Human-likeness of LLMs in the Mental Lexicon

Bei Xiao[1] **Xufeng Duan**[1] **David A. Haslett**[2] **Zhenguang G. Cai**[1,3]
[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[2]Division of Social Science, The Hong Kong University of Science and Technology
[3]Brain and Mind Institute, The Chinese University of Hong Kong
BeiXiao@link.cuhk.edu.hk

Abstract

Recent research has increasingly focused on the extent to which large language models (LLMs) exhibit human-like behavior. In this study, we investigate whether the mental lexicon in LLMs resembles that of humans in terms of lexical organization. Using a word association task—a direct and widely used method for probing word meaning and relationships in the human mind—we evaluated the lexical representations of GPT-4 and Llama-3.1. Our findings reveal that LLMs closely emulate human mental lexicons in capturing semantic relatedness but exhibit notable differences in other properties, such as association frequency and dominant lexical patterns (e.g., top associates). Specifically, LLM lexicons demonstrate greater clustering and reduced diversity compared to the human lexicon, with KL divergence analysis confirming significant deviations in word association patterns. Additionally, LLMs fail to fully capture word association response patterns in different demographic human groups. Among the models, GPT-4 consistently exhibited a slightly higher degree of human-likeness than Llama-3.1. This study highlights both the potential and limitations of LLMs in replicating human mental lexicons, offering valuable insights for applications in natural language processing and cognitive science research involving LLMs.

1 Introduction

Large language models (LLMs) have made significant progress in capturing complex linguistic patterns through self-supervised learning on vast corpora (Brown et al., 2020). Nevertheless, the question remains whether these models merely approximate language based on surface regularities or if they meaningfully align with the deeper cognitive mechanisms underlying human language processing (Cai et al., 2024; Chomsky et al., 2023). Investigating their internal lexical organization—what psycholinguists call the "mental lexicon"—can shed light on whether LLMs' representations go beyond statistical pattern matching to reflect how humans store and retrieve word meanings.

In this study, we examine whether two leading LLMs (at the time of testing, GPT-4o and Llama-3.1) replicate essential properties of the human mental lexicon by leveraging a classic psycholinguistic paradigm: the word association task. By systematically comparing LLM-generated word associations to large-scale human data from the Small World of Words (SWOW) project (De Deyne et al., 2019), we explore how closely lexical organization in LLMs resembles that in humans. In addition, we investigate whether LLMs can accurately reproduce the lexical characteristics unique to different demographic groups when instructed to generate text from these perspectives.

1.1 The Mental Lexicon and Word Association

The mental lexicon is commonly understood as a highly structured, internal system that stores and organizes word-related information, thereby facilitating language comprehension and production (Aitchison, 2012). It encompasses numerous properties of words—including their semantic content, phonological and orthographic representations, syntactic roles, morphological forms, and frequency of use (Jarema and Libben, 2007). Scholars often describe the mental lexicon as a network-like structure, wherein words are interconnected through semantic, phonological, and collocational links (Monakhov and Diessel, 2024; Vitevitch et al., 2014). These networks enable rapid retrieval of lexical information and guide the flow of language processing. Although the mental lexicon cannot be directly observed, a variety of empirical studies—ranging from lexical decision tasks (Balota and Chumbley, 1984) and priming paradigms (Ferrand and New, 2003) to analyses of speech errors

(Stemberger, 1982)—offer converging evidence for its functional organization. Moreover, its structure likely emerges from distributed neural processes underlying language (Jarema and Libben, 2007).

A cornerstone method for probing these lexical connections is the word association task, in which participants list the first words that come to mind given a cue (Rodd et al., 2016; Nelson et al., 2004; Szalay and Deese, 2024). By having participants produce the first word(s) that come to mind, this paradigm helps to reveal associative connections within the mental lexicon (De Deyne and Storms, 2008; Ufimtseva et al., 2020). To capture a richer and more diverse perspective on word relationships, large-scale studies such as the Small World of Words (SWOW) project (De Deyne et al., 2013) employ a multiple-response format in which participants generate three different associative responses for each cue. By assembling extensive datasets from participants of various demographic backgrounds, SWOW enables in-depth investigations of individual and demographic differences in lexical organization (De Deyne et al., 2019). When aggregated across many individuals, these data yield large-scale semantic networks that robustly predict behavioral measures such as lexical decision, naming reaction time, and human-rated word relationships beyond the influence of straightforward lexical statistics like word frequency (Barber et al., 2013; De Deyne et al., 2019; Li et al., 2024). The SWOW norm has proven robust across multiple languages, leading to the construction of mental lexicons for Dutch (De Deyne et al., 2013), English (De Deyne et al., 2019), Mandarin Chinese (Li et al., 2024), and Rioplatense Spanish (Cabana et al., 2024), among others.

1.2 Exploring the Black Box of LLMs Using Behavioral Experimentation

Recent advancements in natural language processing (NLP) benchmarks—including SuperGLUE (Wang et al., 2019) and BIG-bench (Srivastava et al., 2022)—have demonstrated that LLMs excel in tasks such as translation, question answering, cloze tests, textual entailment, and diverse forms of reasoning (Wang, 2018; Srivastava et al., 2022). While these accomplishments highlight the models' versatility and the human-like character of their outputs, they do not clarify whether the underlying processes genuinely resemble human language comprehension or merely represent sophisticated pattern matching (Chomsky et al., 2023; Piantadosi, 2023; Futrell and Mahowald, 2025).

One promising way to bridge this gap is by leveraging behavioral experiments as downstream tasks to evaluate LLMs. These experiments have been instrumental in modeling the cognitive mechanisms that shape human behavior. When adapted for LLMs, they provide a framework to examine whether these models display cognitive patterns comparable to those found in humans. By comparing LLM performance against human responses in well-designed experiments, researchers can gain valuable insights into the language capabilities of these systems. For instance, various psycholinguistic methodologies (e.g., priming) have been employed to explore whether LLMs exhibit language processing patterns akin to human cognition (e.g., Ettinger, 2020; Prasad, 2019; Sinclair et al., 2022).

Several recent studies have applied this methodology to illuminate LLMs' capabilities. Cai et al. (2024) subjected LLMs to a variety of psycholinguistic tasks, finding that the models successfully replicated numerous human-like language processes: forming sound-based associations for unfamiliar words, displaying priming effects in ambiguous word or sentence retrieval, interpreting implausible sentences adaptively, overlooking minor semantic errors, and generating bridging inferences. These models also adjusted causality interpretations in response to verb semantics and tailored language retrieval based on the interlocutor's role. Extending this line of research, Duan et al. (2024b) devised a benchmark to quantify how closely LLMs mirror human language use in phenomena like priming and adaptive sentence interpretation, showing that models such as Llama-3.1 and GPT-4o achieve appreciable levels of human-likeness. Hu et al. (2024) likewise demonstrated that LLMs can replicate human intuitive judgments on diverse grammatical structures.

