Analyzing Cognitive Skills in NLP: A Comparative Study of Model Predictions vs. Human Performance in Language Prediction Tasks

Yaara Shriki[*] Tel Aviv University yaarashriki@mail.tau.ac.il
Itay Tshuva[*] Tel Aviv University itaytshuva@mail.tau.ac.il
Mark Fesenko[*] Tel Aviv University markfesenko@mail.tau.ac.il
Samuel Amouyal[†] Tel Aviv University

Abstract

Natural Language Processing (NLP) models have made remarkable strides in various language-related tasks, demonstrating outstanding linguistic capabilities across a spectrum of tasks, including question answering, sentence comprehension, summarization, common sense reasoning and translation. However, understanding the extent of their cognitive capabilities remains a challenging endeavor. A pivotal question arises: to what degree do the mechanisms that underlie language comprehension in humans correspond to those employed by language models? In this study, we aim to investigate the similarities and disparities between these two entities at different checkpoints of the trained language model. We do this by studying the cognitive skills developed by LLMs, exemplified by MultiBERTs, in analyzing complex sentence structures and compare the models' predictions with human performance. Our preliminary results indicate that models outperformed human participants in most linguistic scenarios we tested. All of our code can be found in our GitHub repository[1].

1 Introduction

Humans exhibit remarkable proficiency in comprehending language. They effortlessly communicate through complex sentences, demonstrating the ability to understand them accurately, even when encountering such sentences for the first time. However, researchers have observed that certain types of sentences, often characterized by similarity-based interference, tend to pose greater difficulties for humans (Fernanda Ferreira, 2001; Lena A. Jäger, 2017).

Neural language models also exhibit exceptional linguistic capabilities, and recent research in the field has uncovered intriguing parallels between humans and state-of-the-art Large Language Models (LLMs) in sentence processing. These LLMs perform remarkably well across a wide range of tasks without the need for prior, task-specific training. Yet, despite their impressive performance on complex tasks, studies have unveiled that they also encounter challenges when confronted with seemingly straightforward linguistic missions (Avia Efrat, 2022). This shows that even though interpretability has shed some light on these models, we are far from understanding the depth and scope of their internal knowledge.

Recent research in the field has provided fascinating insights into the behavior of LLMs that closely resemble human language processing (Rik Huijzer, 2023). Additionally, studies have delved into the correlation between LLM layers and human brain activity during language tasks, revealing intriguing parallels (Charlotte Caucheteux, 2022). These findings strongly suggest a potential correlation between the way humans and LLMs process language.

Finding evidence that LLMs often struggle with tasks that humans find challenging may strengthen the hypothesis that shared linguistic difficulties reflect a deeper connection between human cognition and these models, and that certain aspects of language comprehension and reasoning may be inherently complex, irrespective of whether the entity attempting them is biological or artificial.

2 Related Work

In this section, we review the existing literature and research relevant to our study. Our research focuses on examining the cognitive skills of LLMs, and comparing them to those of humans. Understanding how humans process language and develop cognitive skills has been a subject of extensive research (Frazier, 1987; Gordon, 2001; C., 2012). These studies have provided insights into the neural mechanisms involved in language comprehension and the role of cognitive processes, such as memory

[*]The authors contributed equally.
[†]Advised by Samuel Amouyal, PhD student.
[1]https://github.com/yaaras/CognitiveSkills

and attention, in sentence processing (Martin, 1993; Levy, 2008). Previous studies have also explored the capabilities of NLP models in various language-related tasks, such as question answering, language comprehension, and text generation (Avia Efrat, 2022; Brown, 2020). Comparative studies between human language comprehension and NLP models have gained attention in recent years, as LLMs became more prominent. These studies aim to uncover similarities and disparities in how humans and models process language and answer questions (Schrimpf, 2021; Charlotte Caucheteux, 2022). Despite the valuable insights gained from previous studies, there remain gaps and limitations in our understanding of the cognitive skills of NLP models and their alignment with human cognition. In the following sections, we present our methodology, experimental setup, and results.

3 Method

3.1 Objective

The primary objective of this research is to investigate the capabilities of Large Language Models (LLMs) in comprehending complex sentences, particularly those that are challenging for human comprehension due to their intricate structure.

3.2 Data Collection

We incorporated two existing datasets, originally curated in Hebrew, that aim to measure different aspects of the language in subjects.

