Large Language Models for Psycholinguistic Plausibility Pretesting

Anonymous EMNLP submission

Abstract

In psycholinguistics, the creation of controlled materials is crucial to ensure that research outcomes are solely attributed to the intended manipulations and not influenced by extraneous factors. To achieve this, psycholinguists typically *pretest* linguistic materials, where a common pretest is to solicit plausibility judgments from human evaluators on specific sentences. In this work, we investigate whether Language Models (LMs) can be used to generate these plausibility judgements. We investigate a wide range of LMs across multiple linguistic structures and evaluate whether their plausibility judgements correlate with human judgements. We find that GPT-4 plausibility judgements highly correlate with human judgements across the structures we examine, whereas other LMs correlate well with humans on commonly used syntactic structures. We then test whether this correlation implies that LMs can be used instead of humans for pretesting. We find that when coarse-grained plausibility judgements are needed, this works well, but when fine-grained judgements are necessary, even GPT-4 does not provide satisfactory discriminative power.

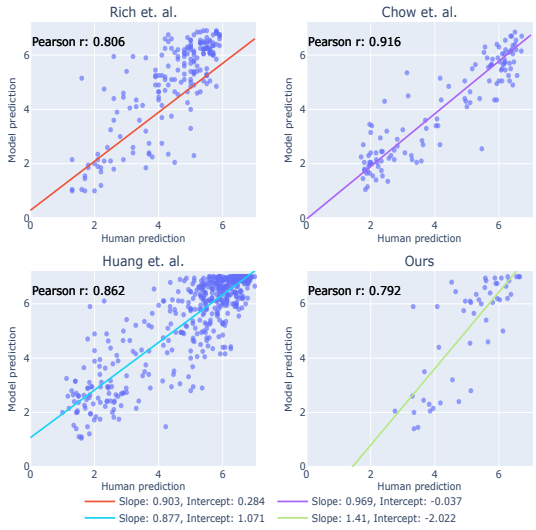

Figure 1: Correlation between average human plausibility ratings and average LLM plausibility ratings across four pretesting datasets, along with the fitted linear regression and Pearson correlation. We plot the LLM with the highest correlation (GPT-4 in all cases, except one where GPT-3.5 is shown).

1 Introduction

Psycholinguistic research explores humans' exceptional language comprehension abilities, aiming to uncover underlying mechanisms through experiments and cognitive modelling (Frazier, 1987; Lewis and Vasishth, 2005; Gibson, 2000; Levy, 2008; MacDonald et al., 1994; Futrell et al., 2020; Tabor and Hutchins, 2004). Researchers use measures such as reading times and comprehension accuracy to compare sentences with distinct processing demands. As an example, Ness and Meltzer-Asscher (2019) investigated reading times to determine if sentences with two animate nouns (e.g., (1a), (2a)) pose greater processing challenges than those with one animate and one inanimate noun (e.g., (1b), (2b)). Longer reading times in the (a) sentences would indicate that similarity between the noun phrases interferes with processing.

1. (a) The photographer that the manager sent was helpful.
   (b) The contract that the manager sent was helpful.
2. (a) The worker that the contractor brought fell down.
   (b) The ladder that the contractor brought fell down.

Careful construction of linguistic stimuli is crucial in psycholinguistic studies to minimize confounding factors. Controlling sentence plausibility ensures that processing differences stem from ex-

perimental manipulations rather than external factors. In our example, making sure that the sentences "the manager sent the photographer" and "the manager sent the contract" have roughly the same plausibility, and likewise that "The photographer was helpful" and "The contract was helpful" have roughly the same plausibility, is necessary to attribute processing variations to the similarity in animacy. Moreover, maintaining overall high sentence plausibility prevents unrelated processing difficulties and reduces data noise.

Controlling sentence plausibility is therefore essential in sentence processing experiments, and is typically accomplished through pretests, where participants rate sentence plausibility on a scale, guiding the selection of materials for the main experiment. However, plausibility pretesting is a time- and resource-consuming process, involving multiple iterations and prolonged data collection with different participant groups.

Recently, Large Language Models (LLMs) (Vaswani et al., 2017; Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Touvron et al., 2023) have shown human-like performance on various language understanding tasks without task-specific training (Brown et al., 2020). Previous studies have established a strong correlation between LMs' predicted probabilities and human reading time (Fernandez Monsalve et al., 2012; Smith and Levy, 2013; Hofmann et al., 2020; Hao et al., 2020; Hollenstein et al., 2021; Shain et al., 2022). Thus, it is natural to ask – *can LMs provide plausibility judgements that are similar to human judgments and consequently be used to reduce the cost of psycholinguistic pretesting?*

In this study, we investigate the correlation between LMs and human plausibility judgments. To accomplish this, we examine four sets of sentences that represent a variety of syntactic structures and plausibility levels, for which human judgements have been collected in prior work in the course of pretesting (Chow et al., 2016; Rich and Wagers, 2020; Huang et al., 2023). We then gather multiple LM judgements for these sets from a wide range of LMs, and compare average human plausibility ratings and average LLM plausibility ratings.

