# Comparing Semantic Navigation in Humans and Large Language Models using Natural Language Processing

**Gabriel Paris-Colombo (gabrielparis461@gmail.com)[1], Rodrigo M. Cabral-Carvalho[1] & Felipe D. Toro-Hernández[1]**

[1]Center for Mathematics, Computing and Cognition, Federal University of ABC

**Abstract**

Semantic memory retrieval can be conceptualized as navigation through conceptual space. We compared semantic search dynamics between humans and three large language models (GPT-4o, Gemini-2.5-Pro, Claude-Sonnet-4.5) using verbal fluency data. By applying trajectory-based NLP metrics to the items generated by 82 human participants and LLM output across eight temperature settings, we quantified three complementary dimensions: entropy (step size predictability), distance to next (successive semantic steps), and distance to centroid (global dispersion). Humans exhibited higher entropy, larger semantic steps and broader dispersion than all LLMs, indicating more variable and exploratory search. Temperature tuning produced only partial alignments, as individual metrics matched between humans and LLMs at specific settings, but no configuration reproduced the complete human profile (in all dimensions). These findings suggest that human semantic search implements a distinctive balance between local exploitation and global exploration that current model architectures fail to reproduce.

**Keywords:** semantic memory; verbal fluency; large language models; semantic navigation; natural language processing

## Introduction

Semantic memory is the long-term storage of general world knowledge, facts, concepts, and meanings (Tulving, 1972; Ralph et al., 2017; Binder & Desai, 2011). Its retrieval can be conceptualized as a navigational process with complex internal dynamics. To investigate the mechanisms underlying this search, researchers have commonly employed the Verbal Fluency Task (VFT), where participants generate as many items as possible from a category (e.g., “animals”) or beginning with a letter within a fixed interval, typically 60 seconds (Bousfield & Sedgewick, 1944). Fluency performance has often been summarized by the total number of words produced (Lezak et al., 2012; Shao et al., 2014; Strauss et al., 2006). However, semantic navigation accounts additionally leverage the temporal organization and semantic structure of responses to characterize how retrieval proceeds and to motivate theoretical models of semantic search (Troyer et al., 1997; Hills et al., 2012; Abbott et al., 2015).

A highly influential framework for understanding semantic fluency is the dual-component model (Troyer et al., 1997), which posits that performance depends on two dissociable processes: clustering (producing words within semantic subcategories, e.g., "pets") and switching (shifting between subcategories). Hills et al. (2012) extended this view by framing semantic retrieval as internal foraging, where effective search balances exploitation (local clustering) and exploration (switching to new semantic regions). Granular search models aim to specify the moment-to-moment mechanisms of retrieval (how agents transition through semantic space and when they make local versus long-range jumps). These include Random Walk models, which capture probabilistic transitions over semantic networks; Lévy Walk variants, which combine short local steps with occasional long jumps; and Area Restricted Search, which proposes that switching emerges from adaptive estimation of local "patch value" based on recent retrieval success (Morales et al., 2025; Rhodes & Turvey, 2007). While these models have been developed primarily to explain human semantic search, recent advances in artificial intelligence offer new opportunities to test whether similar dynamics emerge in systems that navigate semantic space through fundamentally different computational mechanisms.

Large Language Models (LLMs) are a class of artificial intelligence systems built on Transformer architecture (Vaswani et al., 2017), typically pretrained to predict the next token (word or subword unit) in sequences across large text corpora. This training objective leads them to learn statistical regularities that support fluent language generation. Internally, LLMs encode tokens as high-dimensional vectors and transform these representations across layers through self-attention mechanisms (Vaswani et al., 2017), which enable the extraction of multi-scale dependencies beyond mere lexical disambiguation. These architectures encode complex semantic hierarchies that encapsulate the structural and conceptual essence of entire phrases (Jawahar et al., 2019; Devlin et al., 2019), such that tokens with related meanings occupy nearby regions in the model's representational space (e.g., "dog" closer to "cat" than "airplane"). This property enables quantitative analyses of semantic structure and semantic movement during generation (Kalyan et al., 2021). Beyond their practical success across Natural Language Processing (NLP) tasks, LLMs have become increasingly relevant to cognitive science because they generate coherent, context-sensitive language that sometimes mirrors qualitative patterns of human linguistic behavior (Dasgupta et al., 2022; Manning et al., 2020). This has motivated proposals to treat LLMs as experimental comparators "silicon participants" whose outputs can be analyzed with the same behavioral metrics used for humans (Sarstedt et al., 2024). However, interpreting LLM behavior as cognitive



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

evidence raises important methodological concerns, as these models often fail to replicate human strategic behavior in simple experiments and struggle to generalize across semantic variations that humans process intuitively (Gao et al., 2025; Schröder et al., 2025). Furthermore, LLMs' responses exhibit significant instability based on minor phrasing changes and rely on statistical token regularities rather than authentic psychological simulation (Gao et al., 2025; Schröder et al., 2025).