The first dataset (see Appendix A.1) used in our study was drawn from the research conducted by Tal Ness and Aya Meltzer-Asscher (Tal Ness, 2019) , titled 'When is the verb a potential gap site? The influence of filler maintenance on the active search for a gap.' In their paper, they investigated the influence of filler maintenance on the parser's attempt to posit a gap when encountering a verb. Specifically, they examined whether the availability of certain features of the filler, actively maintained in working memory, guides the process of active gap-filling. The findings from their experiments suggest that verbs selecting an argument with features similar to those maintained in the filler's representation are more likely to trigger an attempt to resolve the dependency.

The second dataset (see Appendix A.2) we examine in our research was presented in Koesterich et al. (2021). The study delves into the phenomenon of interference observed in language processing, particularly in the comprehension of filler-gap and filler-resumptive dependencies. The research utilized datasets from Hebrew object relative clauses and conducted two comprehension experiments. The first experiment, involving 64 participants, focused on verbs that take an Indirect Object (IO) complement, where relativization in Hebrew is obligatorily realized by a Resumptive Pronoun (RP). The results showed significant interference effects when a gender-matching distractor was present. The second experiment, with 65 participants, used verbs that take a Direct Object (DO) complement, allowing for relativization either by an RP or a gap. The findings mirrored the first experiment, revealing interference effects in both RP and gap conditions.

In both datasets, sentences contain embedded clauses, where two Noun Phrases within the sentence share a particular feature. An embedded clause is a grammatical construct placed within another clause, providing additional information related to the sentence's topic. This structure is used to offer readers more context and details. However, the inclusion of this additional information can introduce ambiguity and pose challenges in understanding which part of the sentence refers to whom or what. Similar to challenges encountered by humans, we believe that these difficulties may also manifest in the case of LLMs.

3.3 Model

In our experiments, we applied MultiBERTs (Sellam et al., 2021), a collection of BERT-base models that share the same hyper-parameters over 140 intermediate checkpoints captured during pretraining (28 checkpoints for five different seeds). It uses the Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) objectives and is a reproduction of BERT-base uncased, for English. As we will be using a variant of the Masked Language Model, our approach will involve testing various checkpoints of the model with masked sentences, and then comparing the results with those of humans. Our definition of the best model is the model which performs the best at a certain checkpoint.

3.4 Preprocessing

To achieve this, we performed some preprocessing on the acquired datasets. For both datasets, samples included multiple components: the sentence, the question, and parameters under examination, such

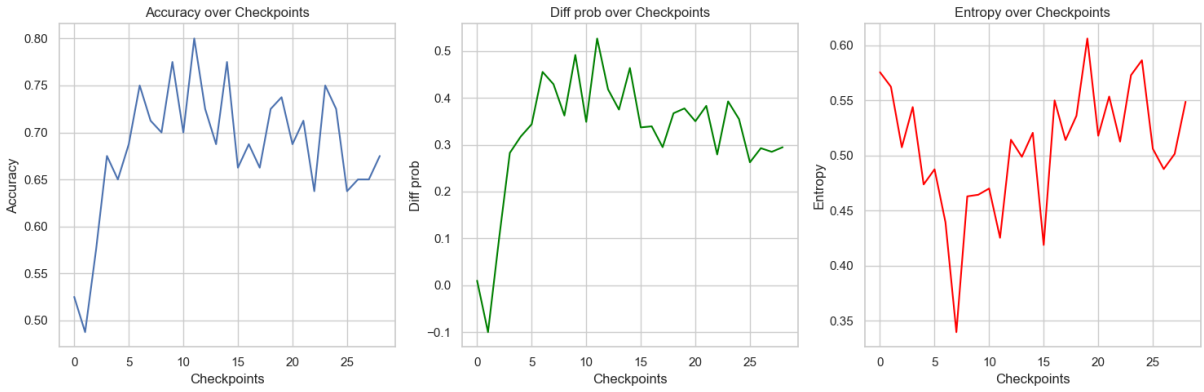

Figure 1: Performance metrics over checkpoints for the dataset presented in Ness et al. (a) Accuracy, (b) Difference in Probability, and (c) Entropy.

as the question type, the animacy, etc. Given that the datasets were originally in Hebrew and later translated into English, the gender parameter lost its contextual significance in the English language, and consequently, we decided to adapt the parameter to reflect singular and plural forms instead. This adjustment was deemed necessary to maintain the contextual relevance and ensure the accuracy of the data analysis in the English language. We then identified both the distractor (the word designed to mislead the human or model) and the label within the dataset.

