, we first collected the affinity matrix, as defined above. We then ran a modified hierarchical clustering algorithm on the affinity matrix with 'average' linkage and with an additional constraint that when joining clusters, they must span adjacent exemplars in the participant's sequence. This constraint ensured that our derived clusters respect the temporal order in which participants generated animals.

We used two methods for determining the number of clusters k. In the 'Fixed-k' method, we set k equal to the number of clusters that the Hills coding scheme assigns to that participant's sequence. This allows direct comparison between attention-derived and Hills-derived boundaries while controlling the granularity of clustering. The second method requires pairwise distances, so we convert the affinity matrix to a distance matrix by subtracting each value from the matrix maximum and setting the diagonal to zero; thus, high-affinity pairs become low-distance pairs. The 'PB' method then chooses k to maximize the point-biserial correlation between these pairwise distances and cluster membership, effectively finding the k that best separates within-cluster pairs (low distance) from

between-cluster pairs (high distance). This method allows for k selection that best matches the inherent structure of the affinity matrix.

Transformer models like GPT-OSS-20B process information through multiple sequential layers, and the attention patterns at each layer can differ—earlier layers may capture surface-level patterns while later layers capture more abstract relationships. To investigate which layers provide the most cognitively-relevant signal, recall that we construct affinity matrices in two ways: per-layer, where we use attention weights from a single layer, and layer-averaged, where we average attention weights across all layers. To the per-layer affinity matrices, we applied both clustering methods (Fixed-k and PB) to each participant's sequence to evaluate how different layers and methods perform at reconstructing subcategory structure.

Measuring Scheme Alignment to the MVT

Research question (2) asks whether when human-produced sequences are coded using an LLM-derived scheme, the data are still consistent with the predictions of the MVT. In particular, is there an increase in IRT between the last exemplar of the prior subcategory and the first exemplar of the new subcategory? Additionally, is there an increase in attention entropy and surprisal at Hills et al. (2012) switch boundaries, indicating that the model also experiences greater uncertainty?

To evaluate this, we examine human and LLM proxies of cognitive effort on human data. For each participant, we first log-transform their IRTs and calculate a z-score for each time. Correspondingly for the LLM, we calculated the surprisal measure defined above and the normalized Attention Head Entropy (AHE). AHE measures how distributed the model's attention is when processing each exemplar and is computed using the Shannon entropy of attention weights across all prior tokens (averaged across each model layer) and normalized by $\log(t)$ (where t is the position in the sequence) to account for the increasing number of possible attention targets. Each metric (IRT z-scores, surprisal, AHE) is computed for each participant for the 5 exemplar positions $-2, -1, Switch, 1, 2$ relative to switch boundaries defined by the Hills et al. (2012) coding ('Hills') and the two LLM-derived codings ('Fixed-k' and 'PB'). For each participant, for each metric, the windows of 5 exemplar positions were averaged together to create an average participant window. These were then averaged across all participants.

We used nonparametric tests to test whether IRTs, AHE, and surprisal are higher at switch boundaries than at surrounding positions as predicted by the MVT. After confirming significant differences among the five positions with a Friedman test, we use a Wilcoxon signed-rank test to compare each surrounding position (-2, -1, +1, +2) against the switch position with the Bonferroni correction applied.

We identified which layers of the model provide the most cognitively-relevant signal. For each participant's output, for the affinity matrix for each layer, we repeated the following analysis: At each switch boundary according to the Hills or LLM-derived coding schemes, we computed the difference between the z-scores at the switch and the mean of the surrounding four exemplars, representing average increase in the participant's IRT z-score at the switch boundary. We averaged these differences for each switch for a participant, and then averaged the participant means to obtain an overall *IRT difference metric*. We compared each scheme by model layer.

