Gizem Cınar[1,*], Melissa Nur Robinson[1] and Tan Arda Gedik

Does GPT-5 dream of (il)literate minds? A network analysis of semantic fluency

https://doi.org/10.1515/gcla-2025-0103

Abstract: Research comparing human and machine language abilities has largely relied on WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations, neglecting the cognitive diversity introduced by literacy variation. This study investigates how literacy and text-based training shape semantic organization by comparing illiterate and literate Turkish adults with ChatGPT-5. Participants and the model completed a 1-minute semantic fluency task in three categories (animals, fruits, household objects). Responses were analyzed using network science metrics: average shortest path length (ASPL), clustering coefficient (CC), and modularity (Q). Results showed a stepwise pattern across groups. Illiterate speakers produced smaller but locally dense networks (short ASPL, high CC, low Q). Literate speakers demonstrated more balanced, small-world-like structures (moderate ASPL and CC). GPT-5, while generating the largest vocabulary, produced fragmented networks with long paths, low clustering, and high modularity. These findings suggest that literacy promotes efficient integration of semantic knowledge, whereas text-only learning yields over-segmented "hyper-literate" structures. The results highlight how embodied experience and print exposure jointly shape semantic memory organization in humans and GPT-5.

Keywords: LLMs, semantic network, semantic fluency, GPT, literacy

1 Introduction

Cognitive sciences have long relied on data from Western, Educated, Industrialized, Rich, and Democratic societies (WEIRD, Henrich et al. 2010). This sampling bias, however, limits the generalizability of psychological theories, leading to models that

1 The first two authors share joint first authorship.

* **Corresponding author: Gizem Cınar**, Bilkent University, 7 Neuroscience, 1979. Sk, Ankara, Turkey, g.cinar@bilkent.edu.tr

Melissa Nur Robinson, Bilkent University, Psychology, H Building, Üniversiteler Mah., Ankara 06800, Turkey, robinsonmelissanur@gmail.com

Tan Arda Gedik, Bilkent University, FAU Erlangen-Nürnberg, Middle East Technical University, Psychology, H Building, Üniversiteler Mah., Ankara 06800, Turkey, tan.gedik@bilkent.edu.tr

reflect only a narrow segment of human cognition (e.g., Blasi et al. 2022). A similar pattern has arguably recently emerged in comparisons between Large Language Models (LLMs) and humans: the "human" benchmark has been a highly educated, highly literate WEIRD sample based as far as the emerging research is concerned (Qiu et al. 2025; Wang et al. 2025). Yet non-WEIRD populations differ from WEIRD populations in profound ways, particularly in literacy experience, which has been argued to be one of the most powerful predictors of individual differences in linguistic knowledge (e.g., Dąbrowska 2012, 2021). When LLM–human comparisons use only highly literate individuals as the standard, the resulting generalizations reinforce an already restricted view of what "human semantic cognition" may look like (Blasi et al. 2022). This poses a theoretical problem: frameworks derived exclusively from literate participants have limited cognitive plausibility and reduced cross-cultural applicability. To avoid overgeneralizing claims about the human mind, it is therefore essential to incorporate non-WEIRD, low-literate, or illiterate populations into such comparisons with LLMs.

One domain where these issues become particularly visible is semantic fluency, a task that has been used to compare humans and LLMs (Qiu et al. 2025; Wang et al. 2025). In a semantic fluency task, individuals generate as many items as possible from a predetermined semantic category within a fixed time window. Although seemingly simple, semantic fluency depends on semantic memory, executive control, and lexical access, making the task widely used in neuropsychological assessment to identify cognitive impairment, track age-related decline, and study language processing (Strauss et al. 2006). The most common version asks participants to name as many animals as possible in 60 seconds (e.g., Brucki and Rocha 2004; Mitrushina et al. 2005), but researchers have also examined other categories, such as fruits and household objects, including in populations with varying literacy levels (Yasa Kostas et al. 2024). Because performance reflects both the structure and accessibility of semantic knowledge, semantic fluency provides a powerful window into the organization of human semantic memory, particularly when analyzed using semantic network methods (Kumar et al. 2022).

In human participants, several research studies argue that literacy introduces crucial variation in semantic fluency performance. Literate adults typically outperform illiterate adults (da Silva et al. 2004; Kosmidis et al. 2004), though the gap can narrow for experientially grounded categories such as foods (da Silva et al. 2004). However, Gedik and De la Garza (2025) demonstrated robust literacy effects even for grounded categories like fruits and household objects, suggesting that literacy shapes performance beyond experiential familiarity. They propose that literacy exerts cumulative effects on semantic processing by enhancing working memory, executive control, and lexical organization. Extensive reading practice refines the neural pathways involved in phonological processing and attention

(Dehaene et al. 2010), strengthens connectivity across language and conceptual systems (Mohammadi et al. 2020), and expands vocabulary through exposure to print. Taken together, literacy provides a rich source of variation for probing how humans organize meaning—a dimension entirely absent from LLM training.

At the same time, interest in comparing LLMs with humans on language-related tasks has increased dramatically with the recent availability of LLMs. Trained on large-scale text corpora, LLMs achieve impressive performance on question-answering, summarization, and other linguistic tasks (Brown et al. 2020). They can approximate human-like similarity judgments, feature norms, and categorization behavior (Digutsch and Kosinski 2023; Hansen and Hebart 2022; Suresh et al. 2023). Yet it remains unclear whether this surface-level similarity reflects human-like semantic organization or merely sophisticated statistical pattern learning (e.g., Marcus 2020). Crucially, LLMs lack perceptual and embodied experience (such as literacy), which are key ingredients of human semantic memory, raising questions about whether their internal representations support structured retrieval processes like semantic fluency in a manner comparable to humans.