Our findings indicate that while several LLMs exhibit high correlation with human judgments on common syntactic structures, only GPT-4 shows strong correlation on the rarer syntactic structures. Figure 1 displays the average plausibility ratings of the LLM with the highest correlation against average human ratings, along with a linear regression model. The Pearson correlation between LLM and human judgments is consistently high across all the datasets. Interestingly, the fitted linear regressions are quite similar across three of the datasets, indicating robustness in the translation of LLM judgements into human judgements.

Based on these findings, we examine if using LLMs instead of humans can lead to similar outcomes when filtering materials in the course of pretesting. We find that when pretesting requires coarse-grained plausibility judgements, i.e., when it is used to filter out implausible sentences, LLMs perform well. However, when fine-grained plausibility judgements are needed, e.g., to ensure that a pair of sentences has similar plausibilty ratings, even GPT-4's performance is not satisfactory yet.

To summarize, in this work we thoroughly investigate the correlation between human and LM plausibility judgements across a wide range of LMs and syntactic structures. We find that many LLMs perform well on simple syntactic structures, and GPT-4 performs well across-the-board. We translate this finding into a method for using LLMs to provide plausibility judgements, and find that performance is high when coarse-grained judgements are needed, but still lagging behind when fine-grained judgements are necessary.

2 Experimental Setup

An experiment is defined by instantiating three parameters: (a) the LM used for eliciting plausibility judgements, (b) the prompt provided as input to the LM, and (c) the linguistic dataset used. We leverage data from existing pretests for which human plausibility ratings were already collected (Chow et al., 2016; Rich and Wagers, 2020; Huang et al., 2023), and also create our own pretest materials and collect human plausibility judgements for them.

In all experiments, we generate 20 plausibility ratings per sentence per LM, using a scale from 1 to 7. We now describe the datasets (§2.1), LMs (§2.2), and prompts (§2.3).

2.1 Datasets

We use four datasets, which cover a wide range of linguistic phenomena. Table 1 provides examples from all datasets.

Dataset	Structure	Plaus.	Example	Num.
Chow et al. (2016)	Emb. Obj. Quest.	Plaus	The park ranger documented which eagle the hunter had shot.	60
	Emb. Obj. Quest.	Implaus	The park ranger documented which hunter the eagle had shot.	60
Huang et al. (2023)	Emb. Decl.	Plaus	The suspect showed that the file deserved further investigation during the murder trial.	24
	Emb. Decl.	Implaus	The new doctor demonstrated that the melon appeared increasingly likely to succeed.	24
	Adj. Cl.	Plaus	Once the new chef started, the restaurant separated mediocre cooks from gifted ones.	24
	Adj. Cl.	Implaus	After the technician called, the smile stopped working almost immediately to his surprise.	24
	Pass. Rel. Cl.	Plaus	The patient who was refused the treatment continued causing uncomfortable scenes in the ER.	24
	Pass. Rel. Cl.	Implaus	The yoga instructor who was offered the beard demanded immense physical effort from everyone.	24
	Adj. Cl.	Plaus	After the esteemed reviewer reads, the book gains more attention due to his glowing praise.	18
	Adj. Cl.	Implaus	Even if the mother calls, her boys continue causing problems with the other kids on the playground.	18
	Sim. Trans. Cl.	Plaus	The suspect changed the file.	108
	Sim. Cl. w. Mod.	Plaus	The technician stopped working almost immediately after the argument.	81
	Sim. Cl. w. Mod.	Implaus	The tournaments remain essentially the same for the rest of the year.	18
	Intrans. Cl.	Plaus	The producer starts.	24
	Intrans. Cl.	Implaus	The dog hatched.	6
	Ditrans. Pass.	Plaus	The operator was brought the machine.	42
	Ditrans. Pass.	Implaus	The clerk was granted the finger.	6
	Trans. Cl.	Implaus	The cleaner ate the book.	15
	Mul. Mod.	Implaus	A prodigious profile quietly lay ahead of the unstoppable crowd.	11
Rich and Wagers (2020)	Passive	Plaus	The knife had been recently sharpened.	144
	Passive	Implaus	The shirt had been recently sharpened.	48
Ours	Simple	Plaus	The nurse fetched the patient.	10
	Simple	Plaus	The nurse fetched the intern.	40

Table 1: Breakdown of the data we used based on origin, syntactic structure, plausibility, and number of items, along with examples for each type. Emb. : Embedded, Obj.: Object, Quest.: Question, Decl.: Declarative, Adj.: Adjoined, Cl.: Clause, Pass: Passive, Rel.: relative, Sim.: Simple, Trans.: Transitive, Mod.: Modification, Mul.: Multiple

1. Chow et al. (2016): 60 sentence pairs from Experiment 1 in Chow et al. (2016), consisting of semantically plausible and implausible sentences with an embedded object question structure. Each sentence has 30 plausibility ratings, collected for a subsequent experiment.
2. Huang et al. (2023): 491 sentences from the Syntactic Ambiguity Processing benchmark (Huang et al., 2023), consisting of disambiguated garden-path sentences or parts of these sentences. Each sentence has 19.6 plausibility ratings on average.
3. Rich and Wagers (2020): 48 sets of 4 sentences each consisting of three semantically plausible and one semantically implausible sentences with a common syntactic structure. Each sentence has 10 plausibility ratings.
4. Our data: 50 plausible sentences with a simple syntactic structure, composed for a future experiment on similarity-based interference. These materials consist of 40 sentence pairs (one sentence is shared among 4 pairs). Each sentence has 40 plausibility ratings.