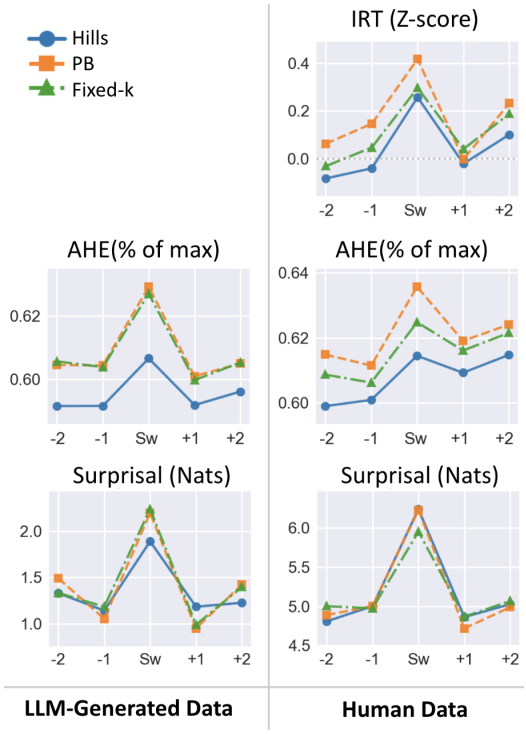

Figure 2: Behavior of schemes at subcategory switch boundaries for Hills dataset and LLM-generated dataset, using a layer-averaged affinity matrix for LLM-derived schemes.

Results

Performance of LLM-Derived Coding Schemes

We evaluate the new schemes by examining similarity between LLM-derived and Hills codings using Adjusted Mutual Information (AMI), subcategory switch distributions, and cases where codings differ.

Across participants, we find the AMI between the Hills and PB schemes (μ=0.59, σ=0.11) and the AMI between the Hills and Fixed-k schemes (μ=0.63, σ=0.11) are both moderately high, indicating some similarity between the Hills and LLM-derived clusterings. This suggests that our new schemes are semantically meaningful yet capture some different information than the Hills scheme.

Figure 1 shows the affinity matrix and dendrogram for one participant, illustrating that exemplars within the same PB cluster generally have higher attentional weights connecting them. The dendrogram is cut at a distance threshold using the PB method to produce $k = 9$ clusters, as compared to the $k = 17$

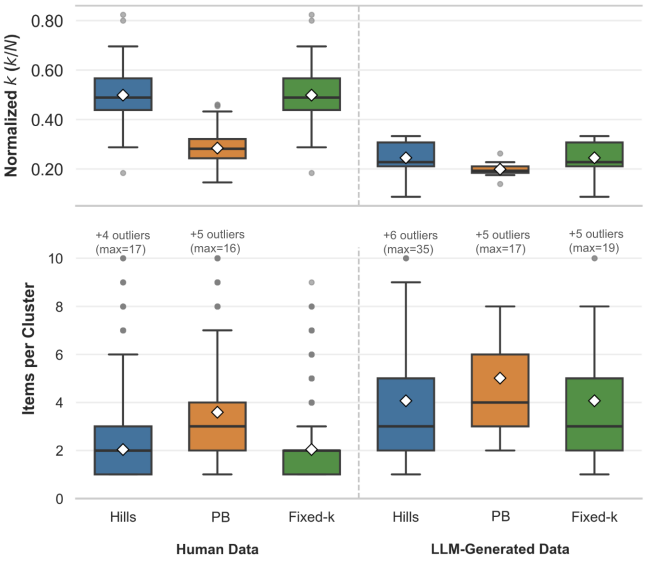

Figure 3: Comparison of cluster sizes and number of clusters (normalized by sequence length) for all schemes on both human and LLM-generated data.

clusters produced by the Hills clustering. In fact, the left half of Figure 3 shows that using the Hills scheme on human data creates clusters with a median of just 2 exemplars and produces almost double the median number of subcategory switches per participant compared to the PB scheme. This shows how LLM-derived k-selection can reduce spurious switches.