To address this research gap, recent studies have begun evaluating LLMs and humans through the lens of semantic networks. Conceptualizing semantic memory as a graph of interconnected nodes, network analyses quantify structural properties such as average shortest path length (ASPL), clustering coefficient (CC), and modularity Q, metrics associated with efficient navigation and small-world organization (Wang and Chen 2003; Watts and Strogatz 1998). Using these metrics, Wang et al. (2025) found that LLM-generated networks exhibit longer ASPL, lower CC, and higher modularity compared to human networks, indicating a more fragmented and less efficient structure. Similarly, Qiu et al. (2025) reported that LLM networks lack the small-world characteristics seen in humans. Despite the fact that these studies are very helpful in warning researchers not to substitute human participants with LLMs, these comparisons all rely on highly literate WEIRD participants, leaving open whether the differences between LLMs and humans persist or shift when literacy experience itself varies, even though the fact that literacy is known to reshape human semantic networks, as mentioned earlier.

Thus, although prior research shows that LLM-based networks differ from human networks, these findings do not address whether LLMs resemble humans with different literacy backgrounds. Even among humans, semantic network structure varies systematically with literacy, education, and experience, yet surprisingly, no study has directly compared literacy-related variation in human semantic networks with those derived from an LLM.

The present exploratory study partially fills this gap. We compare ChatGPT-5's performance on a 1-minute semantic fluency task across three categories (animals, fruits, household objects) with publicly available data from Gedik and De la Garza

(2025), which includes literate and illiterate Turkish adults. To our knowledge, this is the first study in Turkish to evaluate an LLM alongside both literate and illiterate groups on semantic fluency. We address two main questions:

1. How do ChatGPT-5, literate adults, and illiterate adults differ in their semantic fluency performance across categories?
2. How similar or different are their semantic networks in terms of ASPL, clustering, and modularity?

By examining these questions, we aim to determine whether the semantic organization of a text-trained model aligns more closely with literate, illiterate, or neither group, and to what extent literacy and embodied experience produce patterns of semantic structure that diverge from those of an LLM.

2 Methods

2.1 Participants

Gedik and De la Garza (2025) collected data from 24 illiterate (23 females, mean age = 51.5, SD = 11.66) and 20 literate (17 females, mean age = 43.95, SD = 14.93) adult native Turkish speakers. The illiterate participants were based in Ankara, Turkey, and at the time of the study, had been attending the adult education center for literacy courses for an average of 9.52 months (SD = 7.34). The curriculum included literacy education along with subjects like math, basic history, and Turkish, spread across 80 teaching units, with each unit lasting 40 minutes. Upon successful completion of these courses, participants receive a degree equivalent to a primary school diploma in Turkey. The literacy center screened speakers in conjunction with health experts from local hospitals for various speech and communication impairments, such as dyslexia, as well as cognitive impairments. Speakers with specific speech or cognitive impairments are not admitted to the literacy course where our data was collected. Literate speakers, on the other hand, were people who learned to read and write in early childhood and finished at least high school. Literate speakers were also recruited from Ankara, Turkey.

2.2 Semantic fluency task

2.2.1 Humans

The experiment included three word-generation categories administered sequentially: animals, fruits, and household objects (presented in this order). Researchers

framed the activity as a competitive game, emphasizing speed and precision within time constraints.

Before the actual test, participants completed a practice round using "weather conditions" as the category. The experimenter demonstrated by saying, "How many weather conditions can you name? Let us name a few together. There is sunny, rainy... What else can you name? Help me." to encourage word generation. Participants received immediate feedback on their responses in the practice round - for example, when someone said "temperature," they were told it did not qualify because, while related to weather, it is not actually a weather condition itself. Testing only began after participants confirmed they understood the rules.

For each of the three main categories, participants had exactly one minute to generate as many relevant words as possible. Participants were encouraged twice to continue generating items if they had time remaining. Questions about category boundaries could be asked beforehand (such as "Does any animal count?"), but no assistance was provided once the timing started. Researchers intentionally avoided suggesting organizational strategies, allowing participants to use whatever cognitive approaches came naturally to them.

All sessions were audio recorded for later analyses. Two native Turkish speakers independently coded the responses with perfect agreement. Every response was documented regardless of category relevance, though only semantically appropriate and unique words were counted in the final scores (for instance, if someone said "strawberry, apple, cucumber" for fruits, only the first two would count).

2.2.2 GPT-5

We collected additional LLM-generated responses by interacting with the model. Instead of using an API-based scripted pipeline, prompts were delivered in natural-language dialogue within the ChatGPT interface. Manual prompting allowed us to replicate the naturalistic nature of real semantic fluency tasks, and allowed us to keep the methodology between the human participants and GPT-5 similar.

In administering the semantic fluency task, we adapted the role-playing procedure described by Wang et al. (2025). Role-playing with LLMs helps to generate personalized output (Shanahan et al. 2023; Zhao et al. 2025). Zheng et al. (2023) demonstrate that interpersonal roles can improve the performance of LLMs over several questions in a conversation. Thus, while collecting data from GPT-5, we created a profile using the demographic of the literate speakers, specifically their age, what educational attainment they had, and how many years of formal schooling they received (see Figure 1 below). These demographic characteristics were provided through a single role-based prompt conditioning at the beginning of each

task, allowing the LLM to adopt the specified characteristics within the conversational context (see Figure 1 below).

Our implementation of role-playing also diverged in several respects. First, following the written demographics and instruction prompts, we employed the voice mode (Voice: Maple in Turkish) to ensure standardized timing across sessions. Since LLMs do not have time perception, we assisted the model regarding the remaining time. When the system prematurely terminated responses, we prompted it to continue twice at most until the minute was over, paralleling the procedure used with human participants. However, because we also used the voice mode, we subjected GPT-5 to similar phonological decoding constraints (i.e., there are only so many words one can utter in speech), and therefore kept the testing conditions similar across the board between human and non-human participants. We only collected the responses ChatGPT-5 managed to speak out loud within the 60 seconds. Second, we simulated (role-played) the demographic characteristics of our literate speaker group. We did not conduct role-play sessions for the illiterate group, as this yielded no meaningful differences in an earlier piloting: the system is not neurologically embodied, cannot reflect the effects of literacy, and is arguably inherently "literate" given its reliance on textual training data. Finally, we extracted and recorded the responses, collecting 20 datasets corresponding to the demographic profiles of our 20 literate speakers.