Furthermore, as part of our experimental setup, we transformed the sentence-question pairs into masked-sentences. This involved replacing the labels in the sentences with mask tokens and then assessing the model's predictive ability for the masked words. This transformation allowed us to explore how well the model performed in predicting missing information within the sentences, a critical aspect of understanding the language processing capabilities of the model.

3.5 Evaluation Metrics

In our analysis, we employed a selection of evaluation metrics to comprehensively assess the performance of the selected model. For a better understanding of what each metric signifies, we've provided a brief description alongside. Table 1 outlines these metrics and their descriptions.

Metric	Description
Accuracy	Proportion of correct predictions out of total predictions.
Diff Prob	Measure of the model's differentiation capacity between labels and distractors, ranges from -1 to 1.
Entropy	Measure of uncertainty in predictions of labels.

Table 1: Overview of Evaluation Metrics.

4 Experiments

In this section, we present plots that depict various metrics computed at different checkpoints of the model collection for a specific seed. The metrics being examined include accuracy, 'diff prob', which measures the disparity between label and distractor probabilities, and entropy derived from the 'diff prob' metric, offering insights into the datasets' information content and uncertainty.

These metrics are evaluated on two distinct datasets: Ness et al. and Koesterich et al. and across various characteristics subgroups in each dataset. It's important to emphasize that while our illustrations focus on a specific seed, the same behavior was consistently observed across different seeds, reinforcing our findings.

4.1 Ness et al. Dataset

In Figure 1, three distinct metrics are presented side-by-side. The leftmost plot, depicted in blue, showcases the model's accuracy. The x-axis represents the progression of checkpoints, while the y-axis indicates the accuracy values. Adjacent to this, the center plot in green outlines the 'diff prob' metric across the same checkpoints, which illustrates the difference between label and distractor probabilities. Lastly, the rightmost plot, rendered in red, displays the entropy values.

We begin by examining the accuracy trends. We notice some unexpected behavior. Our initial an-

ticipation was that we would witness mainly a steady increase in accuracy as the model progresses through the various checkpoints. The model accuracy began at a low rate and showed an incremental trend over the first 10 checkpoints. However, following this, there was a decline in the model's performance. In the following sections, we discuss possible explanations for this phenomenon.

Next, we take a look at a plot that explores the difference between the probabilities assigned to the true label and the distractor. This plot shares similarities with our previous accuracy plot. The difference values in this plot range from -1 to 1, where a negative value indicates a higher probability assigned to the distractor, while a positive value signifies a higher probability assigned to the true label. It's important to note that we calculate the mean difference for each checkpoint, over all sentences. The presented graph indicates a predominant assignment of higher probabilities to the label over the distractor during the initial 10 checkpoints. This suggests effective learning by the model in the early stages. However, post this period, there's a noticeable shift as the model starts giving reduced probabilities to the label—consistent with the trends observed in the accuracy plot.

Lastly, we also present entropy graph, trying to understand further the language model's decision-making process. Observing the entropy provides insights into the model's level of certainty when making predictions. During the first 10 checkpoints, the level of certainty of the model improves and reaches 0.35, but as we saw in the previous graph, as the training progresses, the model starts to exhibit increased uncertainty. This increase in entropy aligns with our prior observations where the model displayed a shift in probability assignment away from the label towards the distractor.

Accordingly, we observe that as the accuracy increases, the certainty also rises. Conversely, a slight dip in accuracy results in a significant drop in certainty. Thus, when the model is predominantly correct, it is confident in its predictions; however, when it is slightly off, its confidence in its predictions diminishes considerably.

Based on the model's performance of the first 10 checkpoints on the Ness et al. dataset, it becomes clear that there exists some similarity in feature distribution between the model's training data and this dataset. Typically, the accuracy on the training set displays a near-monotonic increase, with minor fluctuations attributable to noise, reflecting the model's capacity to comprehend the intrinsic patterns of its training data. Yet, around the 10^{th} checkpoint, we observe a decrease in accuracy on the Ness et al. dataset. This suggests that as the model becomes increasingly refined to the nuances of its training data, it compromises its generalization ability on the external dataset. Such behavior is a distinct sign of overfitting, further enhanced by the differing data distributions between the model's training set and the Ness et al. dataset. Essentially, the model has become overly specialized for one dataset, undermining its adaptability to datasets with varied properties.

