Semantic Category's Effects on the Human-Likeness of LLM Word Associations

Habbie Dem
University of Toronto Mississauga

INTRODUCTION

Semantic memory, or "memory for word meanings, facts, concepts, and general world knowledge" (Jones et al., 2015), is as central to human language production as it is elusive. One of the foremost methods of investigating it are **free association tasks** - given a word, what is the first thing to come to mind? The answers to such prompts allow the creation of **semantic networks**, which can elucidate psycholinguistic phenomena through network-spreading experiments. But what can be shown through attempting such methods on LLMs? That is what the nascent field of **machine psychology** seeks to answer.

RESEARCH QUESTION

How do the semantic associations of three LLMs (Mistral, Llama3 and Claude Haiku) differ from those of humans, particularly by semantic category?

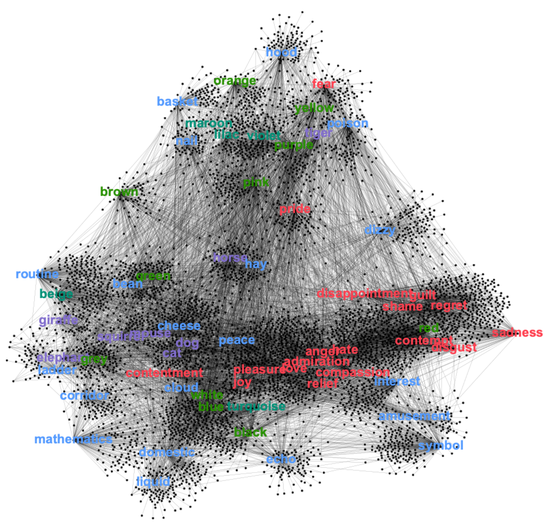

Figure 1: Semantic network constructed from human answers.

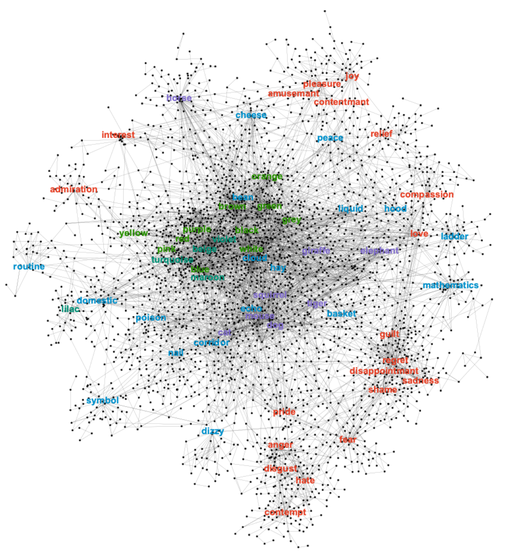

Figure 2: Semantic network constructed from Llama3's answers.

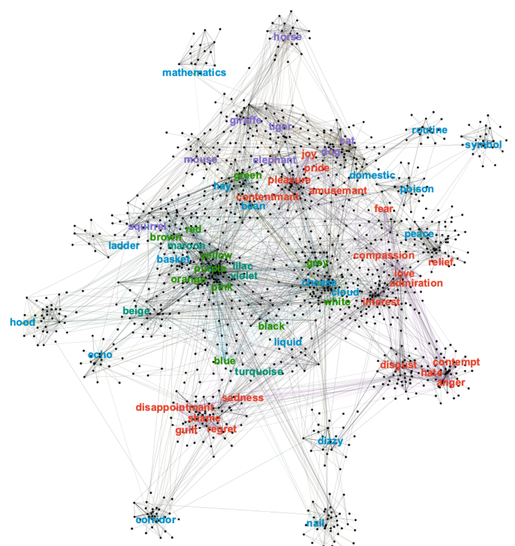

Figure 3: Semantic network constructed from Mistral's answers.

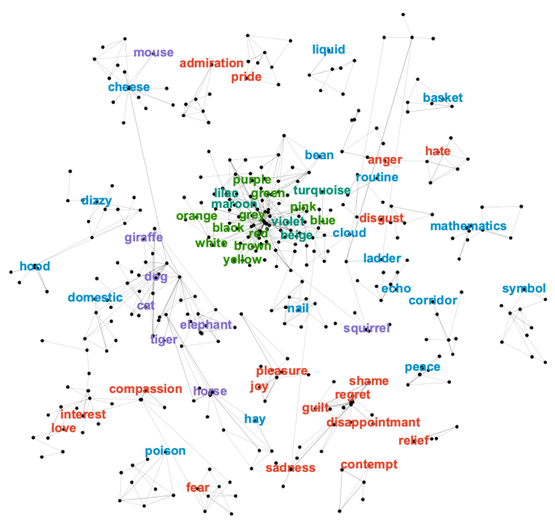

Figure 4: Semantic network constructed from Haiku's answers.

Figure 5: A grouped bar chart featuring the average JSD for each model in each semantic category.

METHODS

DATA: LLM data was taken from the **"LLM World of Words" (LWOW)**, a large-scale English LLM free association database (Abramski et al., 2025). Although the LWOW contains over 12 000 cue words, this study focuses on the 62 from **Jonauskaite et al.'s 2025 human-generated free associations database**, the English answers of which make up this study's human data.

ANALYSES: The first analysis focuses on the reweighed **Jensen-Shannon divergence (JSD)** between the probability distribution of a model's answers and that of the human responses. The second analysis explores **five distinct properties of each semantic network**, namely sparsity, connectedness, path-lengths, neighbourhood clustering and degree distributions.

RESULTS

JSD: The reweighed average JSD between humans and all models indicates a clear hierarchy in human-likeness, with **Mistral taking the lead (0.60), followed by Haiku (0.64) then Llama3 (0.77)**. An analysis by semantic category further reveals the order in which they affect human-likeness, from most to least human-like: **Animals, Basic Colours, Emotions, and Non-Basic Colours**.

NETWORKS: The network analysis, done by running a linear mixed model for each property under investigation, revealed that **Llama3 had the most human-like network, followed by Mistral and Haiku**. Furthermore, the **Basic Colours** category resulted in the most human-like networks, followed by **Animals, Non-Basic Colours, and Emotions**.

CONCLUSIONS

LLMs struggle the most with recreating human semantic associations for **abstract concepts**. Moreover, their production of divergent semantic associations **does not imply similarly divergent semantic networks**.

ACKNOWLEDGMENTS

Many thanks to Professors Barend Beekhuizen and Mai Ha Vu for their invaluable mentorship and assistance.

REFERENCES

Abramski et al. (2025). *Scientific Data, 12*(1), 803. https://doi.org/qvng

Jonauskaite et al. (2025). *Journal of Open Psychology Data, 13*(1), 4. https://doi.org/qvnh

Jones et al. (2015). *The Oxford Handbook of Computational and Mathematical Psychology (p. 232–254)*. https://doi.org/d8vp

Contact: habbie.dem@mail.utoronto.ca | habbie321.github.io