Despite these promising parallels, researchers have identified key divergences from human cognition. Qiu et al. (2023) reported that LLMs encounter difficulties in pragmatic reasoning, while Cai et al. (2024) highlighted issues such as a failure to prefer shorter words for less informative content and an inability to optimally use context to resolve syntactic ambiguities. Likewise, Dentella et al. (2023) noted that LLMs fall short of humans in accuracy and consistency of grammatical judgments.

Taken together, behavioural experimentation has deepened our understanding of LLMs' language

processing abilities and underscored both their human-like traits and their limitations. The mixed results highlight the importance of continued research aimed at refining our grasp of these models' strengths and shortcomings, particularly through systematic examinations of foundational aspects of language cognition, such as lexical organization.

1.3 Exploring the Mental Lexicon in LLMs Using Word Association

Since LLMs are trained on vast amounts of text data but lack embodied sensory experience, an intriguing question arises: can they understand word relationships purely through textual associations, or is there a crucial role for non-linguistic sensory experience in forming a rich, human-like mental lexicon? Unlike humans, who accumulate word associations through multisensory interactions with the world, LLMs can only infer relationships from the patterns present in the text they are trained on. This raises the central challenge of whether LLMs can approximate the depth of human lexical organization without shared lived experiences.

A well-established approach for probing lexical structure is the word association paradigm (Kumar et al., 2021), which offers a window into the associative networks underlying lexical access. The Small World of Words–English (SWOW-EN) corpus (De Deyne et al., 2019), comprising over 12,000 cue words and responses from approximately 80,000 participants, serves as a robust benchmark for such comparisons. Recent studies by Abramski et al. (2024, 2025) adapted this paradigm to LLMs such as Llama 3, Claude Haiku, and Mistral, generating large-scale word association datasets. Their work investigated lexical diversity, concreteness effects, and bias patterns, and evaluated model-derived semantic networks via priming simulations. Vintar et al. (2024) explored word associations in multilingual and monolingual LLMs (e.g., mT5, SloT5) for Slovene and English, focusing primarily on lexical overlap with human data and categorizing response types

While our study adopts a similar SWOW-style elicitation method, our analytic focus diverges in important ways. We evaluate the extent to which LLMs capture core psycholinguistic dimensions of the mental lexicon—semantic relatedness, associative frequency, lexical entropy, and network clustering—and assess their alignment with human data. We also use KL divergence to quantify distributional differences. Beyond structural comparisons, we further examine whether LLMs reflect sociolinguistic variability observed in human lexical representations. Specifically, we test whether model responses vary systematically across demographic groups, including education level, gender, and age, based on significant sociolinguistic divergence patterns reported in prior work (Garimella et al., 2016, 2017). By integrating structural and sociocognitive perspectives, our study provides a comprehensive assessment of the extent to which LLMs approximate both the organization and variability of the human mental lexicon.

Building on these open questions, the current study examines:

1. To what extent does the mental lexicon in LLMs resemble that of humans in terms of their associative structure and organization?

2. How do different LLM architectures and training approaches influence the human-likeness of their mental lexicon?

3.To what extent does the mental lexicon of LLMs capture demographic variability, akin to the way human word associations vary across factors such as age, cultural background, and personal experience?

To address these questions, we adapted the SWOW-EN word association paradigm for LLMs, using identical cue words and controlling for demographic factors wherever possible. We then modeled each LLM's mental lexicon, with a focus on association frequency, semantic relationships, network properties (such as clustering coefficients), and vocabulary diversity. Our comparisons extended across different LLMs (e.g., GPT-4o and Llama-3.1), as well as between LLMs and human participants. We also examined how demographic aspects might be encoded or omitted in their associative structures.

2 Method

2.1 Models and Human Data

Two state-of-the-art transformer-based language models (at the time of testing) were employed for data collection: GPT-4o, developed by OpenAI, and Llama 3.1-70b-instruct, developed by Meta. For simplicity, these models are referred to as GPT and Llama, respectively, throughout this paper. Human responses were drawn from the SWOW-EN dataset (SWOW-EN.R100.20180827.csv). Only trials contributed by native English speakers were retained, thereby excluding data from non-native

speakers. Trials included in the analysis aligned precisely with those replicated in the model experiments.

2.2 Stimuli and Procedure

A total of 12,281 cue words from the SWOW-EN project (De Deyne et al., 2019) served as stimuli. [1] In the original SWOW-EN dataset, thousands of participants each provided responses to 14–18 of these cue words, resulting in over one million trials.

LLM data were collected in two experiments: one using GPT-4o and the other using Llama-3.1. Each experiment encompassed 1,061,729 trials, mirroring the number of trials from native English speakers in the SWOW-EN dataset. In the experiments, each trial consisted of a single cue word embedded in an instruction prompt (e.g. *...You will receive a cue word. Write the first word that comes to mind...The cue word is...*), accompanied by a system prompt specifying the demographic information corresponding to a trial from the SWOW-EN dataset (i.e., educational level, age, gender, English dialect, and location) (e.g. *You are 33 years old. You are a female...*). This demographically targeted prompting strategy was designed, on one hand, to closely mimic human experimentation and, on the other hand, to provide demographic cues for exploring the potential influence of demographic factors on LLM responses, akin to the variability observed in human language processing. Full example of prompt and response are provided in Appendix B.

All model responses were collected using the R MacBehaviour package (Duan et al., 2024a), a toolkit designed to facilitate behavioral experiments on LLMs. Each trial was run as a discrete chat session containing only one cue word to avoid memory effects, and the package automatically recorded all responses. The default temperature settings for each model were retained: temperature = 1 for GPT-4o and temperature = 0.6 for Llama-3.1.

2.3 Data Preprocessing

Preprocessing steps were performed for both LLM-derived and human-derived responses. Each participant—human or model—provided three responses per cue word, labeled R1, R2, and R3 according to their order. Any additional responses beyond the first three were truncated, and missing responses were coded as NA. Cue words that were not recognized (prompting the model to respond with "unknown word") were also coded as NA. Responses in non-ASCII characters and duplicates within the same cue word were removed.

Further cleaning was conducted using the SWOW-EN preprocessing script (preprocessData.R). This script removed repeated responses for specific cue words, corrected inconsistencies in missing responses (for example, NA coded in R2 but not in R3), and standardized spelling variations.