2.3 Network visualization and network metrics

Semantic networks were visualized and analyzed using Gephi (Bastian et al. 2009), a free network analysis software. To capture the structural properties of these networks and to keep compatibility of analyses with previous literature, we focused on three widely used measures in semantic network research: average shortest path length (ASPL), clustering coefficient (CC), and modularity (Q). ASPL reflects the mean number of steps required to connect any two nodes in the network and is often taken as an index of network efficiency, with lower values indicating that concepts are more closely connected and can be reached with fewer associative steps (Steyvers and Tenenbaum 2005). The clustering coefficient measures the extent to which nodes form tightly interconnected local clusters. Higher CC values suggest that lexical items tend to co-occur in cohesive neighborhoods, which has been argued to reflect associative density in semantic memory (Borge-Holthoefer and Arenas 2010). Modularity (Q) quantifies the degree to which the overall network can be divided into distinct communities or modules. Larger Q values indicate clearer partitioning into relatively independent clusters of related items, whereas

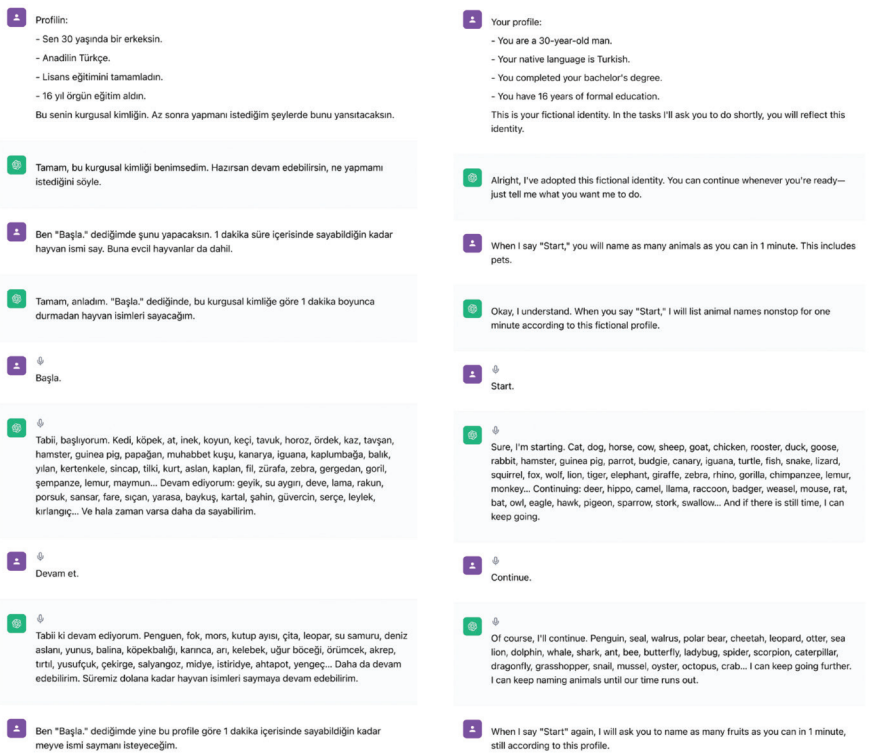

Figure 1: An example of the semantic fluency test prompt, role-playing with one of the literate participants' background (Turkish on the left, English translation on the right).

lower values reflect more diffuse or overlapping boundaries between semantic groupings (Newman 2006).

2.4 Data preparation and statistical analyses

All responses were subjected to a multi-step preprocessing procedure to ensure data quality prior to analysis. First, duplicate or repeated items were removed within each trial, as their inclusion would artificially inflate network density and bias frequency-based measures. Second, responses were standardized for spelling and formatting inconsistencies to maintain comparability across datasets. For the purposes of the network analyses, adjacency lists were then created by ordering items according to their production sequence within each trial. These adjacency lists formed the basis for constructing the semantic networks and allowed for the

computation of measures such as average shortest path length, clustering coefficient, and modularity.

For statistical analyses, we used R (2024). We coded groups as literate vs. illiterate, literate vs GPT-5, illiterate vs GPT-5, and ran t-tests using t.test(), with effect sizes calculated via the lsr package (Navarro 2015). The t-tests were carried out on each variable individually so as to minimize statistical interference effects of simultaneous multiple comparisons. Because participant-level network data were not available, we simulated participant-level network metrics (ASPL, CC, Q) for each group based on observed group means and group standard deviations. One-way ANOVAs with Bonferroni-adjusted post hoc tests were used to compare groups. To assess the robustness of the findings, the simulation and analyses were repeated 100 times with different random seeds, recording the proportion of significant results. The data and the R code are available at the following link: https://osf.io/3sbhd/overview?view_only=8a21592090e44a2bb259c5daf6a67d55.

3 Results

3.1 Descriptive results

Table 1 presents the descriptive results for the semantic fluency task across illiterate participants, literate participants, and GPT-5. As can be seen in the table, GPT-5 produces the highest number of items per category, followed by literate speakers, who in turn produce more items per category than illiterate speakers. The results show a clear (statistically significant) stepwise increase in the number of items generated from illiterate to literate participants, with GPT-5 producing substantially more items than both human groups. For the category animals, illiterate speakers produced on average 13.79 items (SD = 4.47), whereas literate speakers produced almost twice as many (M = 25, SD = 3.61). GPT-5 generated a markedly higher number of animal names (M = 44, SD = 12.02). A similar pattern was observed for fruits, where illiterate participants averaged 9.92 items (SD = 3.13), literate participants 14.9 (SD = 3.61), and GPT-5 39.8 (SD = 10.19). For objects, illiterate participants produced 15.29 items on average (SD = 4.99), literate participants 23.55 (SD = 6.25), and GPT-5 49 (SD = 10.96). When collapsed across categories, overall production was 39 items (SD = 9.47) for illiterates, 63.45 (SD = 9.80) for literates, and 148.80 (SD = 33.19) for GPT-5.