4.1.1 Syntactic and Semantic Parameters

Transitioning from the broad overview, we now delve deeper, segmenting our data to focus on specific aspects. By categorizing the results according to distinct characteristics under investigation, we aim to shed light on the subtleties and relationships within the data.

Our analysis is anchored around two pivotal parameters. The first parameter delves into the syntactic intricacies of sentence structures, specifically contrasting the Filler-Gap dependencies with the Subject-Verb agreements (see Table 2). This exploration seeks to understand how these syntactic constructs influence the model's performance and comprehension capabilities.

The second parameter is focused on the semantic aspect, examining the match versus mismatch of animacy between the label and its distractor. By investigating this, we aim to discern how the congruence or incongruence of animacy attributes affects the model's decision-making and accuracy. Together, these parameters provide a comprehensive view of both the syntactic and semantic challenges presented by the dataset.

Analyzing the results from the Ness et al. dataset, as seen in Figure 2, offers insightful conclusions about the model's linguistic capabilities in comparison to human cognition. The model's performance in Filler-Gap (F-G) structures reveals a distinct sensitivity to animacy. While human performance appears to show weak sensitivity to animacy between animacy match and mismatch (0.69 vs. 0.75) compared to the model performance.

In comparing human and model performances on sentences with filler-gap dependencies, there is a discernible difference in how each handles animacy congruence. For animacy matches, hu-

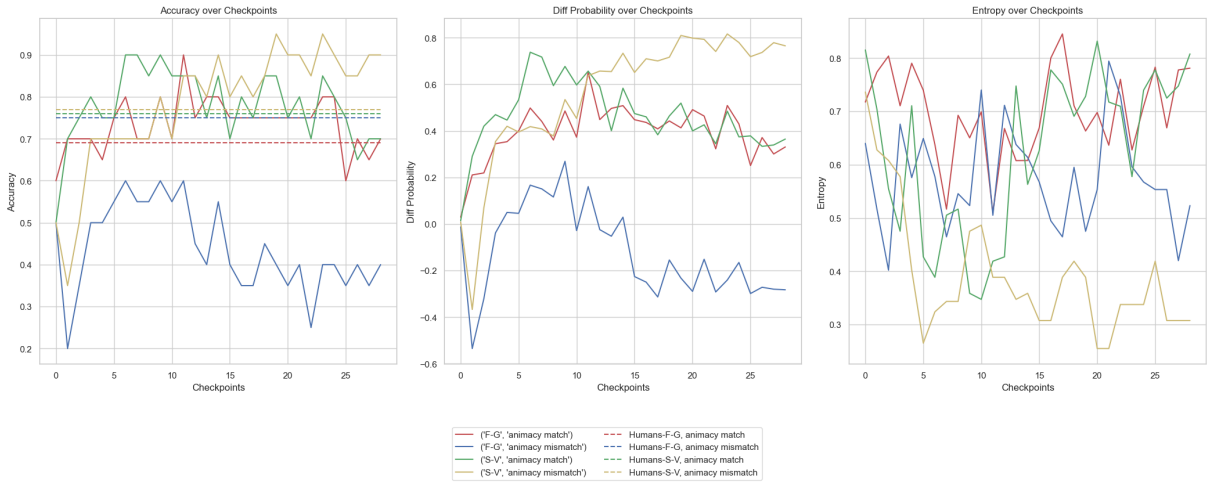

Figure 2: Detailed analysis of performance metrics categorized by linguistic constructs for the dataset from Ness et al. (a) Accuracy, (b) Difference in Probability, and (c) Entropy for each linguistic construct.

mans exhibit a lower accuracy (0.69) compared to the model, which achieves a high accuracy of 0.9. This suggests that the model is better at processing sentences where the animacy between label and distractor is consistent. Conversely, for animacy mismatches, humans outperform the model with an accuracy of 0.76 against the model's 0.6. This reversal highlights a human strength in dealing with semantic inconsistencies that the model has not yet mastered.