Table 1 showcases examples of sentences from the different datasets for each syntactic structure and plausibility variation that was tested. The table also includes the corresponding item counts for each sentence structure.

2.2 Models

We test the following LMs:

Closed-source models:

- **GPT-4** (OpenAI, 2023), a LLM released by OpenAI, available through an API.[1] This LM is widely considered to be one of the best existing LMs, if not the best (Bubeck et al., 2023).
- **ChatGPT** (GPT-3.5), a chat LLM released by OpenAI, available through an API
- **InstructGPT** (text-davinci-003) (Ouyang et al., 2022), an instruction-finetuned LLM released by OpenAI, available through an API

Open-source models: We also used several open-source models available on the HuggingFace Hub (Wolf et al., 2019), through the FastChat (Zheng et al., 2023) servers (allowing simulating the OpenAI API):

- **LLaMa** (Touvron et al., 2023), a foundation model released by Meta Research, trained on non-proprietary open-domain data.
- **Alpaca** (Taori et al., 2023), a model based on LLaMa, instruction fine-tuned based on instruction data generated by InstructGPT.
- **Vicuna** (Chiang et al., 2023), a model based on LLaMa, fine-tuned on chat data from ChatGPT, available through ShareGPT.[2]
- **Falcon-Instruct** (Almazrouei et al., 2023), based on the Falcon foundation model released by Abu Dhabi TII, fine-tuned on a mix of chat and instruction data.
- **StableLM**,[3] a model released by Stability AI, fine-tuned on instruction and chat data.
- **MPT Chat**,[4] a model based on MosaicML's MPT foundation model, finetuned on chat and instruction data.

We decode from the LMs by sampling with a temperature,which is set to 1.5 for closed-source models and 0.3 for open-source models.

2.3 Prompts

Our prompts start with an instruction for the LM to provide a plausibility score on a scale from 1 to 7 (see exact prompts in Appendix A). We then provide examples for plausibility judgements, which are either global and fixed across datasets, or specific for each dataset:

- **Global**: We provide four examples for each possible plausibility score (28 examples overall). Examples include a wide range of syntactic structures, inspired by the four datasets, but including additional structures.
- **Specific**: For each dataset, we provide three examples (21 overall) that illustrate syntactic structures that appear in this dataset.

Data	Best corr.	Model	Prompt	SH
Chow et al.	0.850	GPT-4	Glob.	0.943
Rich et al.	0.793	GPT-4	Glob.	0.868
Huang et al.	0.835	GPT-4	Glob.	0.898
Ours	**0.792**	GPT-3.5	Glob.	0.912
Chow et al.	**0.916**	GPT-4	Spec.	0.943
Rich et al.	**0.806**	GPT-4	Spec.	0.868
Huang et al.	**0.852**	GPT-4	Spec.	0.898
Ours	0.778	GPT-4	Spec.	0.912

Table 2: Highest Pearson correlation achieved for each of the datasets along with the split-half (SH) correlation analysis of human judgements, which provides an approximate upper bound. GPT-4 is the best LM in all cases, except for our dataset with a global prompt. In that case the correlation of GPT-4 is 0.761.

3 Results

Table 2 presents the highest Pearson correlation between average human and LLM ratings for each dataset and each prompt. The top half presents the highest correlation using the global prompt, whereas the bottom half uses the specific prompt. Additionally, the table includes the *split-half correlation* of human plausibility judgments, i.e, we randomly split human data in each example into two halves and measure the correlation between simulated sets of humans. This provides a rough upper bound on the correlation that can be achieved with a model.

Overall, The correlation of the highest-scoring model with human judgements is high, hovering around 0.8-0.9. Moreover, this correlation is typically just a few points under the split-half correlation.

Table 2 also shows that GPT-4 is a strong and robust baseline for human judgements, since it achieves the highest correlation in almost all the setups. When using our dataset with global prompts, the best model is GPT-3.5, where GPT-4 is slightly behind with a correlation of 0.761.

Finally, the results suggest an advantage to the specific prompt, with the highest correlation

[1] https://openai.com/blog/openai-api
[2] https://sharegpt.com/
[3] https://huggingface.co/stabilityai/stablelm-tuned-alpha-7b
[4] https://huggingface.co/mosaicml/mpt-7b-chat