2.4 Data Analysis

Following data collection and preprocessing, we obtained three datasets—Human, GPT, and Llama—each containing the same cue words, up to three associated responses per cue, and demographic information. Multiple metrics were computed to assess how closely model outputs aligned with human data. These metrics capture distinct yet interrelated key aspects of lexical representation, including word prominence, semantic organization, network topology, and lexical diversity.

Association Frequency. Association frequency, defined as the number of times a word appears as an associate (De Deyne et al., 2019). This measure reflects a word's prominence in the mental lexicon and predicts reaction time (RT) in tasks such as lexical decision, naming, and semantic judgment. We conducted three analyses: (1) correlating association frequencies across datasets, (2) examining correlations between association frequencies and RTs (Balota et al., 2007; Pexman et al., 2017), using both Pearson correlations and partial correlations that controlled for word frequency (English SUBTLEX-US (Brysbaert and New, 2009)), and (3) comparing the top 100 most frequent associates across datasets to evaluate overlap and relative lexical prominence.

Semantic Relatedness. We computed semantic relatedness using a random-walk algorithm applied to cue–associate networks derived from word association data (De Deyne et al., 2016, 2019). Random-walk values for the human dataset were obtained from SWOW-EN, while those for GPT and Llama were generated using the original SWOW-EN script (`graphRandomWalk.R`). Note that semantic relatedness, as measured in this context, encompasses not only taxonomic similarity (e.g., *car–automobile*) but also broader associative relationships, including functional, thematic, orthographic (e.g., *favor–flavor*), and collocational

[1] We excluded the cue "none" from the original 12,282-cue list due to its potential to confound analyses.

links (e.g., *duty–free*) (De Deyne et al., 2019). Because word association networks naturally encode this diverse range of connections, the resulting random-walk scores reflect the associative structure of the mental lexicon beyond pure similarity. To assess the extent to which model-based relatedness aligns with human intuitions, we conducted two analyses: (1) correlating random-walk scores across datasets, and (2) comparing random-walk values with human judgments of semantic similarity from benchmark datasets including MEN (Bruni et al., 2012), MTURK-771 (Halawi et al., 2012), and SimLex-999 (Hill et al., 2015). While these benchmarks specifically target similarity, previous work has shown that random-walk relatedness correlates strongly with human similarity judgments (De Deyne et al., 2019), making them a useful point of comparison.

Network Attributes. Network science offers a systematic framework for analyzing structural properties across diverse domains (Barabási, 2013; Lewis, 2011), including semantic networks (Steyvers and Tenenbaum, 2005). The clustering coefficient is a key metric within this framework, indicating how tightly interconnected the neighbors of a given node are (Newman, 2003; Saramäki et al., 2007). In semantic networks, higher clustering coefficients signify denser interconnections among words, resulting in community-like structures (Palla et al., 2005), as illustrated by Figure 1. In this study, cue–response data were transformed into a weighted directed graph using the *igraph* package in R, creating edges for every cue–response pair. The local clustering coefficient for each node was then computed using the standard formula:

$$C(v) = \frac{2 \times e_i}{k_i(k_i - 1)}$$

where e_i represents the number of edges among neighbors of node *i*, and k_i denotes the degree of node *i*. The distributions of clustering coefficients were compared across human, GPT and Llama networks to assess similarities and differences in structural connectivity.

Vocabulary Diversity. Vocabulary diversity gauges the breadth and variety of words produced, reflecting linguistic adaptability and flexibility (Malvern et al., 2004; Laufer and Nation, 1995). To assess this property, we calculated association entropy for each cue word to evaluate variability in word associations. Shannon entropy

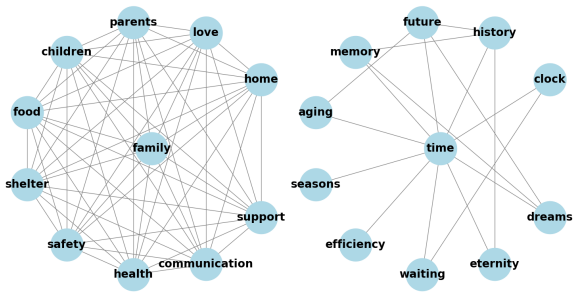

Figure 1: Examples of high and low clustering coefficients. "Family" (left) demonstrates a high clustering coefficient, reflecting dense interconnections among its neighbors, whereas "time" (right) has a low coefficient, indicating sparse connections. Although both words share the same number of immediate neighbors (degree), their internal connectivity differs markedly.

H was computed as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

where $p(x_i)$ is the proportion of a particular word *i* among all responses to a given cue. Higher entropy values reflected a greater spread of responses, whereas lower entropy indicated stronger consensus. These entropy distributions were then compared across the human data and each LLM dataset. Furthermore, we analyzed demographic variability by incorporating demographic factors (e.g., education level, gender) into entropy calculations. We examined interaction effects between demographic levels and groups (human, GPT, Llama) to determine whether demographic factors influence association variability similarly in humans and LLMs or exhibit distinct patterns.

2.5 KL Divergence

In addition to the aforementioned metrics, we computed Kullback–Leibler (KL) divergence to assess the degree of divergence between human-generated and model-generated word association distributions. KL divergence quantifies how much one probability distribution P differs from a reference distribution Q, with lower values indicating greater similarity. It is defined as:

$$\mathrm{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

In all calculations, we defined the human-generated distribution as P and the model-generated distribution (GPT or Llama) as Q, i.e.,