Recent work has examined whether LLMs can serve as useful behavioral comparators in cognitive tasks and to what extent their internal representations align with human semantic representations. For example, Hansen and Hebart (2022) showed that GPT-3 can generate semantic feature norms for a large set of objects that resemble human feature norms in the distribution of feature types (e.g., taxonomic, functional, encyclopedic) and that these features can predict human judgments of similarity, relatedness, and category membership. Furthermore, LLMs show sensitivity to human category structures, performing well in basic-level category induction, though they often struggle with more abstract superordinate levels (Pedrotti et al., 2025). Recent neuroscientific investigations using functional neuroimaging have provided additional evidence for this alignment: Ren et al. (2024) demonstrated that pre-training data size, model scaling, and alignment training positively correlate with LLM-brain similarity (i.e., the similarity between LLM representations and brain cognitive language processing signals), with instruction-tuned models showing significantly stronger alignment with fMRI signals during language processing. Lei et al. (2025) further showed that intermediate layers of LLMs align more strongly with brain activity than final layers, particularly in language-selective regions, and that this alignment respects hemispheric lateralization patterns observed in human language processing. Critically, when semantic search dynamics are compared using networks derived from fluency tasks, systematic structural differences emerge. Relative to humans, LLM-generated networks tend to show weaker local clustering and poorer global integration (Qiu et al., 2025), with semantic associates being more dispersed and organized into more isolated subgroups in semantic fluency networks (Wang et al., 2025). These patterns are consistent with less efficient navigation through semantic space, motivating trajectory-based step-by-step comparisons between humans and LLMs.

Despite recent advances in mapping the structural properties of LLM-derived semantic networks, a key gap remains in quantifying how these systems progress through semantic content on a step-by-step basis. Prior work has reported systematic differences between human and LLM semantic networks, such as lower clustering and higher modularity for the language models (Wang et al., 2025). Furthermore, while previous studies have successfully employed sequential distance metrics between static word embeddings to evaluate cognitive states in verbal fluency tasks (Linz et al., 2017; van Noort et al., 2025), these approaches assess semantic transitions by treating each generated word independently of its broader retrieval history. The present work extends the computational framework introduced by Toro-Hernández et al. (2026) by using their physics-inspired NLP metrics to assess semantic representations in LLMs. Instead of relying on sequential distances between static word representations, this approach utilizes sequential distances between cumulative contextual embeddings. We specifically chose this cumulative approach because it allows the semantic state to unfold over time, better approximating the accumulating working memory demands and contextual constraints inherent to human memory search (Baddeley, 2000). Following Toro-Hernández, Vieira & Cabral-Carvalho (2026), we quantify (i) distance to next, capturing semantic "jumps" between successive properties; (ii) semantic entropy, summarizing the predictability of step sizes across the sequence; and (iii) distance to centroid, indexing the degree of clusterization. Together, these metrics provide a fine-grained characterization of semantic navigation that complements former approaches.

With these analyses, we aim to characterize systematic differences in semantic navigation between humans and LLMs. Guided by prior work, we first expect LLM outputs to vary across temperature settings. Because temperature modulates sampling stochasticity during generation (Li et al., 2025), we consider that higher temperatures will yield more variable trajectories, reflected in larger distance to next values (broader "semantic jumps" between consecutive items), higher semantic entropy (greater diversity in exploration), and increased distance to centroid (production of more peripheral category exemplars), whereas lower temperatures will produce more predictable, potentially more rigid trajectories with reduced semantic variation. We further expect the closest correspondence to human navigation to emerge at intermediate temperatures, where generation is neither overly deterministic nor excessively noisy. Across model families, we anticipate that LLMs will generally show higher distance to next and entropy values than humans, consistent with their more dispersed semantic organization (Qiu et al., 2025; Wang et al., 2025), though meaningful variation may emerge due to differences in training and alignment procedures (Ren et al., 2024; Lei et al., 2025). Finally, to benchmark the generality of these effects across model families, we repeat the fluency-generation procedure with three state-of-the-art LLMs: gpt-4o, gemini-2.5-pro, and claude-sonnet-4.5, and compare the resulting trajectory profiles. By comparing temperature settings and model configurations, we test which conditions yield trajectory metrics most similar to humans, and we use these matches to constrain candidate mechanisms underlying human-like semantic navigation. This approach provides a framework for generating testable hypotheses regarding the constraints and search dynamics of human semantic memory.