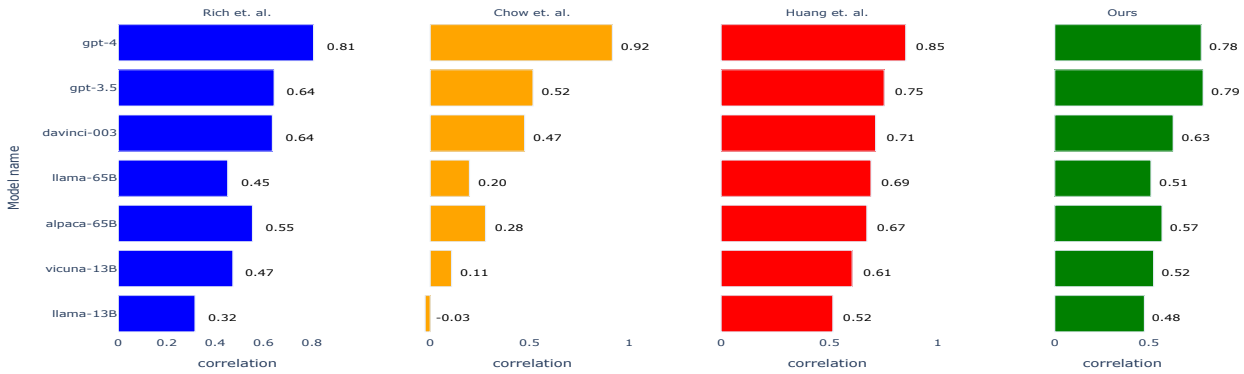

Figure 2: A breakdown of the correlation for the specific prompt for a subset of the models.

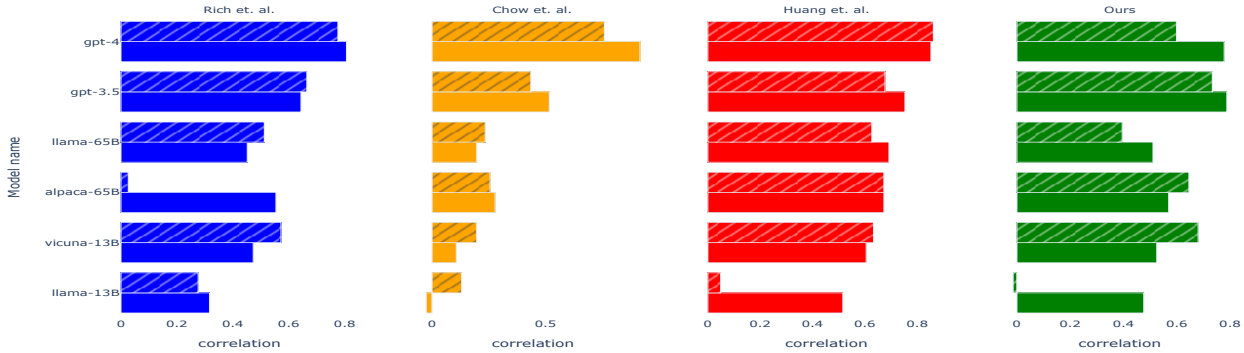

Figure 3: The correlation of the model that uses specific prompt when examples are included (full bar) versus when they are excluded (hatched bar).

achieved by prompts with examples resembling the judged sentences for almost all datasets.

Next, we will further analyse the performance of the different models and the importance of having examples in the prompt.

3.1 Model breakdown

Figure 2 shows the Pearson correlation with the specific prompt for 7 selected models across our 4 datasets (Results for all models and for the global prompt are provided in Appendix B).

First, as previously evidenced in Table 2, GPT-4 is a strong baseline, with a high correlation with human performance across all datasets. The other models from OpenAI also perform well, except on Chow et al. (2016) where a big drop in performance is noted for all the models that are not GPT-4. We conjecture that this is due to rarity of the syntactic structure of the sentences from Chow et al. (2016).

Figure 2 also shows that Alpaca and Vicuna have a better performance than LLaMa, their base model, at equivalent sizes, showing that instruction or chat fine-tuning improves correlation with human judgements.

Falcon-40B-Instruct is the best open source model, with performance comparable to text-davinci-003 model which is 4.5 times larger. Alpaca-65B, LlaMa-65B and Vicuna-13B also have a decent correlation with human judgements for the datasets with simple syntactic structures but perform poorly on data from Chow et al. (2016). The correlation of all the other open source models with human judgements is relatively low across all the datasets and is reported in Appendix B.

3.2 Importance of prompt examples

To analyze the importance of examples in the prompt, we ran experiments on a prompt that includes only the instruction, without examples, and compared its correlation to the correlation achieved with the specific prompt. Results for this experiment are in Figure 3.

Unsurprisingly, for most of the models and datasets, the prompt with examples has higher correlation with human judgments than the prompt without examples.