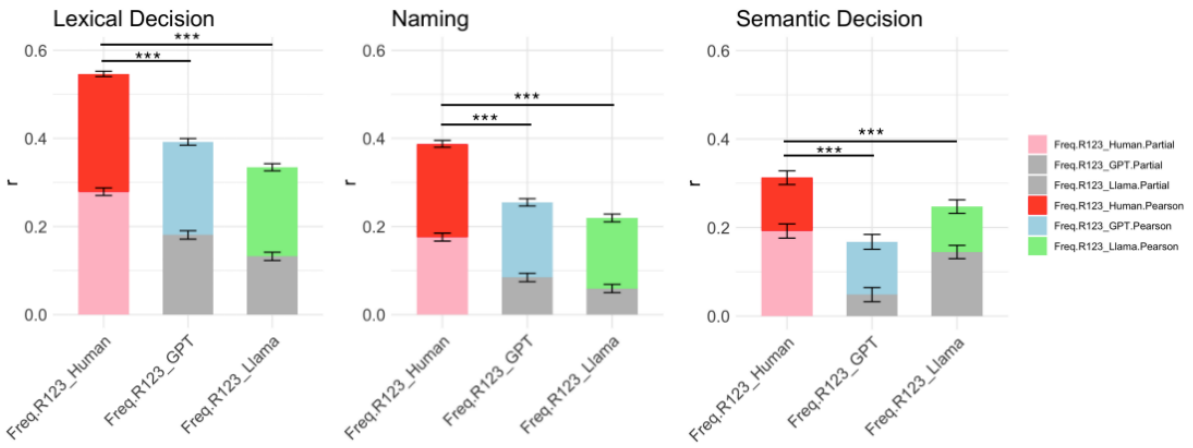

Figure 2: Pearson correlations of association frequencies with lexical decision, naming, and semantic decision RTs. Pink and gray bars depict partial correlations controlling for word frequency (SUBTLEX-US). Freq.R123 is defined as the number of times being an associate, regardless of cue(s), across all associates (R1, R2, and R3) collected in the experiment. For readability, RTs were z-transformed and log-transformed and then shifted to positive values by adding the minimal z-score, while association frequencies were log-transformed after adding a constant of 1. The key finding is that model-derived correlations were significantly weaker than human-derived ones, as indicated by Steiger's Z test ($p < 0.001$ for most comparisons, except for the partial correlation between Llama and human association frequency-RT correlations, where $p = 0.03$). Significance levels: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

we computed KL(Human||Model). This direction reflects the information loss incurred when using model outputs to approximate the human mental lexicon—a standard approach in cognitive modeling. For each cue word, relative-frequency-based probability distributions were derived separately from the Human, GPT, and Llama datasets, and KL divergence was computed accordingly.

3 Results

3.1 Association Frequency

Both GPT and Llama exhibited substantial correlations with human association frequencies, though GPT's association frequency correlated more closely with human data compared to Llama's, a difference confirmed by Steiger's Z test ($Z = 21.43$, $p < 0.001$). See Figure 7 in Appendix C for detail illustration.

Despite the overall correlation among datasets, model-human misalignment emerged when assessing the relationship between association frequency and lexical processing speeds (lexical decision, naming, and semantic decision RTs). Human association frequencies showed the strongest correlations with RT data. While both GPT and Llama significantly predicted RTs, their correlations were consistently weaker than those observed for human data (Figure 2 and Table 1 in Appendix C). The results suggest that while LLM-derived association frequencies capture aspects of lexical processing, they remain less predictive than human-derived frequencies. Partial correlation analyses controlling for word frequency yielded a similar conclusion. While human association frequency continued to show notable correlations with RTs, GPT and Llama each accounted for less variance once word frequency was taken into account (refer to Figure 2 and Table 2 in Appendix C for statistical details).

A comparison of the top 100 words by association frequency (Figure 3 and Figure 4; see Figure 8 Appendix C for Llama's) revealed both overlap and divergence. Words such as "water" and "money" appeared prominently in all lexicons, whereas "sex" was more prominent among humans and "computer" among LLMs. Overall, GPT shared 54% of its top 100 list with humans, compared to Llama's 43%, suggesting that GPT's core associations more closely mirrored human lexical prominence.

3.2 Semantic Relatedness

Random-walk relatedness scores computed using all three associates (R1, R2, R3) revealed that both GPT and Llama correlated strongly with human data, with GPT showing a significantly higher alignment ($Z = 489.38$, $p < 0.001$). See Figure 9 in Appendix C for detailed illustrations.

Figure 3: Top 100 words ranked by association frequency in Human.

Figure 4: Top 100 words ranked by association frequency in GPT.

In the benchmark comparison between model-based relatedness and human semantic similarity judgments, GPT exhibited consistently strong alignment with human responses. According to Steiger's Z test ($p > 0.05$), there was no significant difference between GPT's correlations and those of human random-walk scores across all three benchmarks (MEN, MTurk, and SimLex-999). Llama matched human performance on SimLex-999 alone (see Figure 5). These results suggest that both models—especially GPT—are capable of producing human-like semantic relatedness representations.

3.3 Network Attributes

A linear mixed-effects (LME) model revealed that both GPT and Llama exhibited significantly higher clustering coefficients than humans ($\beta = 0.043$, $t = 36.08$, $p < 0.001$; $\beta = 0.047$, $t = 35.93$, $p < 0.001$). When comparing the models, Llama's clustering coefficient was significantly higher than GPT's ($\beta = 0.004$, $t = 2.58$, $p = 0.01$). See also Figure 10 in Appendix C. These findings suggest that LLM-based semantic networks are more densely interconnected than human networks, with Llama showing the highest degree of local clustering.

3.4 Vocabulary Diversity

An LME analysis showed that both GPT ($\beta = -2.863$, $t = -497.6$, $p < 0.001$) and Llama ($\beta = -2.913$, $t = -506.3$, $p < 0.001$) had significantly lower association entropy compared to humans, indicating reduced lexical diversity. Furthermore, Llama exhibited lower entropy than GPT ($\beta = -0.050$, $t = -8.674$, $p < 0.001$); see Figure 11 in Appendix C.

3.5 KL Divergence

The KL divergence analysis revealed notable differences between human word associations and those generated by GPT and Llama. The average KL divergence—computed as KL(Human||Model)—was 11.09 for GPT and 12.46 for Llama, both indicating substantial deviation from the human distribution. A t-test comparing these values yielded a significant difference ($t = -49.04$, $p < .001$), suggesting that GPT's word association distributions more closely resemble human responses than those of Llama.

3.6 Examining Demographic Variability in LLM Mental Lexicon

A demographic analysis using association entropy and linear regression revealed significant interactions between education level and source group (Human, GPT, or Llama). While models captured general education-related entropy trends (with a visually similar pattern for age in Figure 13, Appendix C), they diverged from human patterns, particularly among higher education groups (Figure 6). In human data, bachelor's degrees exhibited significantly higher entropy than master's ($\beta = 0.136$, $p < 0.001$), a difference absent in GPT and Llama (GPT: $\beta = -0.025$, $p = 0.960$; Llama: $\beta = 0.005$, $p > 0.999$). Llama also failed to replicate entropy differences between high school and bachelor's ($\beta = -0.022$, $p > 0.999$) or master's degrees ($\beta = -0.017$, $p > 0.999$), compared to humans (high school vs bachelor: $\beta = -0.311$, $p < 0.001$; high school vs master: $\beta = -0.174$, $p < 0.001$). GPT captured these differences with slightly smaller effect sizes for the high school-bachelor comparison ($\beta = -0.046$, $p = 0.030$). These findings suggest that while models capture broad demographic-related entropy trends (and align with human data in some aspects, such as gender variability; see Figure 12 in Ap-