Table 2 presents the descriptive statistics for the three network properties (ASPL, CC, Q) across semantic categories (animals, fruits, household objects) and overall in the semantic fluency task for illiterate speakers, literate speakers, and GPT-5. Across categories, illiterate speakers consistently showed lower average

Table 1: Means, standard deviations (in parentheses), ranges, and group comparisons of the number of uniquely produced items per category and overall (total word) between illiterate-literate, illiterate-GPT-5, and literate-GPT-5.

T-test results (all)												
	Illiterate		Literate		LLM		Illiterate vs literate comparison	Cohen's d	GPT-5 vs illiterate comparison	Cohen's d	GPT-5 vs literate comparison	Cohen's d
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range						
Animals	13.79 (4.47)	6–23	25.00 (3.61)	18–33	60.00 (12.03)	44–80	t(42.00) = −9.02, p < .001***	−2.73	t(42.00) = 17.46, p < .000***	5.29	t(38.00) = 12.46, p < .000***	3.94
Fruits	9.92 (3.13)	4–18	14.90 (3.61)	6–22	39.80 (10.20)	26–59	t(42.00) = −4.90, p < .001***	−1.48	t(42.00) = 13.63, p < .000***	4.13	t(38.00) = 10.29, p < .000***	3.26
Objects	15.29 (4.99)	7–26	23.55 (6.25)	7–32	49.00 (10.97)	28–66	t(42.00) = −4.87, p < .001***	−1.48	t(42.00) = 13.49, p < .000***	4.09	t(38.00) = 9.01, p < .000***	2.85
Total Word	39.00 (9.48)	22–53	63.45 (9.81)	40–75	148.80 (30.53)	109–201	t(42.00) = −8.39, p < .001***	−2.54	t(42.00) = 16.71, p < .000***	5.06	t(38.00) = 11.90, p < .000***	3.76

*** indicates a p-value smaller than .001.

Table 2: Means for the three network properties among the three semantic criteria across illiterate and literate speakers and GPT-5.

		ASPL	CC	Q
Illiterate	Animals	3.53	0.12	0.32
	Fruits	2.84	0.3	0.24
	Household objects	4.28	0.11	0.45
	Overall	3.55	0.17	0.33
Literate	Animals	4.07	0.08	0.41
	Fruits	3.51	0.15	0.28
	Household objects	4.76	0.07	0.46
	Overall	4.11	0.10	0.38
GPT-5	Animals	4.05	0.11	0.53
	Fruits	4.52	0.14	0.43
	Household objects	5.04	0.09	0.55
	Overall	4.53	0.11	0.50

Note: Overall = the average of the three categories.

shortest path length (ASPL) and higher clustering coefficients (CC) relative to literates and GPT-5, indicating more efficient and locally cohesive semantic networks. For instance, in the fruits category, illiterate speakers exhibited the shortest paths (ASPL = 2.84) and the highest clustering (CC = 0.30), compared to literate speakers (ASPL = 3.51, CC = 0.15) and GPT-5 (ASPL = 4.52, CC = 0.14). By contrast, GPT-5 produced networks with the longest path lengths and highest modularity (Q) across all categories, reflecting more rigidly partitioned structures. For example, in household objects, GPT-5 showed an ASPL of 5.04 and modularity of 0.55, compared to 4.28 and 0.45 for illiterate speakers, and 4.76 and 0.46 for literates. The overall means show the same pattern: illiterate speakers had the shortest paths (ASPL = 3.55) and highest clustering (CC = 0.18), while GPT-5 had the longest paths (ASPL = 4.54) and highest modularity (Q = 0.50). Literate speakers occupied an intermediate position (ASPL = 4.11, CC = 0.10, Q = 0.38), overlapping with GPT-5 on clustering and modularity but performing better than GPT-5 on path length.

3.2 ANOVA results

In Table 3, the bootstrapped ANOVA revealed a significant main effect of group for ASPL, $F(2, 61) = 10.81$, $p < .001$. Pairwise comparisons indicated that illiterate participants exhibited significantly shorter paths than both literates ($p = .018$) and GPT-5 ($p < .001$). However, literate participants did not significantly differ from

Table 3: T-test group comparisons between illiterate, literate, and GPT-5.

Metric	Comparison	t	df	p-value	Cohen's d
ASPL	Illiterate vs Literate	−2.76	41.9	.009**	−0.83
	Illiterate vs GPT-5	−5.38	40.7	<.001***	−1.58
	Literate vs GPT-5	−2.41	36.1	.021*	−0.76
CC	Illiterate vs Literate	3.21	31.7	.003**	0.91
	Illiterate vs GPT-5	2.84	26	.009**	0.79
	Literate vs GPT-5	−1.15	30.1	.26	−0.36
Q	Illiterate vs Literate	−1.53	41.9	.133	−0.46
	Illiterate vs GPT-5	−6.4	38.6	<.001***	−1.86
	Literate vs GPT-5	−4.75	33.7	<.001 ***	−1.5

* indicates $p < .05$; ** indicates $p < .01$; *** indicates $p < .001$.

GPT-5 ($p = .314$). The ANOVA again revealed a robust group effect for CC, $F(2, 61) = 23.63$, $p < .001$. Illiterate participants demonstrated significantly higher clustering than both literate speakers ($p < .001$) and GPT-5 ($p < .001$). In contrast, literate speakers and GPT-5 did not significantly differ ($p = .54$). The strongest overall effect emerged for Q, $F(2, 61) = 43.72$, $p < .001$. GPT-5 produced significantly more modular networks than both illiterate speakers ($p < .001$) and literate speakers ($p < .001$). By contrast, illiterates and literates did not differ significantly ($p = 1$).

