HLB: Benchmarking LLMs' Humanlikeness in Language Use

Anonymous ACL submission

Abstract

As synthetic data becomes increasingly prevalent in training large language models (LLMs), concerns have emerged that these models may deviate from authentic human language patterns, potentially losing the richness and creativity inherent in human communication. This highlights the critical need to assess the language humanlikeness of LLMs in real-world usage. In this paper, we present a comprehensive **H**uman language **L**ikeness **B**enchmark (HLB)—a comprehensive evaluation of 20 LLMs using psycholinguistic experiments designed to probe core linguistic dimensions: phonology, lexical processing, syntax, semantics, and discourse. To contextualize model performance, we collected responses from over 2,000 human participants as a baseline and compared these to the outputs generated by the models.

For rigorous evaluation, we developed a coding algorithm that accurately identified language use patterns, enabling the extraction of response distributions for each task. By comparing the response distributions between human participants and LLMs, we quantified humanlikeness through distributional similarity. Our results reveal fine-grained differences in how well LLMs replicate human responses across various linguistic levels. Importantly, we found that improvements in other performance metrics did not necessarily lead to greater humanlikeness, and in some cases, even resulted in a decline. By introducing psycholinguistic methods to model evaluation, this benchmark offers the first framework for systematically assessing the humanlikeness of LLMs in language use (see Figure 19 for the leaderboard; Code and data will be released upon acceptance.)

1 Introduction

In recent years, large language models (LLMs) have made significant advancements. Models like OpenAI's GPT series and Meta's Llama family can generate human-like text, engage in coherent dialogues, and answer complex questions, often producing responses that are indistinguishable from those of humans in certain evaluations (Tsubota and Kano, 2024). Cai et al. (2024) conducted a systematic evaluation of human-like language use in models such as ChatGPT and Vicuna, demonstrating that LLMs closely replicate human language patterns in many aspects. However, despite these successes, questions remain about how accurately these models capture the deeper, nuanced patterns of human language use. In other words, the full extent of their similarity to human behavior remains unclear.

The importance of evaluating humanlikeness in language use is further underscored by the increasing reliance on synthetic data for model training, particularly in dialogue models. While synthetic data generation facilitates efficient scaling of model training, it raises concerns about models diverging from real-world human language patterns (del Rio-Chanona et al., 2024). Studies have shown that synthetic data can degrade model performance after retraining (Shumailov et al., 2024). This makes it imperative to assess the humanlikeness of LLMs rigorously across various aspects of language use, to ensure that models do not lose the diversity and richness of human language data.

To address this challenge, we introduce a psycholinguistic benchmark designed to provide a systematic and comprehensive evaluation of how closely LLMs align with human linguistic behavior.

Although numerous benchmarks and leaderboards have been developed to assess the performance of LLMs on downstream NLP tasks, they often fail to capture the finer, human-like qualities of language use. Current NLP benchmarks typically focus on task-based accuracy or performance (Lewkowycz et al., 2022; Zhou et al., 2023; Peng