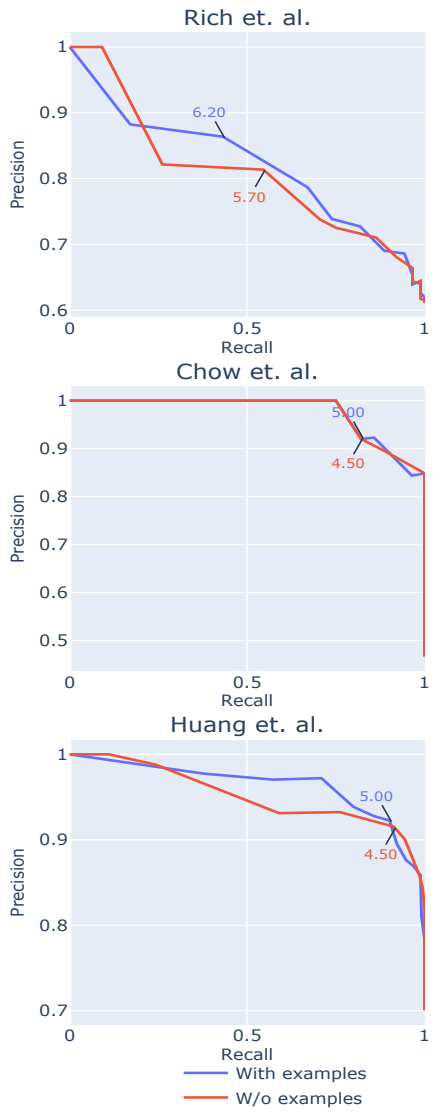

Figure 4: Recall-precision curve when filtering out implausible sentences. Blue is for the specific prompt, red is for the global prompt. We also mark for a few points the threshold value that results in a particular recall-precision result. For Chow et al. and Huang et al. we reach very high precision while keeping a large fraction of the sentences. For Rich et al. we can keep roughly half the sentences with precision of 0.8-0.9.

4 Methodology

In §3, we saw significant correlation between plausibility judgments of humans and GPT-4. We now evaluate directly the performance of LLM judgments when replacing human judgements. Plausibility judgements can be used in different ways for constructing experimental materials.[5]

[5]In some cases, judgements are not used to control experimental materials, but are rather entered as predictors in the analysis of the main experiment, accounting for some of the variability.

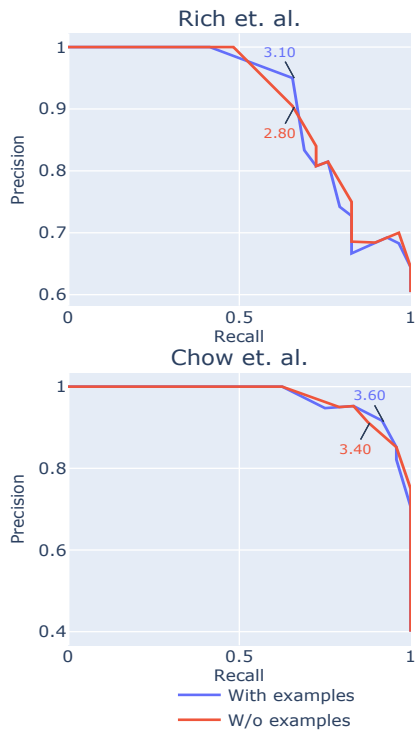

Figure 5: Recall-precision curve when filtering out plausible sentences. Blue is for the specific prompt, red is for the global prompt. We also mark for a few points the threshold value that results in a particular recall-precision result. In both setups, we can obtain very high precision while keeping most of the sentences.

Three common uses are: (a) filtering out implausible sentences by requiring a minimum average plausibility rating, (b) filtering out plausible sentences by requiring a maximum average plausibility rating, and (c) filtering out sentence pairs that have dissimilar average plausibility ratings. We evaluate the performance of LLMs across these operations.

4.1 Mapping LLM judgements to human judgements

We simulate using LLM judgements in two setups: (a) assuming no human ratings are collected, and (b) assuming a minimal amount of human ratings. We then evaluate the performance of LLMs with recall-precision curves, to see if we can achieve high precision (i.e, accepting only "good" sentences), while retaining high recall (i.e., keeping most of the 'good' sentences).

No human ratings: We collect LLM ratings from GPT-4 with the specific prompt. We then linearly map the LLM ratings into human ratings by fitting for every dataset a linear regression model

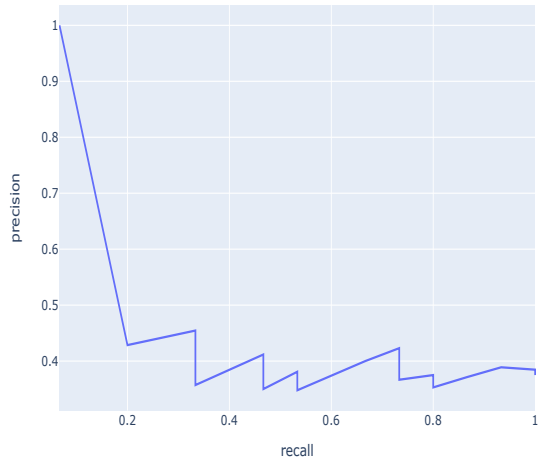

Figure 6: Recall-precision curve for classifying if a pair of sentences has different plausibility ratings.

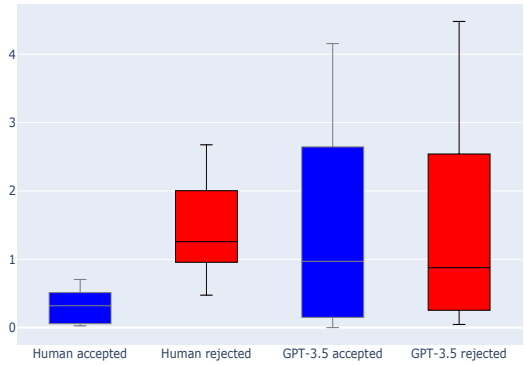

Figure 7: Difference between the average plausibility for pairs of sentences in our dataset. The blue boxes represent pairs that the t-test did not reject, the red represents pairs the t-test rejected.

on data from the other three datasets.