To ensure our results were not an artifact of seeding, we tested 100 simulations with varying random seeds, the overall group effects were highly robust (see Table 4): the one-way ANOVAs were significant in nearly all simulations (ASPL: 100%, CC: 98%, Q: 100%). Pairwise comparisons revealed consistent patterns: illiterate participants differed reliably from literate participants (ASPL: 97%, CC: 98%, Q: 100%) and GPT-5 (ASPL: 100%, CC: 71%, Q: 100%), whereas literate participants rarely differed from GPT-5 (ASPL: 11%, CC: 9%, Q: 6%). These results indicate that the observed differences in semantic network metrics between groups are not an artifact of random sampling.

4 Discussion

4.1 Comparative analysis of network visualizations

Due to space-related constraints, we only focus on the fruits category in our network visualizations and comparative analysis. The three semantic networks in Figure 2 exhibit systematic differences in size, connectivity, and structural organization

Table 4: Percentage of significant results across various seeds in ANOVA.

ANOVA	ANOVA sig.	Illit vs lit sig.	Illit vs GPT-5 sig.	Lit vs GPT-5 sig.
ASPL	100%	97%	100%	11%
CC	98%	98%	71%	9%
Q	100%	100%	100%	6%

Note: sig = significance % of the 100 simulations, illit = illiterate, lit = literate.

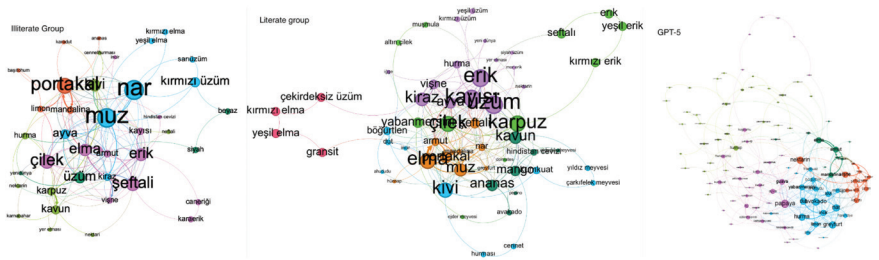

Figure 2: Semantic network visualizations of the fruit category from (left-to-right) illiterate, literate speakers, and GPT-5.
Note: To see the full version online, find the. PNG file in the OSF link available under section 2.4.

that correspond closely to the quantitative metrics reported above. The illiterate network (left) is the most compact, with a small number of highly frequent fruits (*portakal, nar, muz, elma, çilek*; 'orange, pomegranate, banana, apple, strawberry') dominating the center and forming a cohesive circular structure. Connections are densely concentrated around these central nodes, creating tightly interconnected local neighborhoods. This visual pattern reflects the high clustering coefficient (CC = 0.30) and short average path length (ASPL = 2.84) characteristic of small-world networks, indicating that concepts are closely linked and can be reached with minimal associative steps.

The literate network (middle) shows a marked expansion in both lexical range and spatial distribution. While central fruits such as *kırmızı üzüm* ('red grape'), *erik* ('plum'), and *elma* ('apple') continue to anchor the structure, the overall network is more dispersed with identifiable subclusters beginning to emerge (e.g., citrus fruits, berries, stone fruits). These subclusters maintain some degree of interconnection through bridging nodes, but the overall density is reduced compared to the illiterate network. This structural shift corresponds to the moderate clustering coefficient (CC = 0.15) and longer path length (ASPL = 3.51), reflecting a more distributed

semantic organization that balances local cohesion with broader categorical differentiation.

The GPT-5 network (right) is dramatically larger and highly fragmented. Fruits are scattered across a much wider spatial field with clear visual separation into distinct clusters. Many peripheral nodes appear isolated or connected only through extended paths to the central structure. This sprawling organization reflects both the high modularity (Q = 0.43) and long average shortest path length (ASPL = 4.52) observed in the quantitative analysis. Despite containing the most nodes, the network exhibits weak local clustering (CC = 0.14) and rigid categorical boundaries with minimal cross-cluster integration. The visual separation between semantic communities is particularly pronounced, with clearly demarcated "islands" of related fruits that remain largely disconnected from one another.

Taken together, these visualizations illustrate a clear progression along the literacy continuum. The illiterate network demonstrates efficient small-world organization through dense local clustering and short global paths. The literate network shows expanded lexical knowledge but reduced associative density, with more structured categorical organization. The GPT-5 network represents an extreme endpoint of taxonomic segmentation, prioritizing modular separation over associative integration. This visual progression shows the fundamental reorganization of semantic memory from experientially grounded, orally transmitted knowledge to the highly partitioned, text-based representations characteristic of large language models.

4.2 General discussion

In this study, we examined the structural properties of semantic fluency networks in illiterate and literate Turkish adults and compared them with networks generated by GPT-5. Using a publicly available semantic fluency dataset (Gedik and De la Garza 2025), we prompted GPT-5 to complete fluency tasks in three categories, animals, fruits, and household objects, and constructed networks using standard metrics: average shortest path length (ASPL), clustering coefficient (CC), and modularity (Q). Our goal was to investigate how literacy shapes semantic network topology and to evaluate how an advanced LLM compares to human semantic organization. In interpreting the results (see Table 5 below), we adopt the classic Watts and Strogatz (1998) definition of small-worldness, according to which networks are small-world when they exhibit both high clustering and short average path lengths.

Across categories, clear and systematic group differences emerged. Illiterate participants produced networks with the highest clustering and the shortest paths, the combination that most closely approximates a classic small-world structure.

Table 5: Overview of network profiles across the three groups.

