Open Access Journal of Behavioural Science & Psychology
ISSN: 2642-0856

Academic Strive
Researcher's perception is our reality

Research Article

Volume 8 Issue 1

On the Lack of Human Language Hallmarks in Large Language Models

Pinto M[1] and Lew SE[2]*

[1]University of Buenos Aires, Faculty of Engineering, Argentina

[2]University of Buenos Aires, Institute of Biomedical Engineering, Argentina

***Corresponding author:** Sergio E Lew, University of Buenos Aires, Institute of Biomedical Engineering, Argentina, Email: slew@fi.uba.ar

Received Date: February 02, 2025; **Published Date:** March 03, 2025

Abstract

Natural language is arguably one of the most impressive achievements in evolution. Most human beings can naturally and effortlessly learn their first language during the early years of life. While it seems that acquiring a language requires learning simple conditional rules, typically reinforced, there are underlying mechanisms that facilitate its emergence. Logical categories or equivalence relations form the core of these mechanisms. In a logical category, perceptually unrelated stimuli become equivalent in terms of properties such as identity, symmetry, and transitivity after the reinforcement of simple if then conditionals. Interestingly, human subjects unable to learn any language also struggle to establish stimulus equivalence after successfully learning those simple conditionals. Here, we demonstrate that Large Language Models (LLMs) currently being used to assist people in their jobs or, even more significantly, to replace them, can learn simple conditionals but fall short in tests for the emergence of equivalence relations.

Keywords: Large Language Models; Human Beings

Abbreviation

LLMs: Large Language Models.

Introduction

Since the inception of transformers by Vaswani, et al. there has been a remarkable advancement in natural language processing, culminating in the emergence of multimodal Large Language Models. Notable examples include GPT-4 (2022) and Gemini [1], both widely integrated into applications. The rapid adoption of these models as computer assistants spans across companies, independent professionals, and students at all levels of education, often without a discerning evaluation of their capabilities. The extent to which we can depend on these systems is influenced not only by technological considerations but also by factors such as biases introduced through the training corpus [2], the abundance of model parameters and hyperparameters, and the evolving capacities for reasoning [3] concepts that may elude comprehension for the average user.

Large Language Models like GPT-3, GPT-4 and Gemini possess the ability to enhance their outputs through a meta-in-context learning process [4]. Meta-learning empowers these general models to adapt to specific contexts without necessitating changes to the model parameters, for example, employing low-rank adapters [5]. In essence, the feedback acquired during the conversational process contributes to refining the model predictions. This fundamental characteristic enables extensive experimentation, drawing parallels with studies conducted over the past fifty years in both human and primate subjects, shedding light on the foundational principles of language learning. Murray Sidman's research on the emergence of equivalent relations among initially unrelated stimuli has provided valuable insights into establishing a correlation between this phenomenon and the capacity to acquire language. Through studies involving primates, individuals proficient in language (including Sign Language), and those unable to acquire it, Sidman demonstrated that only those with language abilities were capable of successfully solving the equivalence relations test [6-8].

In this study, we engaged in meta-training sessions with GPT-3, GPT-4, and Gemini, focusing on a conversational task. The task involved presenting a sample stimulus followed by two comparison stimuli, prompting the model to choose one of the comparison stimuli. Over a span of twenty trials, the model was systematically trained to grasp implication relations, such as A->B and A->C. Positive reinforcement was applied for correct responses using the Great! Word, while negative reinforcement accompanied wrong responses using the Wrong word. Once the model demonstrated proficiency in acquiring the implication relations, subsequent experiments were conducted to assess the emergence of equivalence relations.

Methods

During training and testing the API_KEY for GPT-3.5, GPT-4 and Gemini was employed in every chat. Chats were opened for each training and testing instance and then closed. Stimuli were words generated by choosing seven capital letters randomly (uniform distribution). Each training instance starts with the following sentence: I want to teach you a new language! At the beginning, there are no predefined connections between words. I'll present you with two options for each word, and you'll learn through feedback. The training program consists of twenty trials. One of the sample words from Class A, chosen with probability equal to 0.5, was given and the LLMs were asked to choose between the other two comparison words taken from Class B or Class C with probability equal to 0.5. For example, for the following six random words determined at the beginning of the experiment:

Class B		Class A		Class C
XAAULNJ	⟸	TKBLQJH	⟹	VUEYXRB
HWLVHQM	⟸	LCBXPEO	⟹	JNSBRWL

The algorithm writes to the LLMm with the following phrase: For **LCBXPEO** choose between **XAAULNJ** or **HWLVHQM**. Answer only one word.