Intriguingly, in Subject-Verb (S-V) agreement structures, the model surpasses human results in both animacy match and mismatch structures. This might suggest the model's ability to discern patterns in the data that aren't as pronounced for humans. Additionally, the model's performance does not correlate with human performance.

Next, we examine the difference between the true label and the distractor's probability. As before, when there is a mismatch in the animacy condition in the context of the F-G dependency (Figure 2), we consistently observe negative difference values. This pattern reaffirms our earlier observations, suggesting that F-G dependencies with animacy mismatch tend to steer the model towards favoring the distractor.

Conversely, in the S-V agreement with mismatch in animacy condition, we observed a significant disparity in the probability allocation between the label and distractor. Specifically, after 20 checkpoints, the difference in probability between the label and the distractor approximated 0.8. However, when there was a match in animacy, the model's difference in probability stabilized around 0.4. This suggests that the model confidently distinguished between the label and distractor in Subject-Verb agreement sentences where the animacy between them was inconsistent, but demonstrated more ambiguity or reduced confidence when the animacy was consistent.

In analyzing the entropy graph (Figure 2) for the S-V agreement sentences, a clear pattern emerges. For cases with an animacy mismatch, the entropy is notably low, stabilizing around a value of 0.2. This low entropy suggests that the model is relatively certain about its decisions when faced with animacy inconsistencies in S-V structures. In contrast, for sentences where there's an animacy match and in F-G dependency sentences, the entropy portrays a different story. It fluctuates considerably, indicating a higher degree of uncertainty in the model's predictions. Specifically, entropy values span a range between 0.4 to 0.8 for these groups. This heightened variability and noise in the entropy graph, particularly for animacy-matching S-V sentences and all F-G sentences, highlight potential areas where the model's confidence wavers. Such observations suggest that the model's cognitive understanding, while proficient in some linguistic constructs, still grapples with consistent confidence in others, mimicking the intricate complexities of human linguistic processing.

Thus, in summary, it's evident that MultiBERTs models and humans exhibit divergent strengths in processing complex sentences. Models outperform humans in Subject-Verb agreement tasks both when animacy is matched and mismatched, possibly due to their ability to detect patterns that humans over-

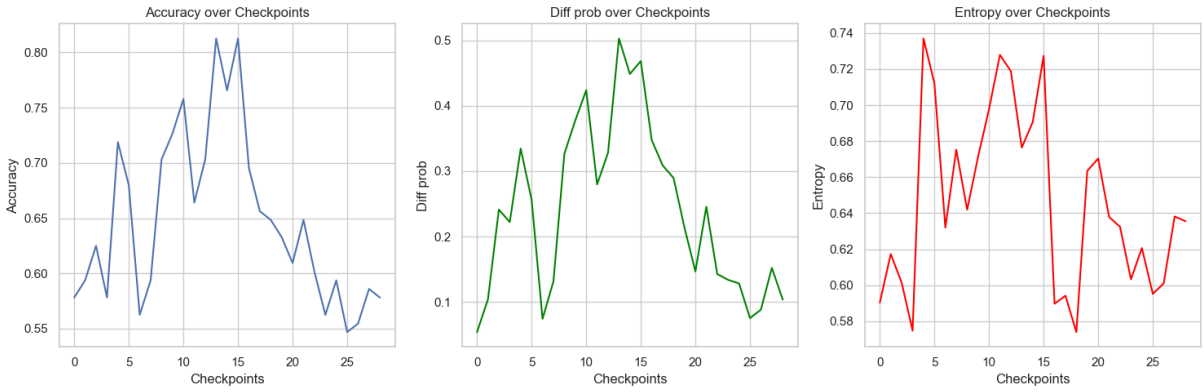

Figure 3: Performance metrics over checkpoints for the dataset presented in Koesterich et al. (a) Accuracy, (b) Difference in Probability, and (c) Entropy.

look. In contrast, models underperform in Filler-Gap dependencies with animacy mismatches, and outperform humans when animacy is matched. Unlike humans who show relative consistency, models are sensitive to animacy in Filler-Gap dependency sentences. These findings demonstrate that while NLP models can sometimes surpass human linguistic capabilities, their understanding is uneven, when complex syntactic and semantic factors interact.