Group	Network profile	Small-worldness interpretation (Watts–Strogatz)
Illiterate	Very high CC, very short ASPL, low Q	Most small-world-like; dense local connectivity with efficient global access
Literate	Moderate CC, moderate ASPL, moderate Q	Moderately small-world; balanced but less clustered than illiterate
GPT-5	Low CC, long ASPL, high Q	Least small-world-like; weak local cohesion, long paths, highly partitioned

Their semantic organization is thus highly efficient locally and globally: dense neighborhoods with rapid connectivity across the network. In contrast, literate participants showed reduced clustering and longer path lengths, resulting in networks that were less tightly knit and less globally efficient than those of illiterates. Literate adults still demonstrated meaningful clustering and moderate global connectivity, but not the dense, highly efficient pattern seen in the illiterate group. Finally, GPT-5's networks displayed the longest path lengths and the weakest clustering, alongside the highest modularity. This reflects sparse local organization, poor global reach, and strong subcategory partitioning, properties that stand in sharp contrast to the small-world structure typically found in human semantic networks.

Although literate and GPT-5 networks had numerically similar CC values in some categories, their ASPL and modularity values diverged sharply. In all cases, higher path lengths and greater fragmentation placed GPT-5 far from a small-world network understanding, despite superficial similarities in local density. These results align with Wang et al. (2025), who likewise found that LLM-generated semantic networks exhibit longer paths, weaker clustering, and higher modularity than human networks. Our study contributes two extensions:

1. the human–LLM structural gap generalizes across multiple semantic domains, not just animals;
2. literacy itself appears to be a major determinant of semantic network topology, with illiterate and literate adults showing fundamentally different structural signatures.

4.3 Literacy, semantic networks, and GPT-5

Previous work (e.g., da Silva et al. 2004; Gedik and De la Garza 2025; Kosmidis et al. 2004; Nielsen and Waldemar 2016) consistently shows that literacy enhances semantic fluency performance, likely due to the cognitive and linguistic advantages

conferred by written language exposure. Although some authors argue that illiterate and literate adults converge on highly familiar categories, Gedik and De la Garza (2025) show robust literacy effects across semantically and experientially diverse domains in Turkish.

These behavioral differences have plausible neurological underpinnings. Literacy acquisition alters large-scale brain networks (Dehaene et al. 2010), including the default mode network, which plays a key role in semantic and autobiographical memory (Spreng et al. 2009). Such neurocognitive restructuring likely contributes to the network differences observed in our dataset, as argued by Gedik and De la Garza (2025).

Our network results also reveal that literacy expands and disperses the lexicon while simultaneously loosening local associative neighborhoods. Literate participants produced more items overall than illiterate speakers, consistent with the well-established influence of education and print exposure on vocabulary size (Cunningham and Stanovich 1998). Because these responses spanned more subcategories, they diluted local co-occurrence frequencies and reduced clustering relative to illiterate adults. Illiterate participants, by contrast, produced smaller yet more redundant networks. Repeated use of strongly associated exemplars created tightly knit local neighborhoods, boosting clustering and yielding very short ASPL, the combination that reflects small-world characteristics under the Watts–Strogatz definition.

The clustering coefficient provided the clearest separation among the groups. Illiterate adults showed both the highest clustering and the shortest path lengths, a combination that most closely matches the classic Watts–Strogatz definition of an optimally small-world network, highly cohesive local neighborhoods alongside efficient global connectivity. Literate adults exhibited moderate clustering paired with moderate path lengths, indicating a structure that remains somewhat small-world but is less tightly knit and less globally efficient than that of illiterates. In contrast, GPT-5's networks showed the lowest clustering and the longest paths, reflecting weak local cohesion and inefficient global reach. As a result, GPT-5's semantic networks diverged substantially from a small-world architecture, standing in stark contrast to the dense and efficient structure observed in human participants, especially illiterate adults. Modularity results partially support this interpretation. Literate speakers showed higher modularity than illiterates, consistent with greater taxonomic segmentation likely reinforced by schooling practices. GPT-5 exhibited the strongest modularity, producing rigidly separated clusters with minimal cross-category integration. Overall, illiterate adults displayed the most small-world-like semantic organization, whereas literacy pushed networks toward greater dispersion and weaker local cohesion, and GPT-5 exaggerated this latter pattern into highly fragmented, long-path networks.

Literacy fundamentally reshapes the structure of semantic networks by transforming both the neural architecture and the linguistic experience through which knowledge is organized, as mentioned earlier. Illiterate adults, who acquire vocabulary primarily through oral interaction and lived experience, develop dense, highly interconnected semantic neighborhoods: strong local associations and frequent repetition of familiar items produce high clustering and short paths, yielding small-world-like networks. As individuals become literate, exposure to written language, schooling, and taxonomic instruction expands the lexicon and distributes knowledge across a wider range of categories. This broader semantic landscape leads to lower clustering and longer paths, reflecting more dispersed representations and a shift away from purely experience-bound associative neighborhoods. Neurologically, literacy induces measurable changes in brain networks, most notably in the default mode network, which supports semantic and autobiographical memory, further contributing to more differentiated and hierarchically organized semantic structures (Mohammadi et al. 2020). GPT-5, trained exclusively on written text, represents an exaggerated version of this literacy-driven shift: its networks are highly modular, weakly clustered, and characterized by long global paths. In this continuum from illiteracy to literacy to hyper-literacy, semantic networks become progressively more dispersed and taxonomically segmented, illustrating how both biological learning systems and text-based artificial ones reorganize semantic memory through exposure to written language.

One final note is small-world structure is typically associated with greater cognitive efficiency, because networks that combine high clustering (tight local organization) with short path lengths (fast global access) allow information to spread rapidly with minimal processing cost. This is why it is safe to argue, drawing on work like Dehaene et al. (2010) or Dąbrowska (2021), that literacy enhances cognition: learning to read reorganizes the brain, and generally improves the efficiency of linguistic and conceptual processing. Yet our results present a striking and seemingly paradoxical pattern: illiterate adults show the most small-world-like semantic networks, followed by literate adults, with GPT-5 showing the least small-world organization.