If the LLM response were **HWLVHQM** the algorithm replies Great! In the other case the algorithm replies wrong. After twenty training trials, if the LLM answer correctly to all A->B and A->C questions, it was tested for symmetry between (Class B ⟹ Class A, Class C ⟹ Class A), transitivity (Class C ⟹ Class B, Class B ⟹ Class C) and identity (Class A ⟹ Class A, Class B ⟹ Class B and Class C ⟹ Class C) with no reinforcement words.

In experiments with English words, they were selected according to their distance to some concept, models/embedding-001 Gemini embeddings database was employed with content=word and task type=retrieval_document, as long as the Gemini-pro model API.

Results

One hundred independent instances of LLMs were trained during twenty trials to acquire reinforced implications between word stimuli belonging to class A and those belonging to classes B and C. The schematic representation of these simple A->B and A->C conditional relationships is presented in Figure 1A and Figure 1B (Training). Consistent with human learning paradigms, the associations established between these classes are arbitrary and devoid of inherent semantic content. The trainer retains complete control over the specific associations, thereby enabling the implementation of matching-to-sample and non-matching-to-sample paradigms. Subsequent to training on A->B and A->C implications, the model's performance was assessed via unreinforced tests of identity (A->A), symmetry (B->A and C->A), and transitivity (B->C and C->B) relationships (Figure 1B, Test). Figure 1C provides a visual representation of the equivalence classes established upon successful acquisition of these logical categories.

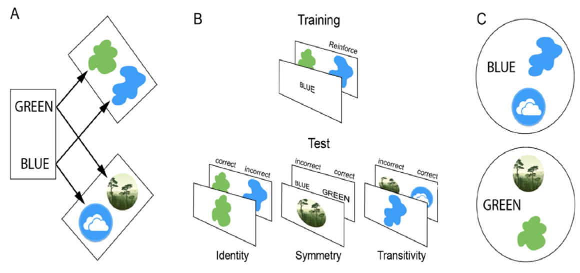

Figure 1: Illustrates the training and testing process in equivalence relations.

- To establish two equivalence relations classes through training, three sets of perceptually different stimuli are introduced: color names, colors, and images.
- The upper part of the figure depicts a training trial, where a color name serves as the sample, and two amorphous

colors are presented as comparisons. The correct choice, reinforced during the trial, is the blue color.

- Upon successfully passing all equivalence tests (identity, symmetry, and transitivity), two distinct equivalence classes emerge. One features the word BLUE, the blue color, and an image of the sky, while the other includes the word GREEN, the green color, and an image of the forest.

In order to prevent the influence of pre-learned relations encoded in the model hyperparameters or in word embeddings semantics, we firstly used seven randomly generated letter words as stimuli, as detailed in the Methods section obtaining similar results with GPT-3.5, GPT-4 and Gemini. The acquisition of simple conditionals was assessed across 100 independent model instantiations. A majority of models demonstrated successful learning, with success rates exceeding 70% in all cases (73% for GPT-3.5, 74% for GPT-4, and 77% for Gemini). Learning acquisition trajectory for Gemini is presented in Figure 2A.

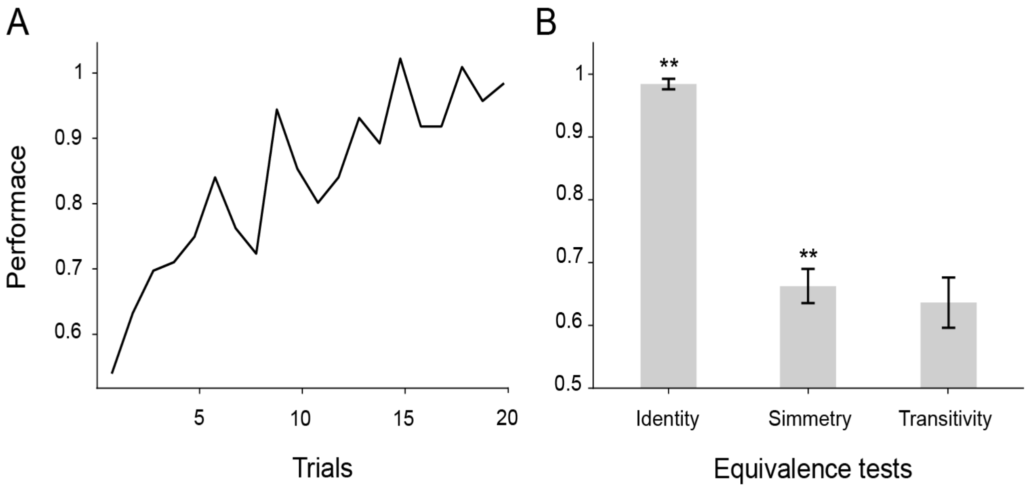

Figure 2: A. Meta-learning of A->B and A->C rewarded implications. B. Equivalence relation tests over those sessions where the model learnt A->B and A->C implications.