With human ratings: We assume access to a small amount of human ratings. Specifically, if D is the size of a dataset, we use human ratings for $\max(0.1 \cdot D, 15)$ sentences. Then, we collect LLM ratings with different OpenAI models and prompts and select the model and prompt combination that leads to the highest correlation with human ratings. We can also learn a linear map from LLM ratings to human ratings with this small amount of data.

4.2 Filtering out implausible sentences

The first pretest use we discuss is filtering implausible sentences by rejecting sentences under a given threshold (e.g. 5, as in Huang et al. (2023). We map LM ratings to human ratings with the linear regression model and then apply a threshold to filter out implausible sentences.[6]

Figure 4 shows recall-precision curves for the aforementioned datasets, varying the threshold for classifying a sentence as plausible (the positive class in the recall-precision curve is plausible sentences). Overall, GPT-4 exhibits high performance in this setup. For Chow et al. (2016) and Huang et al. (2023), we can achieve very high precision, while keeping most of the sentences. For Rich and Wagers (2020), performance is lower, but still we can cover roughly half the dataset with precision around 0.8-0.9. This aligns with the fact that this dataset has the lowest correlation with human judgments and includes rarer syntactic structures compared to the other two datasets.

4.3 Filtering out plausible sentences

The second pretesting scenario is the opposite of the first one – when the experiment requires implausible sentences, plausible sentences are filtered out by rejecting sentences with an average rating over some threshold (e.g. 3). We apply the same procedure for mapping LLM ratings to human ratings.

Figure 5 shows recall-precision curves for these datasets, varying the threshold for classifying a sentence as implausible (here the positive class are implausible sentences). We observe high performance overall, suggesting that predicting implausibility is easier than predicting plausibility.

4.4 Comparing plausibility of sentence pairs

The last pretest use we examine is comparing the plausibility of a pair of sentences and verifying that it is roughly similar. This is typically done by obtaining human ratings for both sentences, and running a t-test to check if the null hypothesis that they originate from the same underlying distribution is rejected, in which case the pair is filtered out.[7]

Using a t-test with LMs is non-trivial, because (as we discuss in §5) the variance in plausibility ratings for LMs is dramatically lower compared to humans, which in turn affects the t-test results. Instead, we propose to set a threshold for the difference between the average plausibility ratings of

[6] Since we evaluate with a recall-precision curve, the linear mapping is not necessary but is helpful for having the output label in a similar scale to humans.

[7] It is also possible to use cumulative link models (Taylor et al., 2021) to test the difference between sentences, but this is currently less common

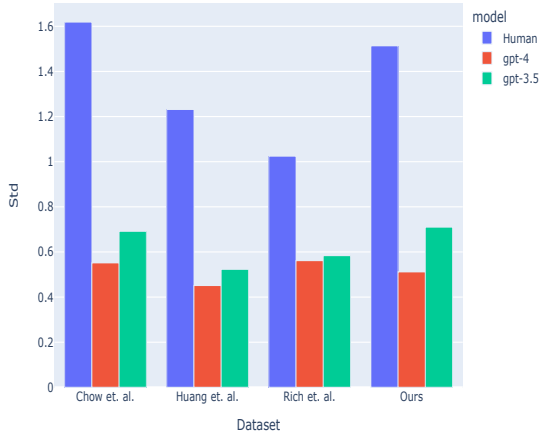

Figure 8: The average standard deviation of the judgements collected with the specific prompt for GPT-4 and GPT-3.5 along with the average standard deviation for human judgements.

the two sentences, and examine if there exists a threshold for which we can reject/accept the same sentence pairs that are rejected/accepted using t-test with human ratings. Specifically, we will draw a recall-precision curve, where the positive class are sentence pairs accepted according to the human rating t-test.

We apply this method for our dataset, using *GPT-3.5-Turbo* with the *global* prompt, which obtained the highest correlation with human judgements (0.792). We find the performance is low – we are unable to find a point on the recall-precision curve where precision is high and recall is substantial. Figure 6 shows the recall-precision curve, and as is evident, precision quickly drops to around 0.4-0.45, and the maximal F_1 obtained is 0.55, which is achieved when the difference between plausibility ratings is larger than 3.69.

To analyze this, we label each pair with its human-based gold label, and plot in Figure 7 the difference in average plausibility judgements for both humans and our LM. Clearly, the difference is a good discriminating feature for human ratings, but is a bad discriminating feature for the LM. This shows that while correlation between human ratings and LM ratings is high (0.792), it captures mostly coarse-grained structure, but is not powerful enough to make fine-grained distinctions like predicting if two sentences have the same level of plausibility. Moreover, when we measure the correlation between the difference in average plausibility ratings between humans and LMs, we find only a moderate Pearson correlation of 0.312.

5 Variance of Humans vs. LMs

Thus far, we saw that the average plausibility ratings of humans and LLMs correlate well. It is important to note that this is not the case w.r.t *variance*. Explicitly, human variance is much higher than the variance of LMs, despite the high temperature used for sampling, which is 1.5. Figure 8 shows the standard deviation for GPT-4 and GPT-3.5 on all the datasets when using the specific prompt, as well as the standard deviation for human judgements. Standard deviation for these LMs is dramatically lower than humans, i.e., we obtain relatively similar plausibility judgements when sampling multiple times from the model.