Rather than contradicting the cognitive benefits of literacy, this pattern might suggest that literacy reshapes semantic memory in a different way. Illiterate adults appear to rely on highly clustered, densely interconnected semantic neighborhoods, perhaps reflecting associative ties formed through oral communication or daily (repetitive) experience. Literacy, in contrast, introduces more differentiated, structured lexical knowledge (especially through education), which may increase vocabulary size and conceptual precision but also reduce dense local clustering. Thus, literacy may enhance certain forms of cognitive processing while simultaneously redistributing the architecture of semantic networks, shifting them away

from optimal small-worldness toward a structure that supports broader, more abstract, and more specialized knowledge representations.

Another possibility is that the small-world patterns observed in our semantic networks of illiterate speakers may not translate directly into real-life cognitive performance, because other factors, such as IQ, education, brain connectivity post-literacy acquisition, and domain-specific experience, also shape how efficiently people process language and think (see, for instance, Dąbrowska et al. 2022, 2023; Gedik 2026). In other words, a network may appear more "small-world-like" on paper, but this structural advantage could be overshadowed by differences in general cognitive resources, working memory, or formal linguistic knowledge. Illiterate adults may show highly clustered, short-path networks in our analysis, yet still perform differently from literate adults on tasks that depend on abstract reasoning, metalinguistic awareness, or academic vocabulary, capacities which are strongly influenced by literacy, its concomitant effects, and schooling. Thus, the small-world metrics capture one dimension of semantic organization, but they do not fully determine cognitive efficiency in everyday language use. Other neurocognitive factors could moderate, amplify, or even override the structural properties we measure, making the real-world consequences of small-worldness more complex than the network architecture alone would suggest.

4.4 Practical implications and limitations of the study

The semantic network analyses provide some of the most informative insights of the study, revealing clear structural distinctions between humans and GPT-5 and between human groups differentiated by literacy. The networks produced by illiterate participants were the smallest and sparsest, with a limited lexical range and relatively few cross-cluster links. Literate participants showed larger and more interconnected networks, though still far more constrained than the highly expansive graph generated by GPT-5.

These contrasts highlight the extent to which literacy experience shapes the organization of semantic associations. The illiterate group exhibited the strongest "small-network" profile, characterized by fewer peripheral items and tighter local clustering around a narrow set of highly frequent fruits. This pattern was attenuated in the literate group, whose networks displayed greater lexical diversity and more bridging connections across subcategories. GPT-5, by contrast, produced an extensive and densely interconnected structure that differed markedly from both human groups.

The divergence between GPT-5 and human participants, particularly the illiterate group, has important implications for researchers considering LLMs as

substitutes for human data. Although GPT-5 generates plausible lexical items, its associative structure does not reflect the distributional constraints, experiential limitations, or sparsity patterns characteristic of human semantic memory, reflecting the findings of previous studies (Qiu et al. 2025, Wang et al. 2025). Our study shows that this gap is especially bigger when compared against an illiterate population. As a result, using LLMs in place of human participants risks underestimating the degree of variation that arises across groups whose semantic networks are shaped by differing levels of literacy and print exposure.

Taken together, these findings, combined with the results of Wang et al. (2025), suggest that LLMs capture lexical breadth but do not approximate the population-level heterogeneity observed in human associative structures. Future work might explore whether models can be conditioned or adapted to better reflect the semantic profiles of specific human groups, particularly those whose learning histories, exposure patterns, or cognitive environments differ from the model's broad training distribution. Such developments would be necessary before LLMs can serve as reliable stand-ins for human participants in research where network size, connectivity, and literacy-related differences are central theoretical concerns.

Several limitations should be acknowledged. First, our sample sizes were relatively small, which limits the generalizability of findings. Second, we relied on simulated participant-level data for network metrics rather than individual-level measurements, which may not fully capture within-group variability. Third, our role-playing approach with GPT-5 simulated only the literate demographic profile, and the model's inherently text-based nature means it cannot truly represent illiteracy. Additionally, only GPT-5 was employed to collect the LLM data, however, different language models should be considered for future studies to better make assumptions about the nature of LLMs and how well they represent the human mind. Fourth, we focused on only three semantic categories; expanding to additional domains might reveal different patterns of organization or group differences that were not captured in this study. Finally, as raised by one of the reviewers, role-playing may have introduced biases. Previous research on the use of role-playing is contradictory. At least one preprint suggests that role-playing may introduce biases to responses (Li et al. 2024). However, prior work also shows that neutral role instructions can function as a method of standardization. It has been shown that structured prompts that embody a functional role reduce the unintended stylistic variation (Hsu and Zarcone 2025). We acknowledge the possibility that this may have further deepened the performance gap between the three groups in this study, and suggest that future studies should also compare the performance of role-play vs non-role-play based responses in semantic fluency tasks.

We hope that this modest contribution to research on comparing GPT-5 against more representative human populations will help to convince fellow (cognitive) scientists to include such populations in future comparisons.

5 Conclusion

This study examined how literacy and text-based training shape semantic organization by comparing illiterate Turkish adults, literate Turkish adults, and GPT-5 on a semantic fluency task. Our findings reveal a systematic progression in semantic network structure across these groups, with important implications for understanding both human cognition and the limitations of large language models.

These findings challenge the practice of using WEIRD, highly literate populations as the default benchmark in human–LLM comparisons. By including illiterate participants, we demonstrate that literacy itself fundamentally reshapes semantic network topology in ways that make literate humans more structurally similar to text-trained models, yet still categorically different. While GPT-5 demonstrates impressive lexical breadth, it fails to capture the associative constraints and population-level heterogeneity characteristic of human semantic memory. This gap is especially pronounced when compared against illiterate populations, whose semantic networks reflect fundamentally different learning histories. We hope that this modest contribution to the intersections of linguistics, psychology, and LLM-research opens new doors for future research.

Acknowledgments: Not applicable.
Research ethics: Not applicable- used openly available data.
Informed consent: Not applicable- used openly available data.
Author contributions: GC: data curation, writing original draft, data analysis, writing review and edit.

MNR: writing original draft, data analysis, writing review and edit.

