Human vs. Large Language Model (LLM) Performance On Sentences With Linguistic Nuances

Will James (wjames38@gatech.edu)
Aniyah Bussey (abussey6@gatech.edu)
Sixu Li (sli941@gatech.edu)

Georgia Institute of Technology
Atlanta, Georgia United States

Abstract

With the rise of Artificial Intelligence (AI) and Large Language Models (LLMs), which are said to mimic human intelligence, how do they compare to real human performance of sentences that are difficult to parse due to linguistic nuances? Solving this question will not only allow us to compare AI and human responses to questions, but also to extract even more information about the responses given and the differences in processing of the human brain compared to different LLMs. For this study, we created a list of sentences with these linguistic nuances and grouped them into five categories of idioms, metaphors, similes, ambiguities, and dialects, along with including normal sentences for a baseline control. With these sentences, we tested humans and four different LLMs with varying parameter sizes to see the performance differences. We found that across all sentence types tested, humans were 93.1% accurate, which was better than the two models with the lowest parameters and worse than the two models with the highest parameter count. In general, this study highlights the differences in human and LLM understanding of sentences, as well as the importance of the parameter count in models.

Keywords: Artificial Intelligence; Large Language Models; Neuro AI; Human Language Processing; Language Ambiguity

Introduction

Motivation

With increasing interest in the intersection of humans and AI models, specifically in terms of performance, they need to be tested against each other in all areas, including sentence understanding. This study will give us the answer to human vs. AI in terms of understanding certain types of sentences. If they are compared rigorously, we can then draw conclusions about possible limitations or extensions of LLMs that should or should not be made for humans to be able to better understand all sentences.

Context

Most people in developing or developed countries now know about AI and LLMs. They use them in daily life and are experimenting with its capabilities, including very detailed individual questions, similar to detailed sentences with ambiguities.

The current research on this particular topic is very slim and only relates to certain parts of the research such as testing LLMs directly to fMRI data (Antonello, Vaidya, & Huth, 2023) or using primarily an audio only LLM to do testing (Cho, Cho, Kang, & Kim, 2019). However, we know from prior research even from the 1990's that constraint satisfaction, which is still used behind the scenes to determine output in LLMs, is not good for sentence processing (Frazier, 1995). Going into the study, we also know that just because the model and human may have the same response, that does not mean they are the same or have the same level of intelligence (Feather, Khosla, Murty, & Nayebi, 2025).

Methods

We started by creating a spreadsheet of 20 sentences for each sentence type including normal, idioms, metaphors, similes, ambiguities, and dialects. These sentences were taken from both our own thinking and experience, as well as from lists online. Both humans and the LLMs tested were given the following prompt for each sentence:

- **Prompt**: Explain what this sentence means and is trying to convey. Use simple and everyday words to explain so that anyone can understand your explanation. If a sentence seems to come from a certain region in the United States, include that in your explanation as well.

The sentences are analyzed by both the human and LLMs and then scored for understanding using the guidelines below, as well as taking one point off if a regional sentence is not or is incorrectly identified.

Table 1: Human & LLM Scoring Rubric

Score	Understanding Level
0	No Attempt/Irrelevant
1	Minimal Understanding
2	Partial Understanding
3	Adequate Understanding
4	Strong Understanding
5	Superior Understanding

Human Responses

To get data points of human understanding of different sentence types, we used a survey to gather responses. They were given 10 random sentences from all of the sentences with the prompt from above. Humans then wrote in their answers to each sentence that was then graded by our scoring rubric above. The 19 respondents of this survey were primarily college aged with English as their primary language.

LLM Responses

To fully compare human responses to LLMs, we used four different models, each with a different parameter amount. We used Google's Gemma 1B, 4B, and 12B models (Team et al., 2025), followed by OpenAI's OSS GPT 20B model (Agarwal et al., 2025). For each model, we gave them each sentence in our list and asked the same prompt as the humans, which is stated above.

Results

After getting the human survey responses and .txt file responses from the LLMs, we went through and scored the results based on our scoring rubric above. Since we scored out of five, we multiplied the number of sentences by five to get the total amount and divided the sum of the correct scores by the total to get the average. Across all sentence types, we see that humans have 93.1% accuracy with the 1B and 4B parameter models being lower and the 12B and 20B models being higher seen in the figure below.

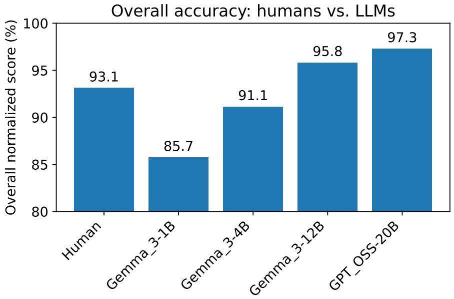


Figure 1: Overall Averages

From the scores, we can also find the accuracy differences for each model with each sentence type compared to human responses as seen in the figure below. We can see that overall, the models with low parameters generally perform worse than humans and vise versa for models with more parameters performing better.