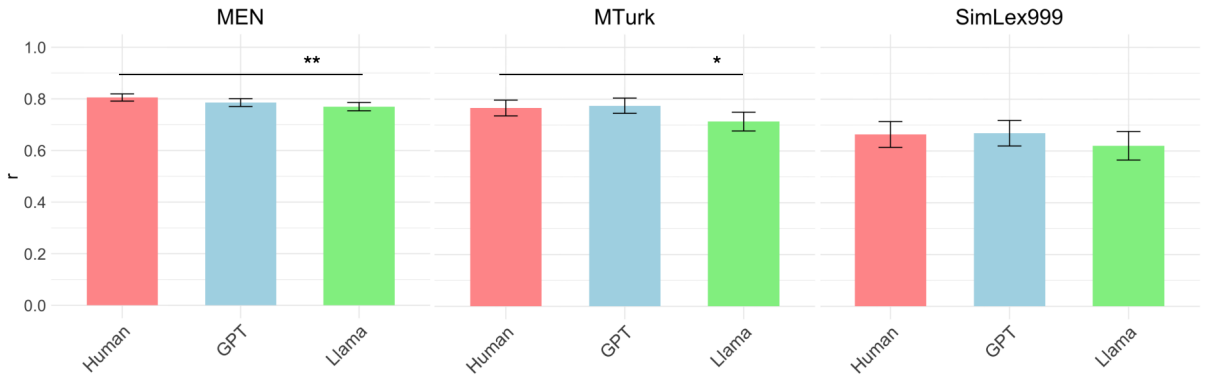

Figure 5: Pearson correlations and 95% confidence intervals between random-walk relatedness scores and direct semantic similarity ratings from MEN, MTurk, and SimLex999. *: $p < 0.05$; **: $p < 0.01$.

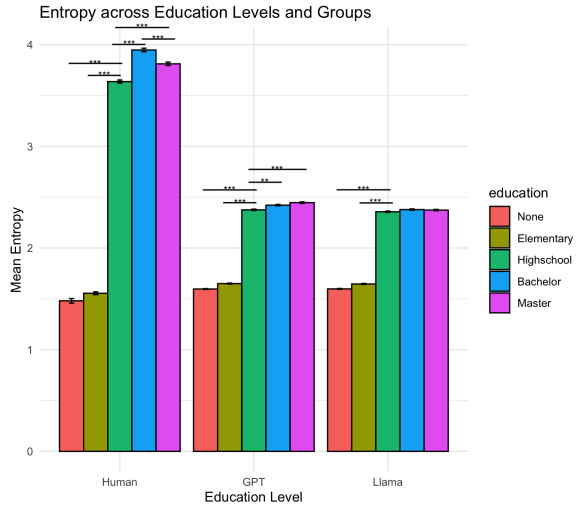

Figure 6: Entropy differences in association for education groups across Human, GPT, and Llama datasets. **: $p < 0.01$; ***: $p < 0.001$.

pendix C), they exhibit limited capacity for capturing fine-grained differences, particularly in educational entropy. Llama deviates more from human patterns in educational contexts than GPT does.

4 Discussion

Our study provides mixed findings regarding the human-likeness of LLMs in replicating the mental lexicon, with semantic relatedness emerging as the most consistent parallel to human performance. This aligns with Abramski et al. (2025), who noted comparable semantic priming effects in both human and model-based networks, highlighting human-like features in LLMs' semantic associations. While association frequency analysis suggests LLMs capture some aspects of human-like prominence in word associations, they primarily encode straightforward lexical statistics like word frequency, rather than deeper cognitive associations.

A significant divergence was observed in the higher clustering coefficient and lower lexical diversity of LLM-based semantic networks compared to human counterparts. Additionally, KL divergence analysis revealed discrepancies between human and model-generated word associations, indicating that while LLMs replicate certain human-like semantic relations, they lack the depth and range of human mental lexicons. This may be due to the absence of embodied sensory experience during model training, which limits their ability to fully capture the complexities of human language cognition.

Our comparison of GPT and Llama highlighted consistent patterns, with GPT generally displaying stronger human-like qualities. This suggests that variations in training strategies and data sources may significantly influence model performance, underscoring the impact of model architecture and training choices on LLM behavior.

Our findings also carry implications for the use of LLMs as surrogate participants in cognitive science research, a notion gaining traction in recent studies (e.g. Duan et al., 2024a; Qin et al., 2024). While LLMs offer a cost-effective alternative for semantic-relatedness studies, their discrepancies with human mental lexicons caution against overreliance on them as surrogates. Issues such as the misrepresentation of social identities, raised by Wang et al. (2025), are particularly relevant here, as our results suggest LLMs fail to fully capture demographic variability and diversity accurately, at least in terms of word association. This reinforces concerns that LLMs may oversimplify or misrepresent human experiences, especially in studies involving

identity and diversity. This concern is further compounded by the growing reliance on synthetic data in model training (del Rio-Chanona et al., 2024; Shumailov et al., 2024), which may lead to even less spontaneous and more constrained language representations, thereby limiting LLMs' ability to reflect nuanced human variability.

A key interpretive challenge is whether the observed demographic insensitivity stems from inherent limitations in model representations or from insufficiently strong persona conditioning. Findings by Hu and Collier (2024) suggest that even structured demographic prompting typically explains less than 10% of the variance in human responses across subjective NLP tasks. This modest effect implies that LLMs may require more detailed and contextually grounded persona descriptions to meaningfully reflect individual-level variation. Thus, our findings likely reflect both limited model responsiveness to demographic cues and the inadequacy of surface-level prompts in shaping behaviorally distinct outputs. Future work should explore more effective strategies for enhancing demographic control and further delineate the conditions under which persona prompting can elicit interpretable variation aligned with human diversity.

A key methodological consideration concerns our reliance on prompting, rather than directly extracting conditional probabilities from the model's output distribution or other internal representations. While prompting provides an intuitive and human-aligned interface that mirrors task formats commonly used in psycholinguistic research, it may introduce a layer of metalinguistic reasoning that obscures the model's underlying semantic representations. Recent work by Hu and Levy (2023) highlights this limitation, arguing that prompting requires models to interpret linguistic input, thereby testing metalinguistic judgment rather than directly revealing internal representations. To explore the feasibility of probability-based evaluation, we conducted preliminary analyses using log-probabilities sampled directly from the model. However, a substantial proportion of high-probability outputs consisted of subword tokens (e.g., `"un"`, `"ther"`), complicating alignment with human lexical data and introducing nontrivial post-processing assumptions for reconstructing full-word responses. Given these practical constraints, we adopted prompting to ensure interpretability and consistency with behavioral baselines. Nonetheless, we acknowledge that this approach may limit access to deeper representational signals within the model. Future work should consider hybrid frameworks that integrate prompting with direct probability-based measures, enabling a more comprehensive assessment of model–human alignment under varying input modalities.