Furthermore, the affinity matrix shows how similar exemplars like "hamster" and "opossum" are grouped together by the model-derived coding scheme but are separated by the strict Hills scheme. Another example is "bat, bird, bee" which the Hills scheme separates into 3 clusters and the PB scheme groups together.

Alignment between LLM-Derived Coding Schemes and Predictions of the MVT

We find that LLM-derived coding schemes show good alignment to the MVT. The right half of Figure 2 shows MVT curves for the human data. We find significantly higher IRT at switch boundaries ($p < 0.05$) for both Hills and LLM-derived schemes. When prefilling the human sequences into GPT-OSS-20B, we see significant increases in AHE and surprisal at switch boundaries across all schemes ($p < 0.05$), with the exception of AHE on Hills scheme boundaries where there was no significant difference between the switch position and +2 position. Between the LLM metrics, surprisal shows the greater rank-biserial effect ($r_{rb} > 0.8$ for all coding schemes and positions). Notably, surprisal – measured before token sampling – indicates switches are less predictable, while AHE – measured after sampling – reflects post-hoc attention diffuseness from the new exemplar.

Our per-layer analysis in Figure 4 shows that the Hills scheme generally generates an equivalent or larger IRT spike at

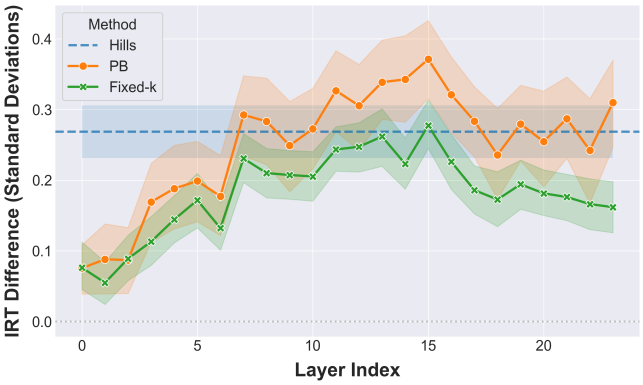

Figure 4: Effect of using different coding schemes on IRT spike at cluster switch by model layer. 95% CI highlighted.

switch boundaries compared to the Fixed-k scheme regardless of model layer, but the PB scheme shows a significantly higher average IRT spike (up to 30% greater) at switch boundaries compared to the Hills scheme in the middle layers (11-16). This indicates the PB scheme's ability to exclude spurious switches that the Hills scheme would include enables it to capture only the larger magnitude IRT differences consistent with the MVT. By contrast, the Fixed-k coding scheme must match the large number of clusters k used by the Hills scheme, leading to a smaller magnitude spike. This further validates the bottom-up derivation of coding schemes using LLMs rather than relying on research intuitions (Troyer et al., 1997; Hills et al., 2012).

Cognitive Plausibility of LLM Generations

Research question (3) asks if LLMs produce human-like exemplar sequences with similar clustering, switching, and MVT congruence.

First, consider the predictions of the MVT. Treating each LLM generation as a participant, we test whether surprisal and AHE increase at switch boundaries under the three clustering schemes. As shown on the left side of Figure 2, there are significant increases in AHE and surprisal at switch boundaries across all schemes, with generally similar magnitudes as in the human data. Despite this alignment with the MVT, key differences remain. The AHE spike is more pronounced compared to human data for all schemes, lacking the trend of increasing AHE present in the human data. This may indicate LLM-generated sequences fit more neatly into the Hills et al. (2012) taxonomy than human-generated sequences. We would expect such a clean AHE spike on generated data using the Hills coding if attention patterns and subcategory structure align well with the Hills coding. On the generated data, the average AMI between the Hills clusterings and the PB clusterings was high ($\mu = 0.69$, $\sigma = 0.09$), supporting the idea that the attention map encapsulates a slightly more Hills-like coding in generated data than in human data, where the average AMI was lower ($\mu = 0.59$, $\sigma = 0.11$). However, on generated sequences, PB coding still produces somewhat greater cluster

sizes (Mdn 4 vs Mdn 3 on bottom right of Figure 3) than the Hills coding, indicating that the Hills coding may still produce some spurious switches and that generated data does not fit perfectly into a Hills scheme. Figure 5 shows an example of a generated sequence where PB creates $k = 12$ clusters while Hills coding produces $k = 19$ clusters due to possible spurious switches. For example, it produces single-item clusters for snail and slug, which PB groups together.