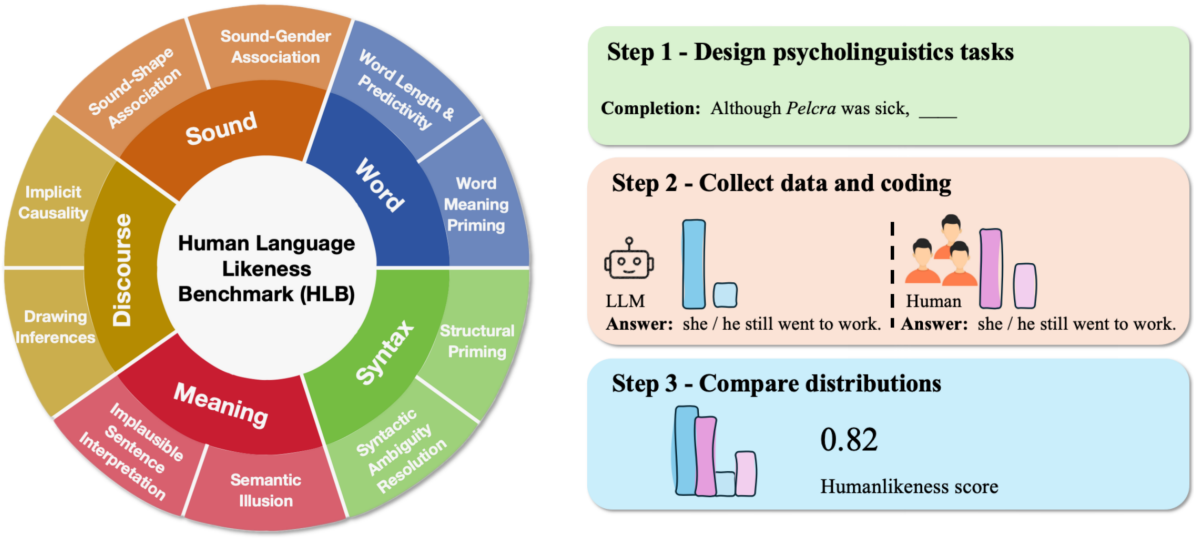


Figure 1: (a) Taxonomy of psycholinguistics experiments in HLB. The experiments and their sources are: sound-shape association (Köhler, 1967), sound-gender association (Cassidy et al., 1999), word length and predictivity (Mahowald et al., 2013), word meaning priming (Rodd et al., 2013), structural priming (Pickering and Branigan, 1998), syntactic ambiguity resolution (Altmann and Steedman, 1988), implausible sentence interpretation (Gibson et al., 2013), semantic illustion (Erickson and Mattson, 1981), implicit causality (Garvey and Caramazza, 1974), drawing inferences (Singer and Spear, 2015). (b) The benchmark framework. The example prompt is taken from the sound-gender association task, where humans can infer the gender of a novel name (e.g., *Pelcra* for Female; *Pelcrad* for Male) based on phonology.

et al., 2024; Hendrycks et al., 2021; Zellers et al., 2019), overlooking the broader psycholinguistic dimensions that characterize how humans process and produce language. Furthermore, few studies have systematically compared the language use of LLMs and human participants across multiple linguistic levels. This gap highlights the need for a new benchmark that can robustly measure the extent to which LLMs replicate human language behavior in real-world, diverse linguistic contexts.

In this paper, we address this gap by presenting a psycholinguistic benchmark study that evaluates the human language likeness of 20 LLMs. Our benchmark consists of 10 representative psycholinguistic experiments, which cover five core linguistic aspects: sound, word, syntax, semantics, and discourse, with two experiments dedicated to each aspect (see Figure 1). We collected approximately 50 to 100 responses per item from over 2,000 human participants. Additionally, we gathered 100 responses per item from each of the 20 LLMs, including well-known models such as GPT-4o, GPT-3.5, Llama 2, Llama 3, Llama 3.1, and other state-of-the-art models (see Table 2). To quantify human language likeness, we developed an auto-coding algorithm that efficiently and reliably extracts language use patterns from responses. The human language likeness metric was then calculated based on the similarity between the response distributions of humans and LLMs, using a comparison of their probability distributions.

Our findings reveal significant, nuanced differences in how LLMs perform across various linguistic aspects, offering a new benchmark for evaluating the humanlikeness of LLMs in natural language use. This benchmark introduces psycholinguistic methods to model evaluation and provides the first framework for systematically assessing the humanlikeness of LLMs in language use.

2 Related Work

Recent advances in LLMs have led to the development of various benchmarks designed to evaluate their linguistic capabilities. Standard benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) assess models across a range of natural language processing (NLP) tasks, including sentence classification, textual entailment, and question answering. However, these benchmarks primarily focus on task-based accuracy and often overlook the more intricate aspects of human-like language processing. While these evaluations provide valuable insights into model performance, they do not fully capture the extent to which LLMs

comprehend and generate language in a humanlike manner. As Manning et al. (2020) note, LLMs are powerful statistical models that can identify patterns in vast datasets, but these benchmarks do not adequately test how well models replicate human patterns of language use due to the interplay of complex cognitive biases.