# Methods

## Participants

To compare LLM-generated fluency with human performance, we used behavioral data from the SNAFU database (Zemla et al., 2020). In this dataset, 82 participants completed a classic semantic verbal fluency task across multiple categories. For each category, participants were instructed to produce as many items as possible within a 3-minute interval, avoiding repetitions and avoiding

responses that shared the same suffix. In the present analyses, we focused on the animals category.

## LLM-generated fluency

We evaluated three LLMs (gemini-2.5-pro, gpt-4o, and claude-sonnet-4.5) on a semantic fluency task. Following Wang et al. (2025), we used a role-playing induction to increase output diversity and reduce repetitive list-like completions (Shanahan et al., 2023; Zhao et al., 2024). We used the 30 different occupations described by Wang et al. (2025). For each trial, the model was first prompted to describe the characteristics of a target occupation ("Do you know a/an [occupation]? Tell me the characteristics of a/an [occupation]."") and then instructed to adopt the role ("Now you need to pretend that you are a person with the above characteristics completely. All the following answers must be in accordance with the above characteristics."). After this induction, the model was asked to generate an animal-fluency list (one animal per line), using standard verbal-fluency constraints (Henry & Crawford, 2004).

To reduce incomparability introduced by human timing constraints versus model generation time (Ding & Wang, 2025; Jain et al., 2023), we did not impose a time limit on the models. Instead, for each LLM trial, we defined a target list length by sampling from the empirical distribution of the number of items produced by human participants in the SNAFU animals category (Zemla et al., 2020). This matched-length stopping rule yields model-generated sequences that are comparable to human sequences in overall output size while avoiding assumptions about the equivalence of response latencies across humans and artificial agents. Model outputs were formatted as one item per line into an Excel sheet to facilitate parsing and downstream analyses.

To examine how stochasticity during generation affects semantic navigation, we varied temperature across trials (Herrera-Poyatos et al., 2025). Temperature modulates sampling randomness, with lower values typically producing more deterministic outputs and higher values producing more diverse outputs (Li et al., 2025; Renze, 2024). Accordingly, model temperature was varied to capture a broader range of responses (Qiu et al., 2025). Each LLM generated responses at each of eight temperature settings: 0.0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9, and 1.0, using each of the occupations.

## Natural language processing analyses

To analyze semantic representations in both human and LLM-generated responses, we used NLP methods to derive metrics that capture the step-by-step granularity of semantic navigation (Nour et al., 2023; Toro-Hernández et al., 2024). We use transformer-based text embeddings, which map text to high-dimensional vector representations that capture contextual meaning (Vaswani et al., 2017). Following recent practices in computational semantics (Hansen & Hebart, 2022; Wang et al., 2025; García et al., 2025; Sanz et al., 2022), we used OpenAI's text-embedding-3-large to encode each produced item into a shared latent semantic space. Treating responses as points in this space enables rigorous quantification of relationships between successive outputs; in turn, distance-based measures can be used to characterize navigational patterns and the semantic structure of the trajectories generated by humans and models (Toro-Hernández et al., 2026). To ensure robustness and address potential concerns about reliance on proprietary embedding models, all metrics were validated using multiple embedding systems, including both commercial and open-source alternatives.

For each verbal fluency trial, we utilized these high-dimensional representations to calculate three complementary metrics (Toro-Hernández et al., 2026). These measures quantify the proximity between successive responses and the overall density of the evoked features, providing a multidimensional view of memory search performance. Importantly, verbal fluency is sequential and depends on executive processes such as maintaining task goals, monitoring prior responses to avoid repetition, and inhibiting intrusions, functions linked to working-memory maintenance and cognitive control (Baddeley, 2000; Troyer et al., 1997). To approximate this accumulating contextual demand, trajectory-based metrics were implemented with a cumulative embedding representation, by which, at step t, the semantic state summarizes the items generated up to that point by combining the current item embedding with embeddings of all previously generated items (Toro-Hernández et al., 2026). This cumulative formulation allows semantic navigation dynamics to be modeled as they unfold over time, rather than treating each response as independent. In contrast, the distance to centroid metric was computed from non-cumulative item embeddings, because it is intended to index dispersion (semantic density) relative to an individual's overall semantic context rather than the evolving cumulative state (Toro-Hernández et al., 2026).

## Natural language processing metrics

To assess step-by-step semantic navigation dynamics, we extracted three NLP metrics from Toro-Hernández et al. (2026) (see that paper for full mathematical details).

**Distance to next.** We mapped each response stream to a semantic trajectory $X = (x_1, \ldots, x_N)$ using cumulative embeddings, where $x_t$ encodes the concatenated prefix of items 1:t (e.g., if the first two items are "cat" then "dog," $x_2$ embeds "cat dog"). We then computed distance to next as the cosine distance between successive unit-normalized states, $x_t$ and $x_{t+1}$, yielding an N - 1 series of step sizes ("semantic jumps"). Larger values indicate larger moment-to-moment shifts in semantic state (e.g., "dog" → "shark" > "cat" → "dog"). For a length-invariant summary, we used the mean step size across all steps in the trajectory.