We next evaluate if LLM generations show similar patterns of clustering and switching as humans. The lower chart on Figure 3 shows that compared to the human sequences, the clusters in generated sequences have an increased median exemplar size (by 1) regardless of coding scheme used. This is consistent with Qiu et al. (2025)'s finding that LLM generations have larger clusters than human sequences.

Discussion

Our findings suggest that LLM attention mechanisms may offer a dynamic, context-sensitive method for decoding semantic memory structure. Attention-derived subcategories perform at least as well as fixed, Hills subcategories at revealing the increasing human cognitive effort at switch boundaries predicted by the MVT. Importantly, subcategories derived from middle layer attention patterns, which represent more abstract reasoning (Goldstein et al., 2025), surpass the Hills scheme. This suggests the LLM might be identifying cognitive boundaries that static norms miss. Previous coding schemes use rigid taxonomies (e.g., Fish vs. Birds), a poor fit for human retrieval, which is more fluid (Kumar et al., 2025). For example, humans sometimes utilize idiosyncratic groupings (e.g., "cute animals") which attention-based methods might capture. To validate this, future work should compare LLM clusters with human-labeled clusters. If validated, this potential ability of LLMs to find latent clusters without supervision would support the hypothesis that semantic switching is driven in part by context-dependent attentional shifts that cannot be captured by fixed categorical schemes (Hills et al., 2015; Kumar et al., 2025; Collins & Loftus, 1975).

How might LLMs capture these idiosyncrasies? Analysis of generated sequences reveals that the model does not strictly adhere to hierarchical clusters but instead produces "associative chains" reminiscent of "local exploitation" in optimal foraging models. For instance, in this LLM-generated sequence cluster – rat, bat, fox, and wolf, deer – sequential pairs show commonalities of different types – rhyme, the flying fox species of bat, predators, and predator-prey, respectively – that signal a single cluster despite no overall commonality. This mirrors the dynamics observed in human productions, where retrieval is sometimes guided by local cues rather than just a global taxonomy, and illustrates the potential of LLMs for deriving individualized coding schemes.

By generating such fluid transitions and grouping such exemplars in the same cluster (via attention), LLMs can potentially help bridge between clustering accounts and random walk models. The attention mechanism may be acting as

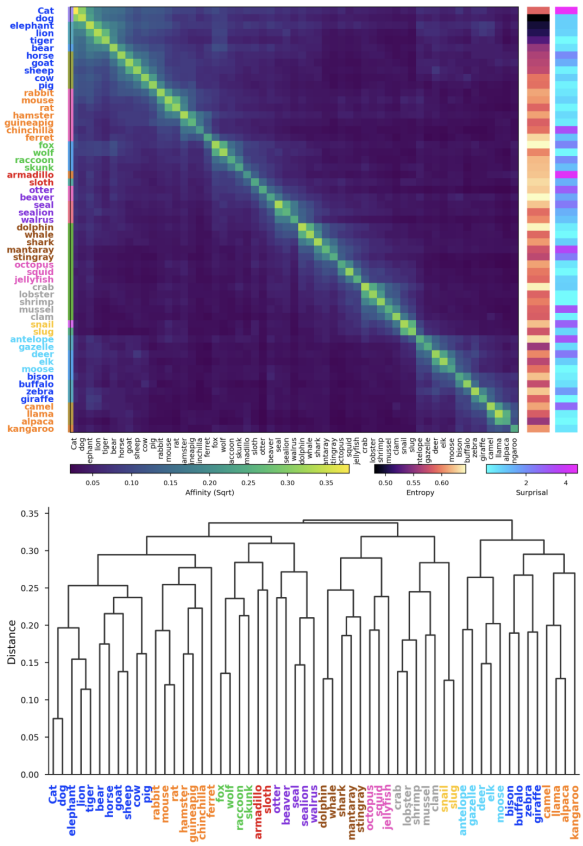

Figure 5: Affinity matrix and dendrogram for an LLM-generated sequence, formatted like Figure 1.