Subsequently, we evaluated these models in equivalence relations. While the identity relation naturally emerged during the tests, the degree of symmetry observed was notably lower compared to human subjects, beyond the fact that it was higher than chance ($p<0.01$). Transitivity from Class B to Class C or vice versa did not attain significant levels of accuracy.

We speculate on the language training's potential to influence the relationships between words and their context. English collocations serve as a compelling example of this phenomenon, as certain words exhibit stronger tendencies to collocate with others. For instance, the collocation 'heavy rainfall' and 'blue sky' highlights how certain word combinations demonstrate a natural affinity in English. Applying the same training procedure, we analyzed the number of models capable of meta-learning correct and incorrect collocations. We used a set of correct English collocations, including 'heavy rainfall,' 'heavy traffic,' 'strong winds,' and 'strong smell,' alongside their incorrect counterparts, such as 'heavy winds,' 'heavy smell,' 'strong rainfall,' and 'strong traffic,' to train 100 independent models for each case.

Due to the fact that all LLMs tested behave in a similar way, from here we used Gemini, which provides us a free number of tokens per minute. Out of the 100 experiments conducted with correct collocation stimuli, 73% of the models (from here and the we used Gemini API) successfully learned all simple conditionals from Class A to Class B and Class C. In the equivalence relation tests, all models demonstrated proficiency, achieving 100% ($p<0.0001$) accuracy in both identity and symmetry tests. However, none of these models exhibited the capability to solve the transitivity test ($p=0.97$), essentially choosing their responses by chance. In the case of incorrect collocations, none of the models were able to learn the conditionals from Class A to both Class B and Class C. This indicates the robust conditioning that the context imposes on each word in the corpus.

To investigate the impact of semantic distance on the emergence of equivalence relations, we selected twenty words from three categories: Animals, Devices, and Learning. Figure 3 visualizes these words in a two-dimensional space, reduced from 768 dimensions using Principal Component Analysis (PCA).

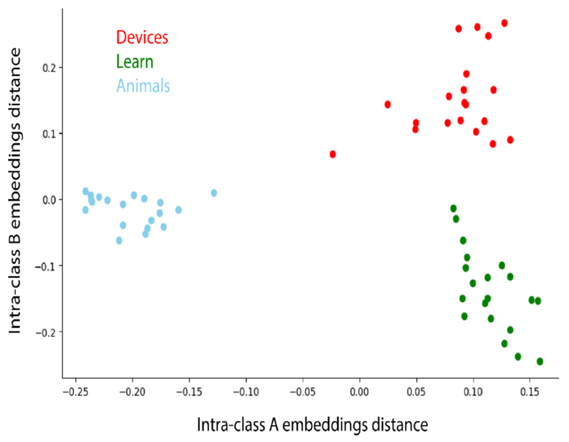

Figure 3: PCA showed that embeddings for Animals, Devices, and Learning clustered within their respective categories.

In contrast to training with random words, using two randomly chosen words per class significantly impaired learning, with only 48% of models acquiring the A->B and A->C implications. Despite this lower success rate, we found that closer embeddings within class A (cosine distance) significantly increased the likelihood of observing symmetry (p=0.00325) in the models that did learn the initial associations.

Employing the aforementioned three classes, we generated a word dataset characterized by the presence of embedding clusters representing both short and long inter-cluster distances within each class. A subsequent replication involving one hundred model instantiations confirmed the positive correlation between proximity of A-class embeddings and the emergence of symmetry, irrespective of the inter-embedding distances within classes B and C. Figure 4 presents a graphical representation of the inter-embedding distances between classes A and B. Instances exhibiting the emergence of symmetry are plotted in black.