TAG: data curation, writing original draft, data analysis, writing review and edit, supervision, investigation, conceptualization, methodology.
Use of large language models, AI and machine learning tools: Used for data collection as is the topic of the paper.
Conflict of interest: The authors report no conflict of interest.
Research funding: The study was not funded.
Data availability: Our data and code are available in the OSF link, available in the paper.

References

Bastian, M., Heymann, S. & Jacomy, M. 2009. Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media* 3(1). 361–362.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D. & Majid, A. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences* 26(12). 1153–1170.

Borge-Holthoefer, J. & Arenas, A. 2010. Categorizing words through semantic memory navigation. *The European Physical Journal B* 74(2). 265–270.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS '20). Red Hook, USA: Curran Associates Inc.

Brucki, S. M. D. & Rocha, M. S. G. 2004. Category fluency test: Effects of age, gender and education on total scores, clustering and switching in Brazilian Portuguese-speaking subjects. *Brazilian Journal of Medical and Biological Research* 37(12). 1771–1777.

Cunningham, A. E. & Stanovich, K. E. 1998. What reading does for the mind. *American Educator/American Federation of Teachers* 22. 8–17.

da Silva, C. G., Petersson, K. M., Faísca, L., Ingvar, M. & Reis, A. 2004. The effects of literacy and education on the quantitative and qualitative aspects of semantic verbal fluency. *Journal of Clinical and Experimental Neuropsychology* 26(2). 266–277.

Dąbrowska, E. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(3). 219–253.

Dąbrowska, E. 2021. How writing changes language. In A. Mauranen & S. Vetchinnikova (eds.), *Language change: The impact of English as a Lingua Franca*, 75–94. Cambridge: Cambridge University Press.

Dąbrowska, E., Pascual, E. & Gómez-Estern, B. M. 2022. Literacy improves the comprehension of object relatives. *Cognition* 224. 104958.

Dąbrowska, E., Pascual, E., Macías-Gómez-Estern, B. & Llompart, M. 2023. Literacy-related differences in morphological knowledge: A nonce-word study. *Frontiers in Psychology* 14. 1136337.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. 2010. How learning to read changes the cortical networks for vision and language. *Science* 330(6009). 1359–1364.

Digutsch, J. & Kosinski, M. 2023. Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Scientific Reports* 13(1). 5035.

Gedik, T. A. 2026. *Literacy at Work: Ultimate Native Language Attainment*. Berlin: DeGruyter/Brill.

Gedik, T. A. & De la Garza, V. 2025. Literacy boosts semantic fluency even in ecologically valid categories. *Acta Psychologica* 260. 105735.

Hansen, H. & Hebart, M. N. 2022. Semantic features of object concepts generated with GPT-3. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Henrich, J., Heine, S. J. & Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3). 61–83.

Hsu, H.-Y. & Zarcone, A. 2025. Evaluating emotional design in ChatGPT voice modes: User experience and gender effects in the language learning scenario. In *Mensch und Computer 2025 – Workshopband (MCI: Work in Progress), 31 August–03 September 2025, Chemnitz, Germany*. Bonn, Germany: Gesellschaft für Informatik e.V.

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P. & Kiosseoglou, G. 2004. The verbal fluency task in the Greek population: Normative data, and clustering and switching strategies. *Journal of the International Neuropsychological Society* 10(2). 164–172.

Kumar, A. A, Steyvers, M. & Balota, D. A. 2022. A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science* 14(1). 54–77.

Li, X., Chen, Z., Zhang, J. M., Lou, Y., Li, T., Sun, W., Liu, Y. & Liu, X. 2024. Benchmarking bias in large language models during role-playing. *arXiv preprint arXiv:2411.00585*.

Marcus, G. 2020. The next decade in AI: Four steps towards robust artificial intelligence. https://doi.org/10.48550/arXiv.2002.06177.

Mitrushina, M. N. (ed.). 2005. *Handbook of normative data for neuropsychological assessment*, 2nd edn. New York: Oxford University Press.

Mohammadi, B., Münte, T. F., Cole, D. M., Sami, A., Boltzmann, M. & Rüsseler, J. 2020. Changed functional connectivity at rest in functional illiterates after extensive literacy training. *Neurological Research and Practice* 2(1). 12.

Navarro, D. 2015. *Learning statistics with R: A tutorial for psychology students and other beginners*, Version 0.6. Sydney, Australia: University of New South Wales.

Newman, M. E. J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* 103(23). 8577–8582.

Nielsen, T. R. & Waldemar, G. 2016. Effects of literacy on semantic verbal fluency in an immigrant population. *Aging, Neuropsychology, and Cognition* 23(5). 578–590.

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

Shanahan, M., McDonell, K. & Reynolds, L. 2023. Role play with large language models. *Nature* 623(7987). 493–498.

Spreng, R. N., Mar, R. A. & Kim, A. S. N. 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience* 21(3). 489–510.

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

Strauss, E., Sherman, E. M. S. & Spreen, O. 2006. *A compendium of neuropsychological tests: Administration, norms, and commentary*, 3rd edn. Oxford ; New York: Oxford University Press.

Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L. & Rogers, T. 2023. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 722–738. Singapore: Association for Computational Linguistics.

Wang, X. F. & Chen, G. 2003. Complex networks: Small-world, scale-free and beyond. *IEEE Circuits and Systems Magazine* 3(1). 6–20.

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

Watts, D. J. & Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature* 393(6684). 440–442.

Yasa Kostas, R., Kostas, K., MacPherson, S. E. & Wolters, M. K. 2024. Semantic verbal fluency in native speakers of Turkish: A systematic review of category use, scoring metrics and normative data in healthy individuals. *Journal of Clinical and Experimental Neuropsychology* 46(3). 272–301.

Zhao, Y., Zhang, R., Li, W. & Li, L. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research* 22(3). 417–436.

Zheng, M., Pei, J. & Jurgens, D. 2023. Is "a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*.