2.1 Psychological Experimentation on LLMs

A growing body of research has begun applying classical psychological experiments to evaluate LLMs in more domain-specific and cognitively demanding tasks. For example, Binz and Schulz (2023) and Dasgupta et al. (2023) used well-known psychological paradigms, such as the Linda problem and the Wason selection task, to probe LLMs' abilities in judgment and decision-making. Similarly, Sap et al. (2023) and Trott et al. (2023) explored whether LLMs exhibit theory of mind, a key component of human social cognition, while Miotto et al. (2022) and Karra et al. (2023) examined LLMs' personality traits. In the domain of behavioral economics, Horton (2023) conducted experiments with GPT-3 to explore its decision-making processes. These studies suggest that LLMs can be treated as cognitive agents in psychological experiments, providing insights into how LLMs align with humans in reasoning, behavior, and decision-making. Moreover, they help shed light on the underlying mechanisms of LLMs, as seen in the work of Huang and Chang (2023)and Qiao et al. (2023), who analyzed reasoning patterns in LLMs. Hagendorff (2023) further provided a comprehensive review of LLM performance in psychological tests, showing that while LLMs demonstrate sophisticated behaviors, they often diverge from human cognition. These divergences highlight the need for more robust frameworks to understand the limitations of LLMs in mimicking human thought processes.

2.2 Psycholinguistic Experimentation on LLMs

Psycholinguistic approaches offer a deeper analysis by testing LLMs on how well they replicate the cognitive processes underlying human language processing. Ettinger (2020) and Futrell (2019) have subjected models like BERT to psycholinguistic tasks such as syntactic ambiguity resolution and structural priming, revealing both the strengths and limitations of LLMs in replicating human language processing. Michaelov and Bergen (2023) used structural priming tasks to investigate how LLMs internalize syntactic structures, while Huang et al. (2024) examined LLMs' ability to resolve syntactic ambiguity. Qiu et al. (2023) explored how well LLMs handle pragmatic reasoning. These studies demonstrate that LLMs can, to some extent, mimic humanlike behavior in controlled experiments. However, divergences in processing reveal the distinctions between machine learning models and humans. A recent review by Demszky et al. (2023) emphasized the need for benchmarks that incorporate psychological paradigms to evaluate LLMs. The authors argue that by applying psycholinguistic methods, researchers can better understand how closely LLMs approximate human cognition and where they fall short. Despite extensive research on LLMs' performance across various tasks, there is still no benchmark that includes human language processing data to reveal the extent to which LLMs resemble humans, particularly in language use. This paper addresses that gap by adapting 10 psycholinguistic experiments to evaluate how closely LLMs align with human language behavior, covering phenomena ranging from sound symbolism to discourse comprehension.

3 Methodology

3.1 Human Experiments

Experimental Design The human experiments were constructed using Qualtrics, an online survey platform (Qualtrics, 2024). The study included ten psycholinguistic tasks that spanned various linguistic levels, from sound, word, syntax, and meaning to discourse comprehension, with two experiments for each level (see Appendix A for details). We exposed a participant to only one trial on each task, with a total of 10 trials across all the tasks. This setup minimized trial-level effects and facilitated direct comparisons with LLMs, which were tested under similar conditions (presenting instructions and stimuli in a single prompt) to avoid context effects within individual conversations.

Procedure After providing consent, participants completed the ten psycholinguistic tasks (presented in a random order); four attention checks were randomly interspersed among the trials to later identify participants for random responding. Each experimental task began with an instructional screen, some of which included examples to clarify task requirements. The examples were carefully designed to differ from the experimental stimuli to

prevent potential priming effects. For instance, in a sentence-completion task, an illustrative example that did not resemble the experimental stimuli and did not induce target words for any stimuli was used. The priming tasks (which included pairs of priming and target stimuli) were spread across multiple pages to avoid strategic responses in case participants realise the relation between the prime and the target. The overall experimental procedure was streamlined for clarity and efficiency, with each session lasting approximately 8 to 10 minutes (mean = 8.336, *SD* = 4.171).