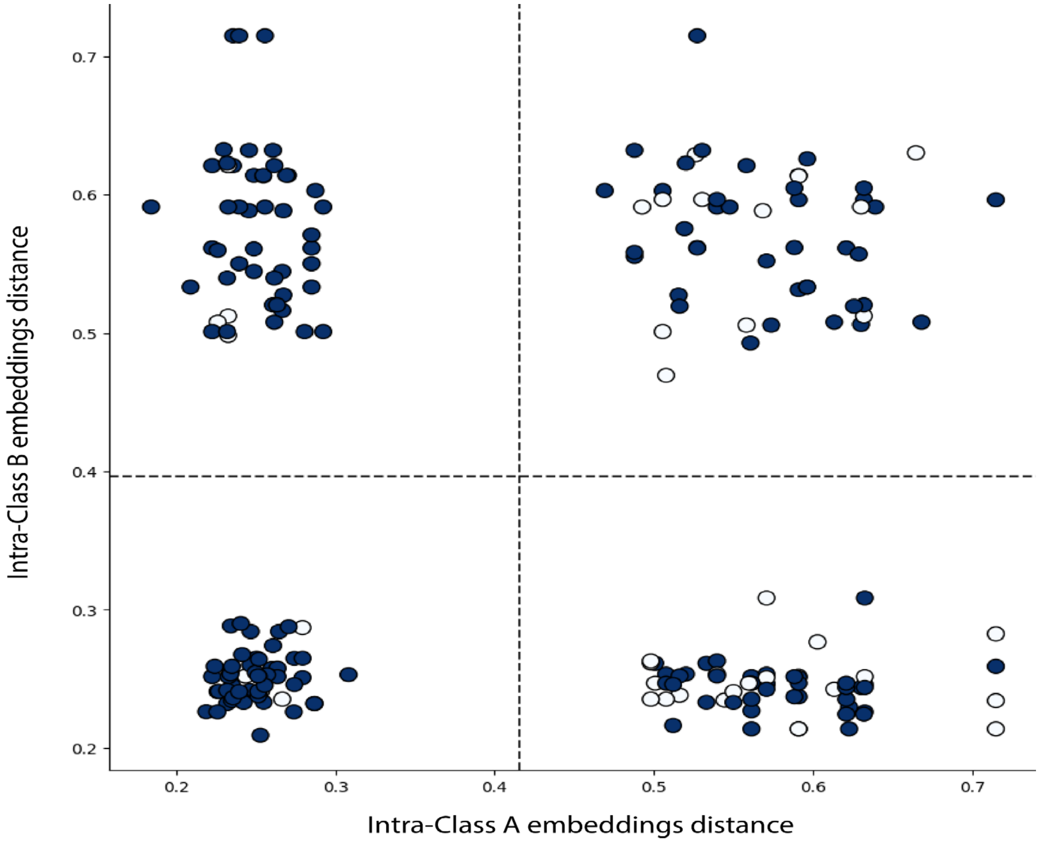

Figure 4: Graphical representation of the relationship between intra-class embedding distances and the emergence of symmetry. Instances exhibiting symmetry following reinforced conditioning are indicated by black markers. Quadrants are defined based on the sample median for each respective axis.

Thus, the closer the word embeddings in class A, the higher the probability of B->A symmetry emergence (0.92 vs 0.69, p < 0.01).

Among the models that passed the symmetry test (B->A or C->A), 58% also exhibited transitivity (B->C or C->B). We then explore how B->A and C->A symmetry impact on B->C and C->B transitivity. We found that 85% of the models where symmetry emerges were able to succeed in transitivity tests, showing that the emergence of symmetry from B->A (C->A) is necessary and almost sufficient to assure transitivity from B->C (C->B).

Discussion

Large Language Models (LLMs) have rapidly permeated nearly all disciplines, transforming established work practices. LLMs automate routine tasks, including guiding customers through menus, quickly summarizing and classifying text [9], and providing an additional layer of quality control in medical image processing among others [10]. However, while the accuracy of LLMs predictions is estimated in base of the training corpus, their performance in real situations, i.e. in the production stage, can only be checked by users who may not be experts in the matter. Here, we investigated the language-learning capabilities of LLMs by adapting established psychological testing paradigms typically employed with human participants. Specifically, we examined performance on symmetry tests, a necessary prerequisite for inferring B->A relationships given prior learning of A->B associations. Symmetry constitutes a necessary condition for the development of equivalence relations. The accurate execution of B->C (and C->B) tests is contingent upon the prior establishment of the symmetric relationship B->A (or C->A) and the subsequent retrieval and application of a previously learned rule, A->C (or A->B) [6].

