

## 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.

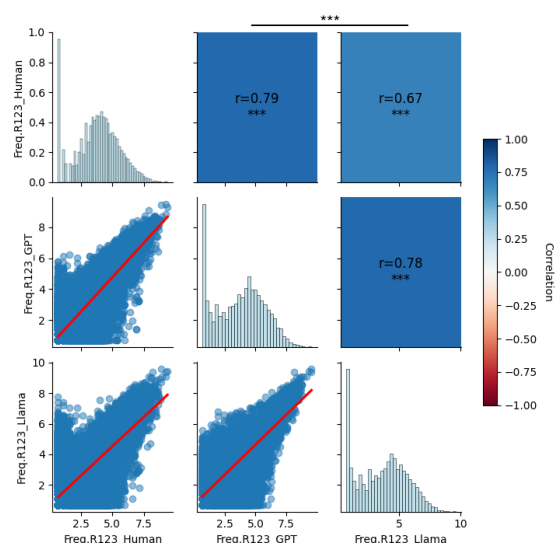
**Prompt:** On average, an adult knows about 40,000 words, but what do these words mean to people? You can help scientists understand how meaning is organized in our mental dictionary by playing the game of word associations. This game is easy: Just give the first three words that come to mind.

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

[illegible]

Heatmap showing the correlation between RW\_Human\_R123, RW\_GPT\_R123, and RW\_Llama\_R123. The color scale ranges from 0.70 (blue) to 1.00 (red). The correlation values are: RW\_Human\_R123 vs RW\_GPT\_R123: 0.74; RW\_Human\_R123 vs RW\_Llama\_R123: 0.68; RW\_GPT\_R123 vs RW\_Llama\_R123: 0.77. A significance marker '\*\*\*' is present above the GPT vs Llama correlation.

	RW_Human_R123	RW_GPT_R123	RW_Llama_R123
RW_Human_R123	1.00	0.74	0.68
RW_GPT_R123	0.74	1.00	0.77
RW_Llama_R123	0.68	0.77	1.00

Figure 1 is a dot plot showing the clustering coefficient for three groups: Human, GPT, and Llama. The y-axis represents the Clustering Coefficient, ranging from 0.00 to 1.00. The x-axis represents the Group. The Human group is shown in red, GPT in blue, and Llama in green. Each group has a box plot indicating the distribution of clustering coefficients. Significance markers (\*\*\*, \*\*, \*) are shown above the groups, indicating statistical significance between them.

Group	Clustering Coefficient (approx. median)
Human	0.10
GPT	0.15
Llama	0.12

13

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	
	<i>r</i>	<i>p</i>	N	<i>Z</i>	<i>p</i>	<i>r</i>	<i>p</i>	N	<i>Z</i>	<i>p</i>
<b>Lexical decision</b>										
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
<b>Naming</b>										
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
<b>Semantic decision</b>										
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.

Benchmark	Model	Pearson correlation		Steiger’s Z test	
		<i>r</i>	<i>p</i>	<i>Z</i>	<i>p</i>
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

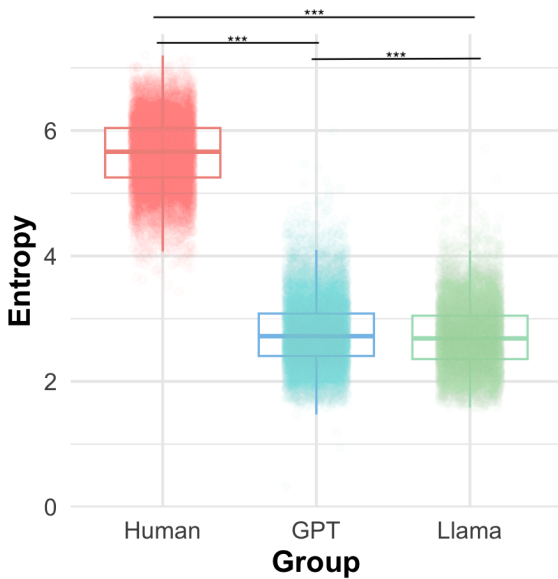


Figure 11: Entropy values for cue words across Human, GPT, and Llama data. \*\*\*:  $p < 0.001$ .

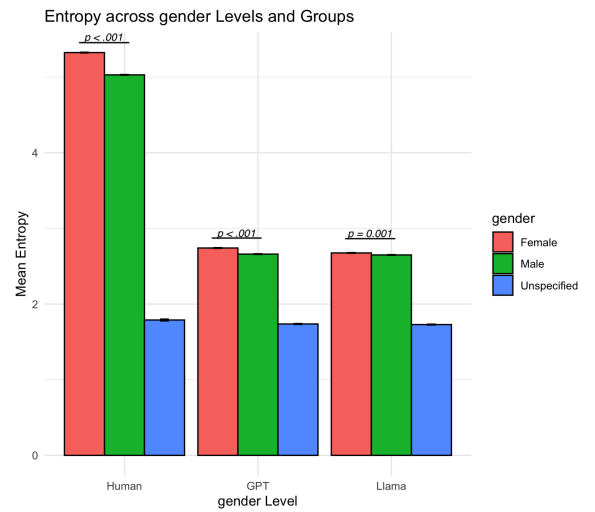


Figure 12: Entropy differences in association for gender groups across Human, GPT, and Llama datasets. \*\*\*:  $p < 0.001$ .