A possible theoretical explanation for this phenomenon is that the outputs of LMs can be viewed as an average over multiple samples, since pretraining is done on texts from many authors. Thus, when sampling plausibility ratings from a LM, we are sampling from an average of plausibility ratings. Let each human rating r_i be a sample from a distribution with mean μ and variance σ^2. We can view each sample from a LM as an average of N human ratings: $\frac{1}{N}\sum_{i=1}^{N} r_i$. This is a random variable with mean μ and variance $\frac{\sigma^2}{N}$. This observation can be used to estimate for a particular sentence what is the number N of humans that the LM is averaging over, by computing the ratio between the observed variance of humans and the observed variance of the LM for that sentence.

6 Conclusion

We investigate the correlation between plausibility judgements of humans and language models and find high correlation for simple syntactic structures overall, and high correlation throughout for GPT-4. We show language models can be used to provide coarse-grained plausibility judgements, which can reduce the cost of and accelerate psycholinguistic research. We view this work as a first step in this direction, where future work can improve the correlation through finetuning and prompt engineering and further investigate the utility of language models for conducting psycholinguistic research.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wing-Yee Chow, Cybelle Smith, Ellen F. Lau, and Colin Phillips. 2016. A "bag-of-arguments" mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31:577 – 596.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Lyn Frazier. 1987. Sentence processing: A tutorial review.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86, Online. Association for Computational Linguistics.

Markus J. Hofmann, Steffen Remus, Chris Biemann, Ralph R. Radach, and Lars Kuchinke. 2020. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2023. Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Richard Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29:375–419.

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Tal Ness and Aya Meltzer-Asscher. 2019. When is the verb a potential gap site? the influence of filler maintenance on the active search for a gap. *Language, Cognition and Neuroscience*, 34(7):936–948.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Stephanie Rich and Matt Wagers. 2020. Semantic similarity and temporal contiguity in subject-verb dependency processing. *Talk at the Human Sentence Processing conference*.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Nathaniel J. Smith and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Whitney Tabor and Sean Hutchins. 2004. Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):431.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model.

Jack Edward Taylor, Guillaume A. Rousselet, Christoph Scheepers, and Sara C. Sereno. 2021. Rating norms should be calculated from cumulative link mixed effects models. *Behavior research methods*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Prompt examples

We experimented with various prompts, some specific for the syntactic structure under study and one **global** prompt meant to range over a wide array of syntactic structures and be general enough to capture all of them. We also experimented with a prompt without examples. The instructions remain the same across the prompts; the only changed elements are the examples.
In all the showcased prompts we show only 1 example per score.

A.1 Global prompt

We created a prompt showcasing a variety of syntactic structures, in an attempt to create a general prompt that will be diverse enough to fit a large number of pretesting samples. There are at most 4 examples per score. Figure 9 shows an example of the prompt.

A.2 Prompt for our data

For our data, we wrote a prompt using the specific syntactic structure used in the materials. There are at most 3 examples per score. Figure 10 shows a prompt with 1 example per score.

A.3 Prompt for Chow et al.

For Chow et al. (2016) data, we wrote a prompt using the specific syntactic structure used in the materials. There are at most 3 examples per score. Figure 11 shows a prompt with 1 example per score.

A.4 Prompt for Huang et al.

For (Huang et al., 2023), given the wide array of syntactic structures present in the data, we covered the different types of syntactic structures in the examples for each of the scores. There are at most 3 examples per score. Figure 12 shows a prompt with 1 example per score.

B Full results

The correlation for all the models and the datasets are presented in Table 3.

```
You will read sentences and judge how natural they sound. You will need to judge,
on a scale from 1 to 7, how natural/plausible the presented sentence sounds, and
explain yourself shortly.
All presented sentences will be grammatically correct.
Important: you are encouraged to use the whole scale.

Here are some examples:

They spent their week-end at the beach, sipping iced tea.
The plausibility score is 6 (it is plausible that people would spend their week-end
at the beach).

The farmer planted the fruits from which the seeds came.
The plausibility score is 3 (it's more likely to plant seeds than fruits).

The table occupied most of the space in the kitchen.
The plausibility score is 5 (it is a somewhat plausible situation, maybe it is a
small kitchen).

Because he slept nine hours, he woke up completely exhausted.
The plausibility score is 1 (sleeping is not supposed to make you tired).

The policeman stopped the plane.
The plausibility score is 4 (it is a situation that might happen but is a bit unlikely).

The witness observed which policeman the robber had caught.
The plausibility score is 2 (in general, policemen catch robbers, not the other way
around).

I'm so thirsty, can you please pour me a glass of water?
The plausibility score is 7 (it is highly plausible that someone thirsty would like
to drink water).

The sentence:
The chef prepared the meal

The plausibility score is:
```