**Entropy.** To quantify the predictability of step-size dynamics, we computed a length-normalized Shannon entropy over the distance to next series (Shannon, 1948). Within each trajectory, step sizes were binarized as "high" vs "low" relative to the within-trajectory median, producing a binary sequence from which Shannon entropy was computed and normalized for the number of valid steps. Entropy was only computed for sequences with at least three steps. Higher entropy indicates a less predictable, more variable pattern of step sizes across the trajectory.

**Distance to centroid.** To index overall dispersion (clustering vs spread) relative to an individual's central semantic context, we computed a centroid-based measure from non-cumulative item embeddings. Repeated items

were collapsed so redundancy did not overweight specific responses; for each unique item, we retained the embedding from its first occurrence and defined the centroid as the average of all unique-item embeddings. Each item was then assigned the cosine distance between its embedding and this centroid. Higher distance to centroid values indicate a more dispersed semantic set, whereas lower values indicate a more focused (more clustered) set of responses.

### Data analysis

We compared human participants with LLM-generated responses across eight temperature conditions (0.0, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90, and 1.0) using generalized linear mixed-effects models (GLMMs). For each NLP metric, we fitted a model with group as a fixed effect and participants/instance ID as a random effect to account for repeated measures and individual variability. Models were fitted according to the most appropriate distribution, assessed using DHARMa residual diagnostic in R. Analyses were performed with the glmmTMB package (Brooks et al., 2017) in R (v.4.3.1).

## Results

Our analyses proceeded in two stages. First, we established a baseline by comparing general performance metrics between each group (gpt-4o, gemini-2.5-pro, claude-sonnet-4-5, human). Second, we conducted a series of contrasts between the human group and LLMs groups performance at each specific temperature level. Full statistical reports for all comparisons, including estimate values and exact p-values, are provided in Supplementary Materials (available at https://osf.io/gkp57)

### Category comparison

**Entropy** Human participants exhibited substantially higher entropy than all three LLMs. Among LLMs, Claude-sonnet-4.5 showed lower entropy than both Gemini-2.5-pro and GPT-4o, whereas Gemini and GPT-4o did not differ reliably from one another (see Table 1 for comparisons).

**Distance to next** Humans produced larger semantic steps than each LLM. Among LLMs, Claude-sonnet-4.5 showed the smallest step sizes, differing reliably from Gemini-2.5-pro but not from GPT-4o.

**Distance to centroid** Human responses were the most semantically dispersed. All three LLMs showed more centroid-focused responses, with GPT-4o exhibiting the strongest clustering, followed by Gemini-2.5-pro. All pairwise contrasts were significant.

**Summary of category comparisons** Human semantic trajectories exhibit greater variability and less predictable step-size dynamics, whereas all LLMs show more regular, constrained entropy profiles consistent with a more rigid navigation regime. Similarly, human navigation involves larger successive semantic shifts between items, whereas LLMs exhibit more "sticky" local trajectories with tighter moment-to-moment transitions through semantic space, particularly for Claude. At a broader scale, LLM responses maintain a more conservative global search profile, exploring a narrower semantic breadth compared to the wider dispersion characteristic of human semantic retrieval.

Table 1: Tukey results for category comparisons.

| Metric | Contrast | Δ | p |
|---|---|---|---|
| Entropy | Claude vs. Human | -0.211 | <0.001 |
| | Gemini vs. Human | -0.187 | <0.001 |
| | GPT vs. Human | -0.192 | <0.001 |
| | Claude vs. Gemini | -0.024 | <0.001 |
| | Claude vs. GPT | -0.019 | 0.003 |
| | Gemini vs. GPT | 0.005 | 0.732 |
| Distance N | Claude vs. Human | -0.106 | <0.001 |
| | Gemini vs. Human | -0.068 | 0.003 |
| | GPT vs. Human | -0.085 | <0.001 |
| | Claude vs. Gemini | -0.038 | <0.001 |
| | Claude vs. GPT | -0.021 | 0.096 |
| | Gemini vs. GPT | 0.017 | 0.257 |
| Distance C | Claude vs. Human | -0.025 | <0.001 |
| | Gemini vs. Human | -0.016 | <0.001 |
| | GPT vs. Human | -0.058 | <0.001 |
| | Claude vs. Gemini | -0.010 | <0.001 |
| | Claude vs. GPT | 0.033 | <0.001 |
| | Gemini vs. GPT | 0.043 | <0.001 |

### Human vs. LLMs temperatures

**Entropy** Results showed that alignment with humans was model-specific (Figure 1 - A): Claude-sonnet-4.5 aligned most closely at high-intermediate temperatures (e.g. T = 0.75 and T = 0.90), whereas Gemini-2.5-pro and GPT-4o aligned most clearly at T = 0, with additional convergence at T = 1.0 and T = 0.60.