4.2 Koesterich et al. Dataset

Moving to the next dataset (see Figure 3), we observe further unusual behavior. In the beginning, the accuracy sets at a notable mark of around 0.6. As we progress through the checkpoints, there's an encouraging ascent, peaking near 0.8. However, towards the latter stages, the accuracy descends back to its initial level of 0.6, resulting in a bell-curve pattern.

Subsequently, we turn our attention to the graph illustrating the difference between the label and the distractor.'s probabilities. This representation too showcases an unusual trend, similar to the accuracy graph. The graph reveals that the model initially gives a notably higher probability to the label over the distractor. However, this precedence diminishes after checkpoint 15, eventually leveling both probabilities at 0.1.

Next, we will delve deeper into the analysis of the entropy graph to gain more comprehensive insights. It is evident that there is a noticeable fluctuation across checkpoints. The entropy values oscillate between 0.58 and 0.74 without manifesting a discernible pattern. A plausible explanation for this observed variability is the potential difference in the data distributions between the training data and the external dataset. Discrepancies in data distributions can lead to such inconsistencies in the entropy values, highlighting the importance of ensuring similarity in distributions when evaluating model performance.

In more detail, the model performance on the previous dataset displayed little overfitting, but the performance on the present dataset shows a more significant concern. Initially, the model demonstrated promising learning. However, after 15 checkpoints, its capacity for generalization diminished considerably, resulting in a nearly complete loss of generalization. This is evident from the bell-shaped accuracy curve. Such a result has led us to ponder the underlying cause of this fluctuation. We have proposed a hypothesis to explain this unusual pattern. It is plausible that the LLMs being trained on conventional data, struggle with difficulties when presented with the task of predicting words in atypical sentence structures. As can be seen in the appendix, the sentence structures in this dataset are indeed unconventional. Consequently, it is conceivable that during the model's training phase, it could compromise its generalization abilities when confronted with such unique data structures.

4.2.1 Syntactic and Semantic Parameters

In our examination of the dataset from Koesterich et al., we direct our attention to two salient linguistic parameters. The first parameter delves into the realm of number agreement, contrasting sentences where there's a match in singular/plural forms between the label and its distractor, against those where a mismatch occurs. This exploration is instrumental in discerning how the model navigates the challenges of number agreement and its implications on semantic understanding.

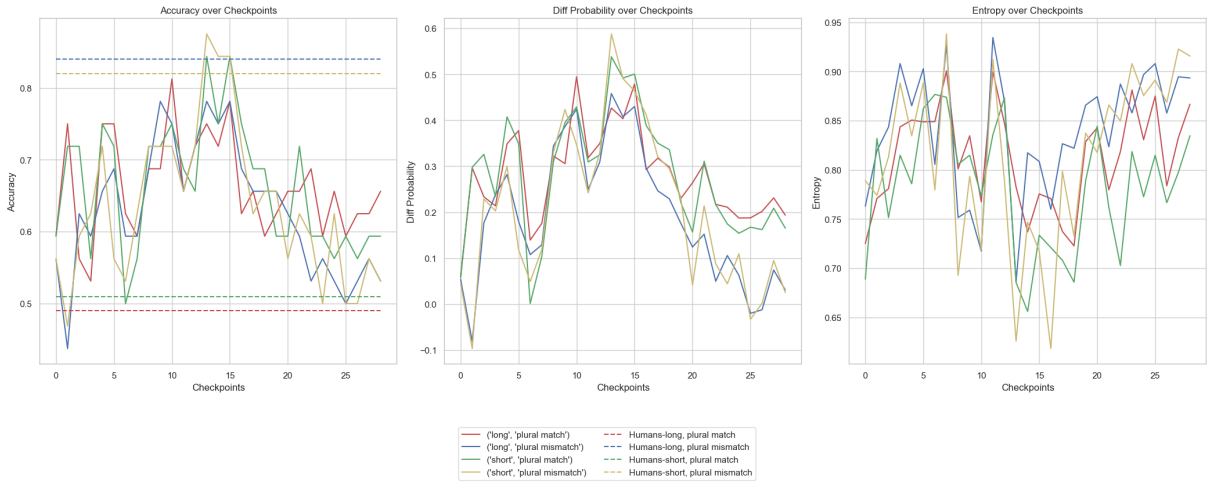

Figure 4: Detailed analysis of performance metrics categorized by linguistic constructs for the dataset from Koesterich et al. (a) Accuracy, (b) Difference in Probability, and (c) Entropy for each linguistic construct.