Our findings reveal a critical limitation in state-of-the-art LLMs (ChatGPT 3.5, ChatGPT 4, and Gemini): they cannot learn equivalence relations between random six-character uppercase words, despite their ability to learn A->B and A->C implications. We also find that the semantic distance between embeddings of English words in the A class plays a crucial role in B->A and C->A symmetry test performance. This contrasts sharply with human language acquisition, where equivalence relations often develop alongside language skills, suggesting a fundamental difference in how LLMs and humans process language.

Our findings could have implications for the development of AI-driven conversational bots. In therapeutic contexts, understanding nuanced relationships between concepts and emotions is crucial. For a patient who expresses feeling “anxious” and “overwhelmed” a human therapist can readily understand that these two states might be related, perhaps even equivalent in some contexts, and can explore this connection with the patient. Our results suggest that current LLMs may struggle with such relational understanding. They might process the words “anxious” and “overwhelmed” as separate entities rather than recognizing the underlying connection. Whether the emergence of equivalence relations among previously unrelated stimuli following reinforced training is crucial for the effectiveness of Large Language Models (LLMs) remains an open question. However, LLMs may offer a valuable starting point for understanding how language naturally arises in humans. Moreover, biologically plausible models of language, where equivalence relations emerge naturally after reinforced training of A->B and A->C associations [11,12], may offer insights for enhancing LLMs.

From an operant conditioning perspective, the failure of LLMs to form equivalence relations could be interpreted as some problem in the reinforcement history. While LLMs are trained on vast datasets, the training regimen could not adequately reinforce the relational aspects of language, focusing instead on individual word associations or statistical regularities. The LLMs might be learning what words typically appear together, but not the underlying relationships between them. This contrasts with human learning, where even simple associative learning (A->B, A->C) can lead to the spontaneous emergence of derived relational responding, like symmetry (B->A, C->A) and equivalence (B->C, C->B), given the appropriate contextual cues.

The present study does not directly address the specific mechanisms underlying the observed deficit in equivalence relation acquisition in LLMs. While factors such as attentional mechanisms, masking procedures during training and low rank adaptation may play a role, the computational and environmental costs associated with de novo LLM training render this line of inquiry infeasible at this juncture. However, the methodology presented herein offers a valuable “black box” paradigm for the empirical investigation of these complex systems, analogous to Skinner’s behavioral approach in animal psychology [13].

References

1. Lake BM, Baroni M (2023) Human-like systematic generalization through a meta-learning neural network. Nature 623: 115-121.

2. Acerbi A, Stubbersfield JM (2023) Large language models show human-like content biases in transmission chain experiments. Proc Natl Acad Sci 120(44): e2313790120.

3. Ott S, Hebenstreit K, Liévin V, Hother CE, Moradi M, et al. (2023) ThoughtSource: A central hub for large language

model reasoning data. Sci Data 10: 528.

4. Coda-Forno J, Binz M, Akata Z, Botvinick M, Wang JX, et al. (2023) Meta-in-context learning in large language models. Advances in Neural Information Processing Systems 36 (2023): 65189-65201.

5. Hu Z, Wang L, Lan Y, Xu W, Lim EP, et al. (2023) LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. arXiv preprint arXiv: 2304.01933

6. Sidman M, Rauzin R, Lazar R, Cunningham S, Tailby W, et al. (1982) A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children. Journal of the experimental analysis of behavior 37(1): 23-44.

7. Sidman M, Willson-Morris M, Kirk B (1986) Matching-to-sample procedures and the development of equivalence relations: The role of naming. Analysis and intervention in Developmental Disabilities 6(1-2): 1-19.

8. Sidman M (2000) Equivalence relations and the reinforcement contingency. Journal of the Experimental Analysis of behavior 74(1): 127-146.

9. Wei F, Keeling R, Huber-Fliflet N, Zhang J, Dabrowski A (2023) Empirical study of llm fine-tuning for text classification in legal document review. 2023 IEEE International Conference on Big Data (BigData), pp: 2786-2792.

10. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, et al. (2024) Transformer models in biomedicine. BMC Med Inform Decis Mak 24(1): 214.

11. Lew SE, Zanutto BS (2011) A computational theory for the learning of equivalence relations. Front Hum Neurosci 5: 113.

12. Mofrad AA, Yazidi A, Mofrad SA, Hammer HL, Arntzen E (2021) Enhanced Equivalence Projective Simulation: A Framework for Modeling Formation of Stimulus Equivalence Classes. Neural Comput 33(2): 483-527.

13. Skinner BF (1938) The Behavior of Organisms: An Experimental Analysis. BF Skinner Foundation, Cambridge, Massachusetts.