**Distance to next** Here, results were also model-specific (Figure 1 - B). Claude-sonnet-4.5 converged toward the human baseline at higher intermediate temperatures (T = 0.75 and 0.90), showing reliably smaller step sizes at lower temperatures. Gemini-2.5-pro showed the most human-like settings at the endpoints and some lower temperatures (T = 0, 0.15, 0.60, 0.75, 1.0), with reduced "semantic jumping" at T = 0.30, 0.45, and 0.90. GPT-4o was indistinguishable from humans at T = 0, 0.45, 0.60, and 1.0, but showed more locally constrained transitions at T = 0.15, 0.30, 0.75, and especially 0.90.

**Distance to centroid** Global dispersion was the hardest dimension to match (Figure 1 - C). Only a narrow setting for Claude (T = 0.60) and a mid-range band for Gemini (T ≈ 0.30 to 0.90, excluding endpoints) were statistically indistinguishable from humans. GPT-4o remained persistently non-human-like across all temperatures, showing consistently reduced distance to centroid values.

**Summary of human vs. temperature comparisons** Overall, the temperature comparisons show that "human-like" behavior does not emerge as a simple monotonic function of added sampling noise; instead, it depends on how each architecture converts temperature into semantic variation. This is most apparent for local trajectory dynamics such as entropy and distance to next, where models exhibit distinct alignment regimes, with Claude approaching humans at higher intermediate temperatures, while Gemini and GPT-4o show more irregular convergence patterns with pockets of locally constrained transitions. By contrast, global dispersion (distance to centroid) is harder to

recover through temperature tuning, suggesting that global semantic breadth is limited by deeper representational constraints that temperature alone may not overcome.

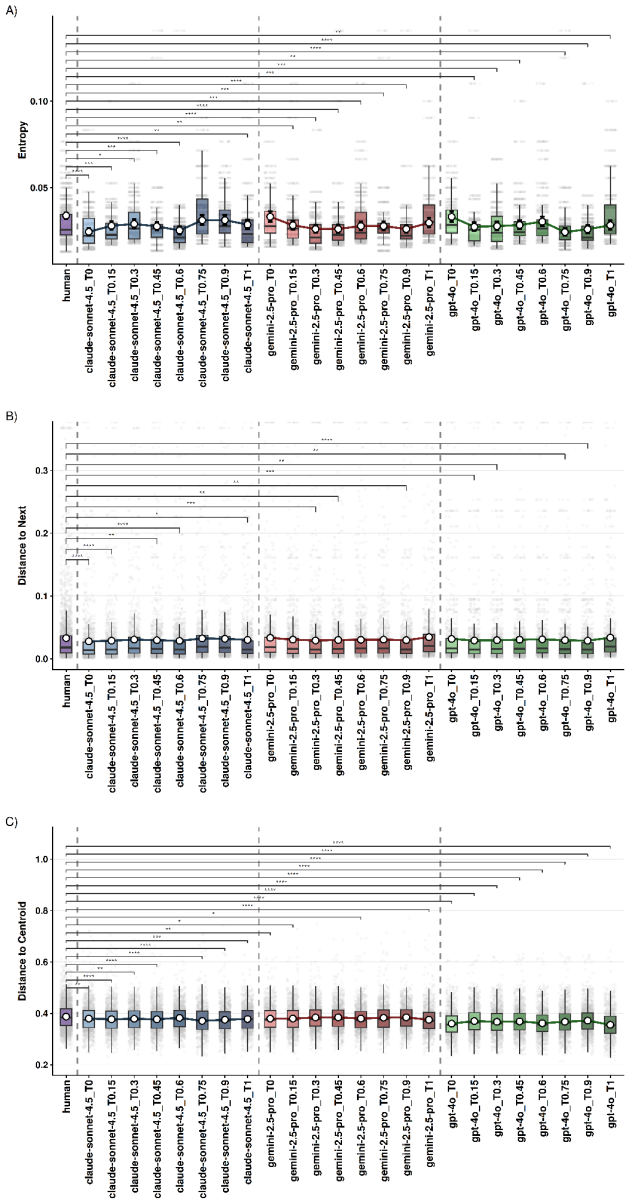


**Figure 1. Human baseline vs. all LLM–temperatures.** Entropy (A), Distance N (B), and Distance C (C). The leftmost group shows humans, followed by LLM outputs grouped by model; Claude Sonnet 4.5; Gemini 2.5 Pro; and GPT-4o. Vertical dashed lines separate model families. Gray points show individual observations; boxplots show medians and interquartile ranges. Circles indicate estimated marginal means (± 95% CI); within each model family, line color intensity increases with temperature (lighter = T = 0.0, darker = T = 1.0). Brackets denote Tukey-adjusted contrasts against the human baseline (* $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$). All plots in high quality are provided in Supplementary Materials.