5 Conclusion

In conclusion, while LLMs demonstrate some human-like properties in their mental lexicons, they fail to fully replicate the complexity of human semantic networks. The observed discrepancies in lexical diversity and network structure reveal fundamental differences between human and machine cognition. As LLMs continue to evolve, further research is essential to refine these models to better capture the nuanced, multimodal nature of human language. Caution is also needed when using LLMs as substitutes for human participants, particularly in studies involving social identity and linguistic diversity.

References

Katherine Abramski, Riccardo Improta, Giulio Rossetti, and Massimo Stella. 2025. The "llm world of words" english free association norms generated by large language models. *Scientific data*, 12(1):1–9.

Katherine Abramski, Clara Lavorati, Giulio Rossetti, and Massimo Stella. 2024. Llm-generated word association norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 3–12. IOS Press.

Jean Aitchison. 2012. *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

David A Balota and James I Chumbley. 1984. Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, 10(3):340.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39:445–459.

Albert-László Barabási. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Horacio A Barber, Leun J Otten, Stavroula-Thaleia Kousta, and Gabriella Vigliocco. 2013. Concreteness in word processing: Erp and behavioral effects in a lexical decision task. *Brain and language*, 125(1):47–53.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Álvaro Cabana, Camila Zugarramurdi, Juan C Valle-Lisboa, and Simon De Deyne. 2024. The" small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, 56(2):968–985.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8.

Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior research methods*, 45:480–498.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 1861–1870.

Simon De Deyne and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231.

R Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2024. Large language models reduce public knowledge sharing on online q&a platforms. *PNAS nexus*, 3(9):pgae400.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Xufeng Duan, Shixuan Li, and Zhenguang G Cai. 2024a. Macbehaviour: An r package for behavioural experimentation on large language models. *Behavior Research Methods*, 57(1):19.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G Cai. 2024b. Hlb: Benchmarking llms' humanlikeness in language use. *arXiv preprint arXiv:2409.15890*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ludovic Ferrand and Boris New. 2003. Semantic and associative priming in the mental lexicon. *Mental lexicon: Some words to talk about words*, pages 25–43.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 conference on empirical*

methods in natural language processing, pages 2285–2295.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers*, pages 674–683.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Gonia Jarema and Gary Libben. 2007. *The mental lexicon: core perspectives*, volume 1. Elsevier Amsterdam.

Abhilasha A Kumar, Mark Steyvers, and David A Balota. 2021. Semantic memory search and retrieval in a novel cooperative word game: A comparison of associative and distributional semantic models. *Cognitive Science*, 45(10):e13053.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied linguistics*, 16(3):307–322.

Ted G Lewis. 2011. *Network science: Theory and applications*. John Wiley & Sons.

Bing Li, Ziyi Ding, Simon De Deyne, and Qing Cai. 2024. A large-scale database of mandarin chinese word associations from the small world of words project. *Behavior Research Methods*, 57(1):34.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development*. Springer.

Sergei Monakhov and Holger Diessel. 2024. Complex words as shortest paths in the network of lexical knowledge. *Cognitive Science*, 48(11):e70005.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814–818.

Penny M Pexman, Alison Heard, Ellen Lloyd, and Melvin J Yap. 2017. The calgary semantic decision project: concrete/abstract decision data for 10,000 english words. *Behavior research methods*, 49:407–417.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414.

G Prasad. 2019. Using priming to uncover the organization of syntactic representations in neural language models. *arXiv preprint arXiv:1909.10579*.

Xin Qin, Mingpeng Huang, and Jie Ding. 2024. Aiturk: Using chatgpt for social science research. *Available at SSRN 4922861*.

Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. 2023. Pragmatic implicature processing in chatgpt.

Jennifer M Rodd, Zhenguang G Cai, Hannah N Betts, Betsy Hanby, Catherine Hutchinson, and Aviva Adler. 2016. The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, 87:16–37.

Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. 2007. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 75(2):027105.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Joseph Paul Stemberger. 1982. The nature of segments in the lexicon: Evidence from speech errors. *Lingua*, 56(3-4):235–259.

Mark Steyvers and Joshua B Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

Lorand B Szalay and James Deese. 2024. *Subjective meaning and culture: An assessment through word associations*. Taylor & Francis.

Natalia V Ufimtseva et al. 2020. Association-verbal network as a model of the linguistic picture of the world. *European Proceedings of Social and Behavioural Sciences*.

Špela Vintar, Mojca Brglez, and Aleš Žagar. 2024. How human-like are word associations in generative models? an experiment in slovene. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024*, pages 42–48.

Michael S Vitevitch, Rutherford Goldstein, Cynthia SQ Siew, and Nichol Castro. 2014. Using complex networks to understand the mental lexicon. In *Yearbook of the Poznań Linguistic Meeting*, volume 1. Uniwersytet im. Adama Mickiewicza w Poznaniu.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

A Appendix A: Limitations

This study uses a psycholinguistic method (word association) to explore the mental lexicon of LLMs and the extent to which it resembles that of humans. A more comprehensive understanding of LLM lexical organization could involve additional metrics, such as network attributes that capture both local and global properties. Furthermore, a philosophical or theoretical grasp of an LLM's human-like capabilities in language understanding, production, and acquisition necessitates broader examination frameworks and careful analysis of internal mechanisms.

Our significant finding is that the divergence between LLM and human mental lexicons in terms of lexical diversity may be partly constrained by technical factors, such as the temperature parameter used to ensure consistent output. In addition, during model training, "meta-controls" are added to regulate content generation (e.g., overly vulgar content), which is crucial for safe use but objectively limits word association divergence. This might explain why certain words prominent in human mental lexicons, such as "sex," are less so in LLMs according to our results. Some immediate associations might have been restricted based on these factors. Nonetheless, we believe these factors do not account for all divergences and likely represent only a small portion influencing our results.

Further limitations arise from the demographic variability analysis where certain groups—like those with "no formal education," "elementary school," or specific accents—had limited data. This reduced sample size weakens statistical comparisons and underscores the need for more balanced datasets reflecting diverse human profiles. Additionally, filtering for native English speakers led to an imbalanced word association dataset with 63 to 100 valid trials per cue ($M = 86$, $SD = 6.55$). Although both human and model groups faced similar testing conditions, future research would benefit from more evenly distributed data to enhance reliability and detail. Despite these constraints, our findings offer preliminary insights into how LLMs resemble and differ from human mental lexicons and suggest promising avenues for further investigation.

B Appendix B Sample Prompts and Response

System Prompt: You are 33 years old. You are a female. You are a native speaker of English who grew up in Australia.

Prompt: On average, an adult knows about 40,000 words, but what do these words mean to people? You can help scientists understand how meaning is organized in our mental dictionary by playing the game of word associations. This game is easy: Just give the first three words that come to mind.