The second parameter centers on syntactic complexity, differentiating between short and long sentences. This distinction allows us to assess the model's capacity to process and comprehend sentences of varying lengths and inherent structural complexities.

When examining the 'match' and 'mismatch' conditions for 'long' and 'short' sentences (see Figure 4), we again observe similar plots to the accuracy plots and the Diff plots as in (Figure 3). The atypical graph is the entropy distribution. We observed earlier, that the entropy typically appeared as noise, devoid of any discernible pattern. However, when we examine subgroups, we observe inverted bell curve patterns. These patterns suggest that as the model achieves higher accuracy, its certainty increases; conversely, with decreased accuracy, there is heightened uncertainty. This positive behavior aligns well with the model's performance on prior datasets.

We observe that the performance of humans on sentences with plural matches is low, approximating 0.5. However, the model's peak performance at checkpoint 15 ranges between 0.83 and 0.85. These results are particularly remarkable, as the model's performance on such sentences outperforms human performance.

Regarding the other category of sentences — those with plural mismatches — there are distinct variations based on sentence length. For shorter sentences, the model demonstrates superior performance. Conversely, for longer sentences, humans perform slightly better than models. Nonetheless, the outcomes across both categories are generally comparable. Further, human performance does not correlate with that of the model.

5 Discussion

As previously mentioned, sentences from both datasets feature embedded clauses, where two noun phrases within the sentence exhibit shared characteristics, including distinctions such as animacy versus inanimacy or plural versus singular. These characteristics can present processing challenges for humans, and as observed, they also pose challenges for LLMs. It is worth noting that these types of sentences are unique and not commonly encountered in everyday life. Therefore, generalizing the conclusions to other sentence structure may not be appropriate.

We recognize that the distribution of the data on which an LLM is trained plays a crucial role in its ability to perform well on new, unseen data. We assume that sentences with embedded clauses and specific characteristics, like those we've discussed, may not be as prevalent in the training data for these models, potentially contributing to the observed loss of generalization capabilities dramatically when processing such sentences. We suggest to add more complex sentence structures to the training data when training large language models. As a result, models would be able to generalize more effectively to complex sentences.

After discussing our findings with our tutor and reviewing results obtained on other models such as GPT-3 (Brown, 2020) and models from the T5 family (Chung, 2022), we learned that our MultiBERT model behaves similarly to GPT-3, especially with

the Ness et al. dataset. GPT-3 typically exhibits a behavior that is often the opposite of human behavior, which we also observe when analyzing the accuracy with the Ness et al. dataset.

Another potential research direction is comparing different checkpoints of LLMs to different levels of human reading comprehension abilities, possibly based on subjects' age. This approach could shed light on whether humans and LLMs share similar learning processes. However, it's important to acknowledge that this might not be an easy task, as available human data is often scarce, especially when considering younger age groups.

6 Conclusion

In our comprehensive study, we observed intriguing disparities between model predictions and human performance in language prediction tasks. Notably, under many scenarios, the models demonstrated superior performance compared to human participants. In some specific scenarios, however, the models performed worse than human, or achieved comparable performance. Both datasets did not show a correlation between model performance and human performance.

The results we got draw some conclusions on the dissimilarity between humans and MultiBERTs models sentence understanding and processing, but we believe these findings represent preliminary results, and drawing conclusions about the correlation between humans and LLMs in general at this stage may be premature.

Further research is essential in this evolving field of LLMs, and given it is a relatively new and fast evolving field, we are confident that additional studies will continue to emerge, contributing to a deeper and more comprehensive understanding of LLMs.

7 Limitations

Throughout our project, we encountered several noteworthy limitations that required careful consideration. First and foremost, both datasets we utilized originally existed in Hebrew and underwent translation to English. Consequently, one of the parameters examined in the original papers, specifically the gender of words within sentences, lost its linguistic significance when transposed into English. To maintain the integrity of our analysis, we adapted our approach by replacing the gender parameter with an evaluation of plural agreement vs. mismatch.

Additionally, our models were trained to predict masked words, while the original papers involved assessing whether a sentence is true or not. To align with our approach, we undertook the task of reconstructing the sentence-question pairs into a question-followed-by-question format. Furthermore, to adapt to our masked word prediction framework, we masked the label in the question part, ensuring our methodology remained consistent with the goals of our study.