## Discussion

Our findings show reliable differences in how humans and LLMs navigate semantic space in verbal fluency. Contrary to expectations derived from prior reports of broader dispersion in LLM semantic neighborhoods, our trajectory-based metrics reveal the opposite pattern. Humans exhibited higher entropy (less predictable step-size dynamics), larger successive semantic shifts, and broader dispersion around the centroid, whereas all three LLMs showed more regular, locally clustered, and centroid-focused trajectories. Consistent with optimal foraging accounts, human search appears to balance exploitation of local patches with exploration of distant regions (Hills et al., 2012), while LLM outputs appear more biased toward high-probability continuations learned from training data (Lake & Murphy, 2023; Bender et al., 2021). Accordingly, human-like semantic search is not recovered by tuning a single parameter; it requires joint alignment across unpredictability, step magnitude, and global dispersion.

As expected, LLM outputs varied across temperature conditions. However, alignment with humans was architecture-specific. Gemini consistently occupied an intermediate position (less variable than humans but more exploratory than Claude or GPT-4o), GPT-4o showed the most extreme clustering (most conservative global search), and Claude showed the tightest local clustering. Even when collapsing across temperatures, all LLMs differed from humans, suggesting that architectural and training differences create systematic biases in semantic organization that persist regardless of sampling strategy. Through the lens of the dual-component model (Troyer et al., 1997), this pattern is consistent with reduced switching and over-reliance on clustering. Crucially, temperature tuning produced only fragmented alignment: individual metrics could be matched at particular settings, but no configuration reproduced the integrated human profile, with the strongest dissociation in global dispersion (distance to centroid), where GPT-4o remained constrained regardless of stochastic input.

These results challenge the notion that LLMs can serve as straightforward "silicon participants" (Sarstedt et al., 2024) in cognitive research. Most contemporary LLMs undergo extensive post-training with reinforcement learning from human feedback (RLHF) to optimize performance as assistants, not to faithfully replicate human cognitive processes (Ouyang et al., 2022; Bai et al., 2022), which may contribute to introducing systematic deviations from human behavior. In our data, temperature can make models appear human-like on isolated metrics, but not on the joint profile across entropy, step size, and dispersion, creating a central challenge for using LLMs as cognitive models, in which calibrating temperature to match one semantic dimension can distort others. This trade-off suggests that current instruction-tuned LLM architectures may make it difficult to simultaneously capture all dimensions of human semantic search. This also limits cross-study generalizability, because different metrics align with humans at different temperatures, so selecting temperature based on a single outcome may introduce systematic biases. Future work should validate parameter choices against multiple theoretically motivated dimensions rather than optimizing for one measure.

### The Challenge of LLMs as Psychological Surrogates

It is important to acknowledge methodological limitations when interpreting LLM behavior as psychological evidence.

Recent studies show that LLMs often fail to reproduce human strategic behavior in simple experiments and struggle to generalize across semantic variations that humans handle intuitively (Gao et al., 2025; Schröder et al., 2025). Their responses can also be unstable under minor prompt changes and are driven by statistical regularities in token sequences rather than genuine psychological simulation (Gao et al., 2025; Schröder et al., 2025). Our findings contribute to this literature by showing that even when LLMs may produce superficially fluent outputs in semantic fluency, the underlying navigational dynamics diverge systematically from human retrieval. Relatedly, Lu et al. (2025) report that LLM-generated tasks are markedly less social and less physical than human-generated tasks and show a thematic bias toward abstraction.

From a theoretical perspective, LLMs are trained to capture statistical structure in text (Schüle, 2024), and they lack the embodied experience and goal constraints that shape human cognition (Gao et al., 2025). Schröder et al. (2025) further show that even CENTAUR (a model fine-tuned on millions of human responses across many psychological experiments; Binz et al., 2025) fails to preserve human behavioral patterns when items are semantically reworded, despite surface-level similarity. This parallels our observation that temperature manipulation, although it modulates output variability, does not fundamentally change the underlying navigational profile in a way that simultaneously matches humans across multiple dimensions of semantic search. Accordingly, researchers should be cautious when using LLM behavior as evidence for or against cognitive theories without extensive validation across tasks and parameter settings (Niu et al., 2024). Our results underscore this point: similarity on a single metric does not imply equivalence of the full behavioral signature, a dissociation that would be theoretically informative in human data but may instead reflect architectural constraints in current LLMs.