Instructions: You will receive a cue word. Write the first word that comes to mind when reading this word. If you don't know this word, write 'unknown word'. Then write a second and third word, or write 'unknown word' if you can't think of any.

Please respond in the following format: [FIRST WORD; SECOND WORD; THIRD WORD]. Please don't ask any questions or give any other information.

The cue word is: although

Response: but; however; yet

C Appendix C Supplementary Figures and Tables for Results

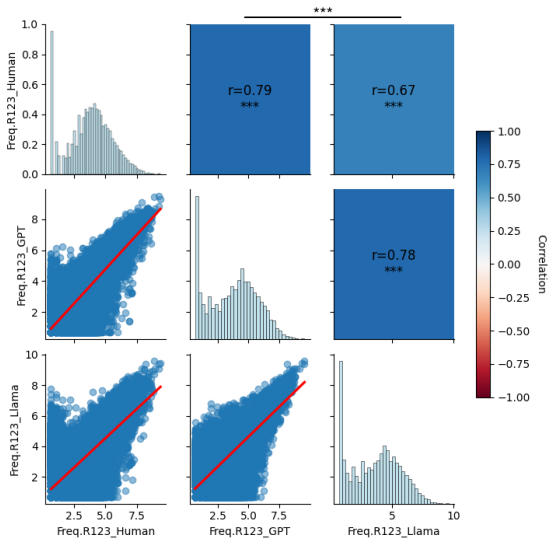

Figure 7: Correlation of association frequencies among Humans, GPT, and Llama. For readability, values were log1p-transformed (adding 1 before taking the natural logarithm). The upper triangle displays Pearson correlation heatmaps, the lower triangle shows scatter plots with fitted regression lines, and the diagonal provides histograms of Freq.R123 distributions. (Freq.R123 is defined as the number of times being an associate, regardless of cue(s), across all associates (R1, R2, and R3) collected in the experiment). ***: $p < 0.001$.

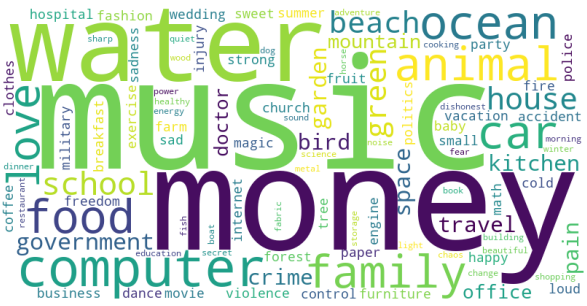

Figure 8: Top 100 words ranked by association frequency in Llama.

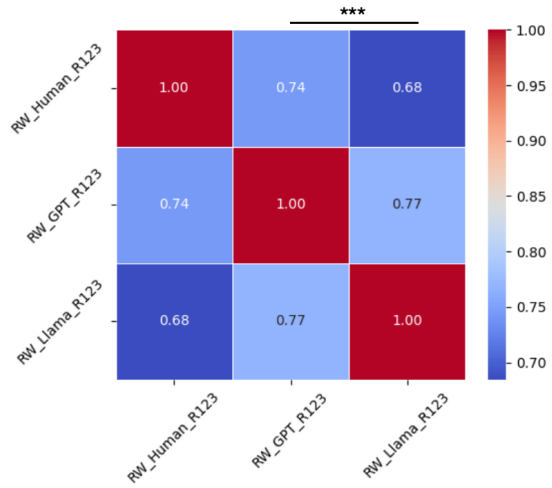

Figure 9: Pearson correlation coefficients for random-walk measures based on all associates (R1, R2, R3) from the Human, GPT, and Llama datasets. ***: $p < 0.001$.

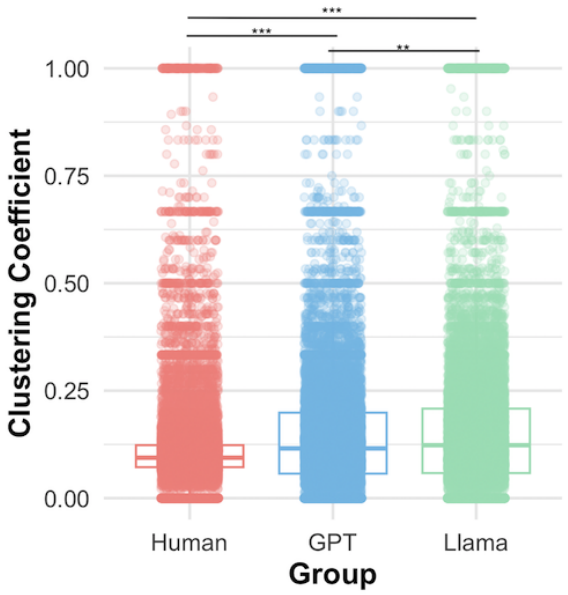

Figure 10: Clustering coefficients in the semantic networks of Human, GPT, and Llama. **: $p < 0.01$; ***: $p < 0.001$.

Table 1: Pearson and partial correlations between association frequency and lexical processing RTs, along with Steiger's Z tests comparing model correlations and human correlations. Significance in Steiger's Z tests indicates misalignment with human association frequency–RT correlation size.

	Pearson correlation			**Steiger's Z test**		**Partial correlation**			**Steiger's Z test**	
	r	*p*	**N**	*Z*	*p*	*r*	*p*	**N**	*Z*	*p*
Lexical decision										
Human	0.54	<0.001	11,928	–	–	0.27	<0.001	11,928	–	–
GPT	0.39	<0.001	11,928	21.18	<0.001	0.18	<0.001	11,928	18.99	<0.001
Llama	0.33	<0.001	11,928	27.10	<0.001	0.13	<0.001	11,928	12.48	<0.001
Naming										
Human	0.39	<0.001	11,968	–	–	0.18	<0.001	11,968	–	–
GPT	0.25	<0.001	11,968	12.96	<0.001	0.08	<0.001	11,968	7.35	<0.001
Llama	0.22	<0.001	11,968	16.07	<0.001	0.05	<0.001	11,968	9.34	<0.001
Semantic decision										
Human	0.31	<0.001	3,932	–	–	0.19	<0.001	3,932	–	–
GPT	0.17	<0.001	3,932	7.27	<0.001	0.05	0.002	3,932	6.54	<0.001
Llama	0.25	<0.001	3,932	3.82	<0.001	0.15	<0.001	3,932	2.19	0.03

Table 2: Pearson correlation and Steiger's Z test results for random walk measures between Human, GPT, and Llama on MEN, MTurk, and SimLex999 benchmarks.