Acknowledgements

We would like to thank Samuel Amouyal, Ph.D. student, for his invaluable mentorship throughout this project. His guidance, knowledge in the field of linguistics, and insightful discussions greatly contributed to our work.

References

Omer Levy Avia Efrat, Or Honovich. 2022. *LMentry: A Language Model Benchmark of Elementary Language Tasks*. arxiv.

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. . . . Amodei D. Brown, T. B. 2020. *Language models are few-shot learners*. CoRR.

Heyes C. 2012. *New thinking: the evolution of human cognition*. Philosophical transactions of the Royal Society of London. Series B, Biological sciences.

Jean-Rémi King Charlotte Caucheteux. 2022. *Brains and algorithms partially converge in natural language processing*. nature.

Hou L. Longpre S. Zoph B. Tay Y. Fedus W. Li Y. Wang X. Dehghani M. Brahma S. Webson A. Gu-S. S. Dai Z. Suzgun M. Chen X. Chowdhery A. Castro-Ros A. Pellat M. Robinson K. . . . Wei J. Chung, H. W. 2022. *Scaling instruction-finetuned language models*.

Andrew Hollingworth Fernanda Ferreira, Kiel Christianson. 2001. *Misinterpretations of Garden-Path Sentences: Implications for Models of Sentence Processing and Reanalysis*. Journal of Psycholinguistic Research.

L. Frazier. 1987. *Sentence processing: A tutorial review*. Attention and performance 12: The psychology of reading.

Hendrick R. Johnson M. K. Gordon, P. C. 2001. *Memory interference during language processing*. Journal of experimental psychology. Learning, memory, and cognition.

Niki Koesterich, Maayan Keshev, Daria Shamai, and Aya Meltzer-Asscher. 2021. Interference in the comprehension of filler-gap and filler-resumptive dependencies. In *34th annual cuny conference on human sentence processing*.

Shravan Vasishth Lena A. Jäger, Felix Engelmann. 2017. *Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis*. Journal of Memory and Language.

Roger Levy. 2008. *Expectation-based syntactic comprehension*. Cognition.

R. C. Martin. 1993. *Short-term memory and sentence processing: Evidence from neuropsychology*. Memory Cognition.

Yannick Hill Rik Huijzer. 2023. *Large Language Models Show Human Behavior*. psyarxiv.

Blank I. A. Tuckute G. Kauf C. Hosseini E. A. Kanwisher N. Tenenbaum J. B. Fedorenko E. Schrimpf, M. 2021. *The neural architecture of language: Integrative modeling converges on predictive processing*. Proceedings of the National Academy of Sciences.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Aya Meltzer-Asscher Tal Ness. 2019. *When is the verb a potential gap site? The influence offiller maintenance on theactive search for a gap*. Language, Cognition, Neuroscience.

A Dataset examples

A.1 Ness et al. dataset

Sentence	Label	Distractor
The actress who was in the French director's studio this week hit the expensive vase at the entrance. The [MASK] hit the vase.	actress	director
The baby who was in the young kindergartener's daycare today rolled around on the lawn. The [MASK] rolled on the lawn.	baby	teacher
The girl who was at the red-haired comedian's wedding this week amazed all the guests. The [MASK] amazed the guests.	girl	comedian

A.2 Koesterich et al. dataset

Sentence	Label	Distractor
The agent discovered the producers that the musicians excited following the success of the album. Therefore, the musicians excited the [MASK].	producers	agent
The policeman hid the suspects that the criminals threatened after the publicized arrest. Therefore, the criminals threatened the [MASK].	suspects	policeman
The singer met the stars that loyal and devoted fans cared for in recent months during the long absence. Therefore, the fans cared for the [MASK].	stars	singer

B Overview of Syntactic Dependencies

Dependency Type	Description	Example
Filler-Gap	A syntactic construction where a word or phrase is missing, often implied elsewhere in the sentence.	Who did you see _ at the store?
Subject-Verb Agreement	The syntactic rule that subjects and verbs must agree in number (singular vs. plural).	She writes. (correct) vs. She write. (incorrect)

Table 2: Description and examples of Filler-Gap dependencies and Subject-Verb agreement.