### Human Semantic Search: Unpredictability and Exploration

Our results also offer insights into human semantic memory search. The joint pattern of high entropy, large distance to next, and broad distance to centroid suggests that effective human foraging relies on a specific configuration of variability, one that supports exploration without devolving into random wandering. The entropy findings are particularly informative. Traditional accounts emphasize executive control mechanisms that monitor retrieval success and initiate switches when local patches are exhausted (Troyer et al., 1997; Rosen & Engle, 1997). Our data are consistent with the possibility that semantic navigation introduces controlled stochasticity into transitions, enabling occasional "surprising" moves that depart from purely associative retrieval. Elevated distance to next values reinforces this interpretation. Whereas LLMs tend to take smaller, locally clustered steps consistent with their training statistics, humans produce larger step sizes, aligning with Lévy-walk accounts of semantic search in which optimal foraging combines frequent short movements with occasional long jumps (Rhodes & Turvey, 2007). In this sense, human trajectories appear to implement a balance between local exploitation and unpredictability. Furthermore, the broader distance to centroid may indicate that human search covers diverse regions of semantic space in an efficient way, rather than making directionless leaps, consistent with executive control accounts in which working memory helps track retrieved items and supports strategic shifts across subregions of a category (Baddeley, 2000).

Several limitations warrant consideration. First, we examined only one semantic category. Since categories vary in internal structure, future research should explore potentially specific differences across categories in the human-LLM comparison. Second, while we used OpenAI's text-embedding-3-large and validated results with alternative systems, all embedding models are trained artifacts with particular biases. The human-LLM differences we observed could partially reflect misalignment between embedding space geometry and human semantic structure. Future work could explore embeddings derived from human behavioral data (e.g., semantic fluency norms, feature generation tasks, similarity judgments) that may better capture human semantic organization. Finally, our reliance on proprietary, closed-weight LLMs restricts mechanistic interpretation. Future studies should utilize open-weight, locally runnable models to directly investigate how internal layer representations correlate with behavioral outputs, offering deeper insights into LLM semantic navigation across a broader range of datasets and categories.

### Conclusion

This study offers important insights regarding differences in how humans and LLMs navigate semantic space during verbal fluency tasks, highlighting different searching patterns. While specific temperature settings can approximate human-like performance on isolated metrics, no configuration simultaneously captured the integrated signature of human semantic search. These findings challenge the use of current LLMs as straightforward cognitive models, suggesting that their optimization for assistant-like performance through RLHF may introduce systematic deviations from human cognitive processes. Future research should validate LLM parameters across multiple theoretically motivated dimensions and explore whether alternative training objectives or architectural modifications could better capture the computational principles underlying human semantic memory search.

## Acknowledgments

The authors used large language models to assist with limited aspects of code and manuscript editing. All AI-assisted material was checked and revised by the authors as needed. The authors take full responsibility for the final content of this article.

## References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569. https://doi.org/10.1037/a0038693

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Dassarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T.J., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T.B., Clark, J., McCandlish, S., Olah, C., Mann, B., & Kaplan, J. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *ArXiv, abs/2204.05862*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. Association for Computing Machinery . https://doi.org/10.1145/3442188.3445922.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridis, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., . . . Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, *644*(8056), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology*, *30*, 149–165. https://doi.org/10.1080/00221309.1944.10544467

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). *Language models show human-like content effects on reasoning*. arXiv. https://doi.org/10.48550/arXiv.2207.07051

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. North American Chapter of the Association for Computational Linguistics.

Ding, X., & Wang, L. (2024). Do language models understand time? *Companion Proceedings of the ACM on Web Conference 2025*.

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, *122*(24), Article e2501660122. https://doi.org/10.1073/pnas.2501660122

García, A. M., Ferrante, F. J., Pérez, G., Ponferrada, J., Sosa Welford, A., Pelella, N., Caccia, M., Belloli, L. M. L., Calcaterra, C., González Santibáñez, C., Echegoyen, R., Cerrutti, M. J., Johann, F., Hesse, E., & Carrillo, F. (2025). Toolkit to examine lifelike language v.2.0: Optimizing speech biomarkers of neurodegeneration. *Dementia and Geriatric Cognitive Disorders*, *54*(2), 96–108. https://doi.org/10.1159/000541581

Hansen, H. J., & Hebart, M. N. (2022). *Semantic features of object concepts generated with GPT-3*. arXiv. https://doi.org/10.48550/arXiv.2202.03753

Henry, J. D., & Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology*, *18*(2), 284–295. https://doi.org/10.1037/0894-4105.18.2.284

Herrera-Poyatos, D., Peláez-González, C., Zuheros, C., Herrera-Poyatos, A., Tejedor, V., Herrera, F., & Montes-Soldado, R. (2025). An overview of model uncertainty and variability in LLM-based sentiment analysis: Challenges, mitigation strategies, and the role of explainability. *Frontiers in Artificial Intelligence*, *8*.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Jain, R., Sojitra, D., Acharya, A., Saha, S., Jatowt, A., & Dandapat, S. (2023). Do language models have a common sense regarding time? Revisiting temporal commonsense reasoning in the era of large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6750–6774.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1356