		Pearson correlation		**Steiger's Z test**	
Benchmark	**Model**	*r*	*p*	*Z*	*p*
MEN	Human	0.80	<0.001	–	–
	GPT	0.79	<0.001	1.80	0.07
	Llama	0.77	<0.001	3.15	0.002
MTurk	Human	0.77	<0.001	–	–
	GPT	0.77	<0.001	0.39	0.70
	Llama	0.71	<0.001	2.06	0.04
SimLex-999	Human	0.66	<0.001	–	–
	GPT	0.67	<0.001	0.13	0.90
	Llama	0.66	<0.001	1.08	0.27

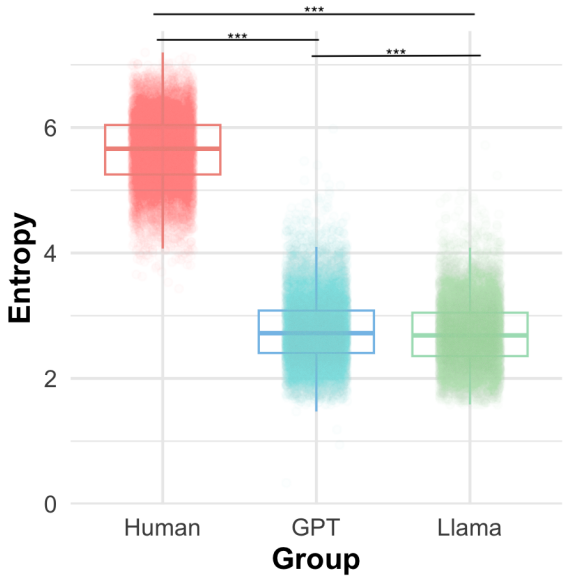

Figure 11: Entropy values for cue words across Human, GPT, and Llama data. ***: $p < 0.001$.

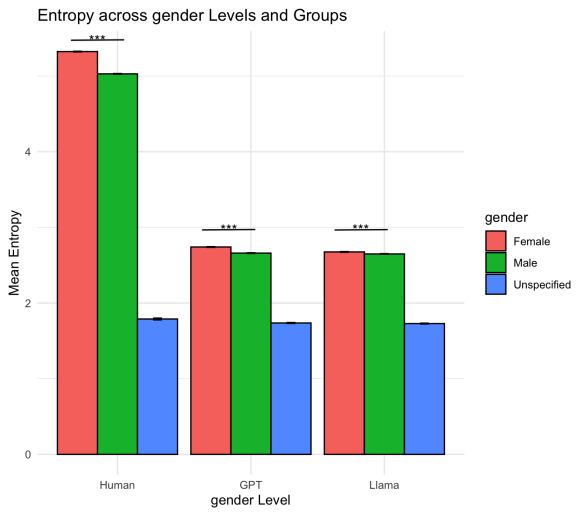

Figure 12: Entropy differences in association for gender groups across Human, GPT, and Llama datasets. ***: $p < 0.001$.

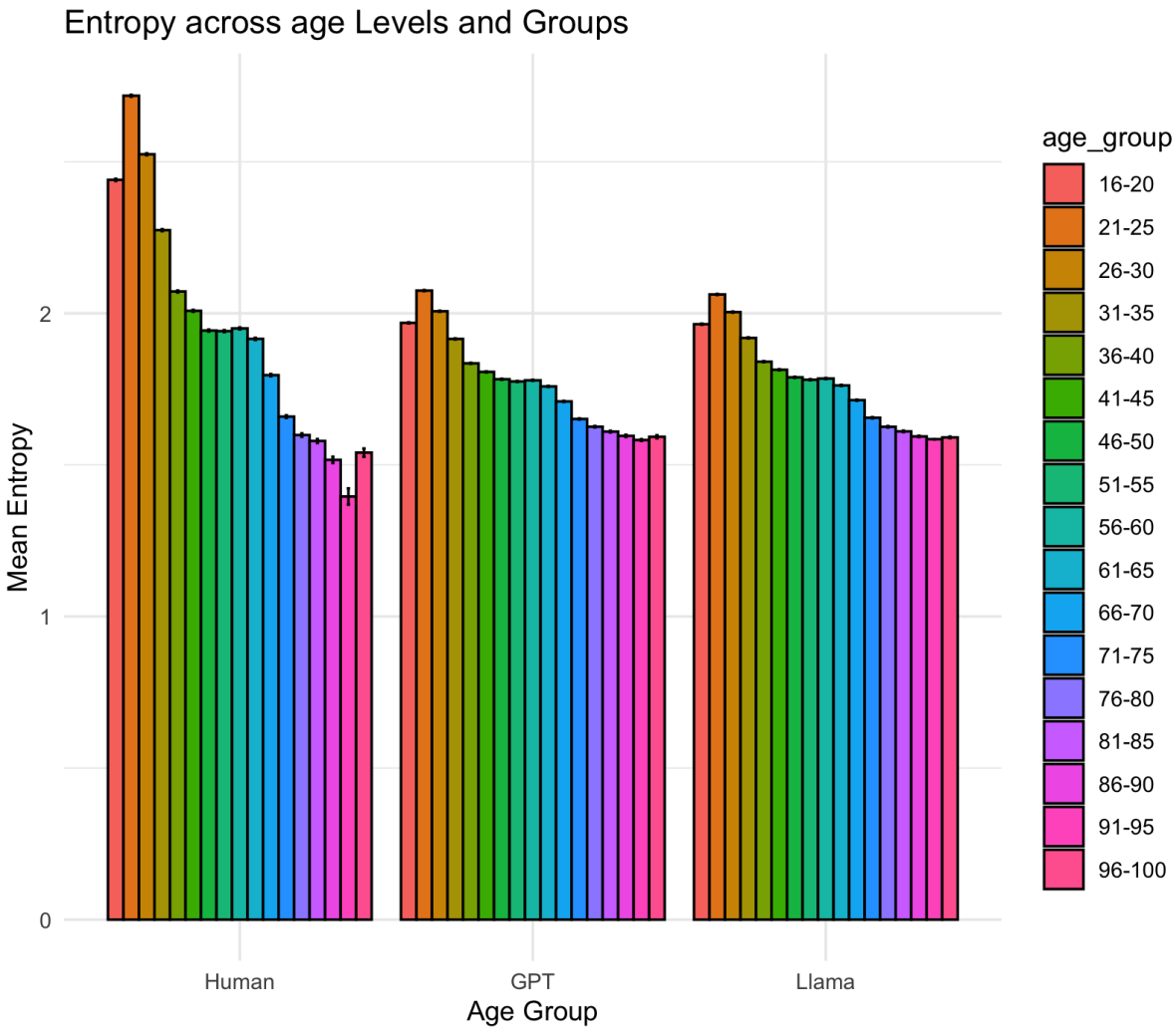

Figure 13: Entropy differences in association for age groups across Human, GPT, and Llama datasets.