Figure 9: Example of a global prompt

```
You will read sentences and judge how natural they sound. You will need to judge,
on a scale from 1 to 7, how natural/plausible the presented sentence sounds, and
explain yourself shortly.
All presented sentences will be grammatically correct.
Important: you are encouraged to use the whole scale.

Here are some examples:

The librarian ordered the audio book.
The naturalness score is 5 (a librarian might order an audio book but in general
they order physical books)

The farmer bought a ski.
The naturalness score is 2 (it is an unnatural/implausible situation)

The handyman repaired the car.
The naturalness score is 3 (it is a somewhat unnatural, handymen repair things in
houses)

The barista prepared the cappuccino.
The naturalness score is 6 (it is likely that a barista would prepare a cappuccino)

The teacher scolded the shoe.
The naturalness score is 1 (it is really unnatural/implausible situation)

The policemen caught the thief.
The naturalness score is 7 (it is highly likely that policemen would try and catch a
thief)

The cook prepared the cocktail.
The naturalness score is 4 (a cook might prepare a cocktail but it is a bit unlikely)

The sentence: The nurse fetched the intern. The plausibility score is:
```

Figure 10: Example of a prompt for our data

```
You will read sentences and judge how natural they sound. You will need to judge, on a scale from 1 to 7, how natural/plausible the presented sentence sounds, and explain yourself shortly.
All presented sentences will be grammatically correct.
Important: you are encouraged to use the whole scale.

Here are some examples:

The director recalled which scene the editor had cut.
The plausibility score is 6 (it is plausible that a director knows which scene has been cut from the movie).

The tour guide guessed which landmark the visitor had photographed.
The plausibility score is 5 (it is relatively plausible that a tour guide might guess which landmark a tourist might photograph).

The detective identified which officer the suspect had recognized.
The plausibility score is 4 (suspects might know some police officer and recognize them)

The zoologist noted which lion the antelopes had hunted.
The plausibility score is 1 (lions hunts antelopes, not the other way around).

The journalist revealed which politician the lobbyist had influenced.
The plausibility score is 3 (it can happen that politicians influence lobbyists but it's supposed to be the other way).

The accountant knew which employee the CEO had promoted.
The plausibility score is 7 (it is highly plausible that an accountant would know who got promoted since he handles the money).

The pilote remembered which plane the airline had represented.
The plausibility score is 2 (planes represent airlines in general, not the opposite).

The sentence:
The park ranger documented which eagle the hunter had shot.

The plausibility score is:
```

Figure 11: Example of a prompt for Chow et al.'s data

```
You will read sentences and judge how natural they sound. You will need to judge, on a scale from 1 to 7, how natural/plausible the presented sentence sounds, and explain yourself shortly.
All presented sentences will be grammatically correct.
Important: you are encouraged to use the whole scale.

Here are some examples:

The firefighter who was denied the transplant went to the moon.
The plausibility score is 2 (people really rarely go to the moon).

The prison guard, which the inmate despised, robbed a bank.
The plausibility score is 4 (a prison guard robbing a bank might happen but is unlikely).

The firefighters put out the fire.
The plausibility score is 7 (it is really plausible, the role of firefighters is to put out fires).

The mechanic fixed the problematic cars with his eyes closed.
The plausibility score is 1 (it is highly unlikely that a mechanic can fix cars without seeing).

The teacher left.
The plausibility score is 5 (it is a somewhat plausible situation, maybe the class is over).

The fish ate the sponge.
The plausibility score is 3 (it is somewhat unlikely that a fish would eat a sponge but it might happen).

The scientist showed that the invention worked well.
The plausibility score is 6 (it is plausible that a scientist would show the efficiency of an invention).

The sentence:
The new chef started.

The plausibility score is:
```

Figure 12: Example of a prompt for Huang et al.

Model	Prompt	Chow et al.	Rich et al.	Huang et al.	Ours
GPT4	Specific	0.916	0.806	0.852	0.778
	Global	0.850	0.793	0.835	0.761
GPT3.5	Specific	0.517	0.644	0.753	0.788
	Global	0.481	0.703	0.794	0.792
Davinci-003	Specific	0.475	0.637	0.713	0.629
	Global	0.323	0.678	0.628	0.729
LlaMa-65b	Specific	0.197	0.452	0.692	0.511
	Global	0.130	0.608	0.634	0.641
Alpaca-65b	Specific	0.278	0.554	0.673	0.570
	Global	0.241	0.652	0.651	0.622
Falcon-40b	Specific	0.379	0.566	0.746	0.675
	Global	0.363	0.665	0.682	0.608
LlaMa-13b	Specific	-0.026	0.317	0.516	0.476
	Global	0.157	0.521	0.464	0.263
Vicuna-13b	Specific	0.107	0.473	0.605	0.525
	Global	0.185	0.582	0.612	0.575
Alpaca-13b	Specific	0.200	0.061	-0.005	-0.081
	Global	-0.140	0.057	-0.063	-0.021
LlaMa-7b	Specific	0.066	0.171	0.248	0.324
	Global	0.034	0.283	0.190	0.086
Vicuna-7b	Specific	0.021	0.313	0.478	0.359
	Global	0.072	0.473	0.496	0.336
Alpaca-7b	Specific	0.067	0.292	0.299	0.409
	Global	-0.043	0.375	0.275	0.430
Falcon-7b	Specific	0.148	0.237	0.238	0.358
	Global	0.167	0.317	0.207	0.203
Mpt-7b	Specific	0.111	0.331	0.395	0.432
	Global	0.034	0.314	0.350	0.455
StableLM-7b	Specific	0.006	0.157	0.000	-0.211
	Global	0.062	0.066	-0.102	-0.123

Table 3: Correlation for all the tested models on all of the datasets