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). *AMMUS: A survey of transformer-based pretrained models in natural language processing*. arXiv. https://doi.org/10.48550/arXiv.2108.05542

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review, 130*(2), 401–431. https://doi.org/10.1037/rev0000297

Lei, Y., Ge, X., Zhang, Y., Yang, Y., & Ma, B. (2025). *Do large language models think like the brain? Sentence-level evidences from layer-wise embeddings and fMRI* [Unpublished manuscript].

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.

Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025). Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, *264*, 242–251. https://doi.org/10.1016/j.procs.2025.07.135

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Morales, D., Chaigneau, S. E., & Canessa, E. (2025). An agent-based model of semantic memory search: Disentangling cognitive control and semantic space organization. *Cognitive Science*, *49*, Article e70155. https://doi.org/10.1111/cogs.70155

Nour, M. M., McNamee, D. C., Liu, Y., & Dolan, R. J. (2023). Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts. *Proceedings of the National Academy of Sciences*, *120*(42), Article e2305290120. https://doi.org/10.1073/pnas.2305290120

Okruszek, Ł., Rutkowska, A., & Wilińska, J. (2013). Clustering and switching strategies during the semantic fluency task in men with frontal lobe lesions and in men with schizophrenia. *Psychology of Language and Communication*, *17*(1), 93–100. https://doi.org/10.2478/plc-2013-0006

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., & Lowe, R.J. (2022). Training language models to follow instructions with human feedback. *ArXiv, abs/2203.02155*.

Pedrotti, A., Rambelli, G., Villani, C., & Bolognesi, M. M. (2025). *How humans and LLMs organize conceptual knowledge: Exploring subordinate categories in Italian*. arXiv. https://doi.org/10.48550/arXiv.2505.21301

Qiu, M., Brisebois, Z., & Sun, S. (2025). *Can LLMs simulate human behavioral variability? A case study in the phonemic fluency task*. arXiv. https://doi.org/10.48550/arXiv.2505.16164

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Ren, Y., Jin, R., Zhang, T., & Xiong, D. (2024). *Do large language models mirror cognitive language processing?* arXiv. https://doi.org/10.48550/arXiv.2402.18023

Renze, M. (2024). The effect of sampling temperature on problem solving in large language models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7346–7356.

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as Lévy foraging. *Physica A: Statistical Mechanics and its Applications*, *385*(1), 255–260. https://doi.org/10.1016/j.physa.2007.07.001

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211

Sanz, C., Carrillo, F., Slachevsky, A., Forno, G., Gorno Tempini, M. L., Villagra, R., Ibáñez, A., Tagliazucchi, E., & García, A. M. (2022). Automated text-level semantic markers of Alzheimer's disease. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *14*(1), Article e12276. https://doi.org/10.1002/dad2.12276

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, *41*, 1254–1270. https://doi.org/10.1002/mar.21982

Schröder, S., Morgenroth, T., Kuhl, U., Vaquet, V., & Paaßen, B. (2025). *Large language models do not simulate human psychology* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.06950

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*, 493–498.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, Article 772. https://doi.org/10.3389/fpsyg.2014.00772

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). Oxford University Press.

Toro-Hernández, F. D., Migeot, J., Marchant, N., Mery, E., El-Hage, W., Toba, M. N., Batrancourt, B., & Volle, E. (2024). Neurocognitive correlates of semantic memory navigation in Parkinson's disease. *npj Parkinson's Disease*, *10*(1), Article 15. https://doi.org/10.1038/s41531-024-00630-4

Toro-Hernández, F. D., Vieira Filho, J., Cabral-Carvalho, R. M. (2026). Characterizing Human Semantic Navigation in Concept Production as Trajectories in Embedding Space. The Fourteenth International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2602.05971

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), 138–146. https://doi.org/10.1037/0894-4105.11.1.138

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. Academic Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

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*, Article 100103. https://doi.org/10.1016/j.chbah.2024.100103

Zemla, J. C., Cao, K., Mueller, K. D., & Austerweil, J. L. (2020). SNAFU: The Semantic Network and Fluency Utility. *Behavior Research Methods*, *52*(4), 1681–1699. https://doi.org/10.3758/s13428-019-01343-w

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., Du, Z., Guo, Q., Li, L., & Chen, Y. (2024). Assessing and understanding creativity in large language models. *Machine Intelligence Research*, *22*, 417–436.

**Supplementary Materials** Supplementary materials for this article, including full statistical reports for all comparisons, are available at https://osf.io/gkp5