a *dynamic similarity metric* that evolves with the sequence, mirroring the navigation of the mental lexicon that Kumar et al. (2025) identifies as a hallmark of human performance. Overall, LLM-derived schemes can potentially capture the contextual groupings that fixed schemes miss.

The current study is limited to simple interpretability methods on only one LLM. Future work should explore advanced probing techniques and different models and category domains, especially with non-English LLMs and datasets, along with comparisons to self-reported clusters of humans.

Conclusion

We examined whether LLMs can serve as cognitive models of category fluency by analyzing attention patterns during processing of human and self-generated sequences. Our findings provide evidence that LLM attention mechanisms capture some aspects of human semantic organization. Hierarchical clustering of attention weights at middle layers produces subcategories corresponding to MVT-predicted IRT elevations at subcategory switch boundaries. AHE and surprisal generally increase at switch boundaries across clustering schemes. Ultimately, this work suggests a shift in the analysis of semantic fluency data from imposing static external structures onto human data to discovering dynamic internal retrieval strategies.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging.

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., ... others (2025). gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology*, *9*(2), 129–136.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Goldstein, A., Ham, E., Schain, M., Nastase, S. A., Aubrey, B., Zada, Z., ... Hasson, U. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, *16*(1), 10529. doi: 10.1038/s41467-025-65518-0

Heineman, D., Koenen, R., & Varma, S. (2024). Towards a path dependent account of category fluency. *arXiv preprint arXiv:2405.06714*.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, *7*(3), 513-534. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12151 doi: https://doi.org/10.1111/tops.12151

Kumar, A. A., Lundin, N. B., & Jones, M. N. (2025). What's in my cluster? evaluating automated clustering methods to understand idiosyncratic search behavior in verbal fluency. *Journal of Memory and Language*, *141*, 104606. Retrieved from https://www.sciencedirect.com/science/article/pii/S0749596X24001098 doi: https://doi.org/10.1016/j.jml.2024.104606

Lundin, N. B., Brown, J. W., Johns, B. T., Jones, M. N., Purcell, J. R., Hetrick, W. P., ... Todd, P. M. (2023). Neural evidence of switch processes during semantic and phonetic foraging in human memory. *Proceedings of the National Academy of Sciences*, *120*(42), e2312462120. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2312462120 doi: 10.1073/pnas.2312462120

Qiu, M., Brisebois, Z., & Sun, S. (2025). Can llms simulate human behavioral variability? a case study in the phonemic fluency task. *arXiv preprint arXiv:2505.16164*.

Sung, K., Gordon, B., Vannorsdall, T. D., Ledoux, K., & Schretlen, D. J. (2013). Impaired retrieval of semantic information in bipolar disorder: a clustering analysis of category-fluency productions. *Journal of abnormal psychology*, *122*(3), 624.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *neuropsychology*, *11*(1), 138.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Wang, Y., Deng, Y., Wang, G., Li, T., Xiao, H., & Zhang, Y. (2025). The fluency-based semantic network of llms differs from humans. *Computers in Human Behavior: Artificial Humans*, *3*, 100103.

Zarrieß, S., Junker, S., Sieker, J., & Alaçam, Ö. (2025). Components of creativity: Language model-based predictors for clustering and switching in verbal fluency. In *Proceedings of the 29th conference on computational natural language learning* (pp. 216–232).