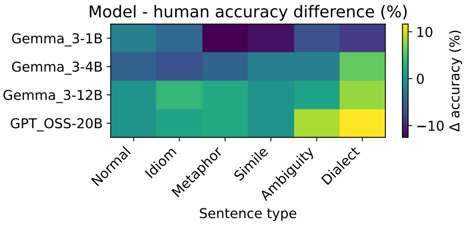
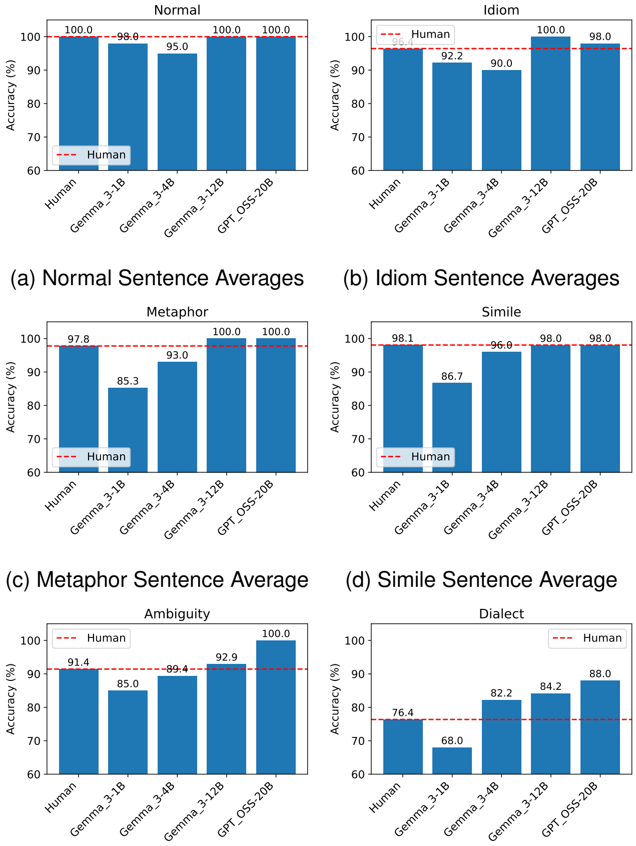


Figure 2: Human & Model Accuracy Comparison

Categorical Results

Below are the accuracy averages for each sentence type.



(a) Normal Sentence Averages

(b) Idiom Sentence Averages

(c) Metaphor Sentence Average

(d) Simile Sentence Average

(e) Ambiguity Sentence Average

(f) Dialect Sentence Average

Figure 3: Human & LLM Averages Per Category

Discussion

What Is Learned

These results show us that small models perform worse than humans across many different sentence types. However, larger models perform just as well or slightly better than humans for the tested sentence types.

Limitations & Future Directions

Even with these promising results, it is still important to notice the overall low understanding of regional dialects by both LLMs and humans. We would naturally expect AI to be better than humans, however it must not be fully trained on specifics and language trends. Another limitation on the LLMs is that they still are hallucinating from time to time, including not answering the question as asked or explaining in too much detail.

Possible future studies and research in these areas could include a deeper dive into specific differences in LLM understanding to see how training could effect the results, and in addition, building a more stable and quantitative measure of accuracy for scoring these models.

Author Contributions

Our team was myself (Will James), Aniyah Bussey, and Sixu Li. We all met together regularly and planned equally in the planning stages. Here is an outline of individual contributions:

- **Will James**: I created the spreadsheet for sentences and contributed to several of the categories of sentences. For the human questionnaire, I fully created the survey by importing the sentences and limited/randomized questions. During the scoring process, I fully scored all of the human and LLM responses based off of our scoring rubric.

- **Aniyah Bussey**: She came up with our hypothesis and research idea from prior experience in a linguistics class including finding prior datasets. Contributed a good amount to the sentence spreadsheet including lots of the dialect section along with other sections. She created a flyer that was distributed and hung up for our human survey.

- **Sixu Li**: He did all of research on possible LLMs to use for our study that could be used with an API to easily send it the sentences. During the testing phase, he tested all of the LLMs by sending the sentence and prompt through and exported the results to be scored. For our results, he created the graphs and charts that are included in this paper.

References

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., ... others (2025). gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Antonello, R. J., Vaidya, A. R., & Huth, A. G. (2023). Scaling laws for language encoding models in fmri. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC11258918

Cho, W. I., Cho, J., Kang, W. H., & Kim, N. S. (2019). Text matters but speech influences: A computational analysis of syntactic ambiguity resolution. Retrieved from https://doi.org/10.48550/arxiv.1910.09275

Feather, J., Khosla, M., Murty, N. A. R., & Nayebi, A. (2025). Brain-model evaluations need the neuroai turing test. Retrieved from https://arxiv.org/abs/2502.16238

Frazier, L. (1995). Constraint satisfaction as a theory of sentence processing. Retrieved from https://doi.org/10.1007/BF02143161

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., ... others (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.