Participants Participants were recruited from the crowd-sourcing platform Prolific and restricted to native English speakers residing in the UK and US, according to their registration on the platform. They were required to use a desktop computer to complete the tasks. Among the 2,205 participants taking part in the experiments, 290 were excluded for not well adhering to the experimental instructions, including completing the study too quickly, showing low effort, or not finishing the experiment, according to the Qualtrics system. The remaining 1,915 participants were further checked for language nativeness and their accuracy with attention checks. After a thorough screening process—excluding those who were not native speakers, failed attention checks, or exhibited irregularities such as excessively short completion times or multiple participation attempts—the final valid sample consisted of 1,905 participants. The sample was composed of participants as follows: female (n = 1,051), male (n = 838), preferred not to disclose (n = 16), with an average age of 44.8 years (range: 18 to 89 years). Educational levels included: no formal education (n = 2), elementary school (n = 12), high school (n = 672), bachelor's degree (n = 862), and master's degree (n = 357). This sample of participants resulted in each item being tested in a minimum average of 24 trials (e.g., Word Length and Predictability) and up to an average of 96 trials (e.g., Sound-Shape Association Task).

3.2 LLM Experiments

Experimental Design To compare human responses with those generated by LLMs, we employed the same 10 psycholinguistic tasks designed for human participants. 20 LLMs (See Table 2) were selected for evaluation, including models from prominent families like OpenAI's GPT series (GPT-4o, GPT-3.5), Meta's Llama series (Llama 2, Llama 3, Llama 3.1) and Mistral series(OpenAI et al., 2024; Touvron et al., 2023; AI, 2024). Each model provided 100 responses per item in each experiment, ensuring that the response data was comparable to the human data. Similar to the human experimental design, LLMs followed a one-trial-per-run paradigm, ensuring that responses were generated independently for each item to prevent context effects. The input format for the LLMs closely mirrored the instructions provided to human participants. Careful modification of human prompts was performed to ensure that task instructions were clear and interpretable by LLMs. This allowed for a direct comparison between human and LLM performance on the same tasks under identical conditions.

Response Collection Procedure This closely mirror that in the human experiments. Each LLM was presented with the task instructions and the stimulus combined into a single prompt. We collected 100 responses (across different conditions) for each stimulus in an experiment in order to ensure a sufficiently large dataset for robust analysis of the response distributions. For OpenAI models, responses were obtained through the OpenAI API, while models hosted on Hugging Face were accessed using the Hugging Face Inference API. All requests to the models were made using their default parameters to encourage variability in responses. The collected responses were stored and processed for subsequent coding and analysis.

3.3 Response Coding

Development and Validation We employed an auto-coding algorithm across 10 experiments to assess agreement between human annotations and machine-generated labels. This algorithm utilized spaCy's *en_core_web_trf-3.7.3* model for syntactic parsing (e.g., structural priming and syntactic ambiguity resolution tasks) and regular expressions to detect answer patterns in others. Across 20,953 trials of human response data, we computed Cohen's Kappa (*kappa*), a measure that corrects for chance agreement between the results from manually coding and auto-coding algorithm, defined as:

$$K = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

where Po is the observed agreement, and Pe is the expected agreement by chance.

The Kappa score was $\varkappa = 0.993$, indicating near-perfect agreement ($z = 451$, $p < 0.001$). This demon-

Model	Overall	Sound		Word		Meaning		Syntax		Discourse	
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
Meta-Llama-3.1-70B-Instruct	**66.50**	**89**	62	61	6	81	77	**80**	67	**80**	63
Meta-Llama-3.1-8B-Instruct	65.89	73	65	60	12	84	**78**	79	**74**	79	56
Phi-3-mini-4k-instruct	64.61	61	68	59	19	**89**	71	76	48	**80**	76
Mistral-7B-Instruct-v0.1	62.77	73	**70**	62	24	87	43	69	36	79	**84**
gpt-4o	58.58	60	63	**68**	2	71	77	47	61	75	62
gpt-3.5-turbo	58.32	55	61	66	3	76	76	71	47	76	50
zephyr-7b-alpha	56.96	57	62	47	23	85	29	44	73	76	75
Mistral-8x7B-Instruct-v0.1	52.80	60	53	48	23	71	46	43	59	73	52
zephyr-7b-beta	47.85	28	53	48	**26**	71	7	38	73	75	60

Table 1: Language human-likeness scores for selected models across 10 experiments. **Bold** values indicate task-wise best performances among all models; Underlined values indicate the weakest. Full data for 20 models are provided in Table 2.

strates the high accuracy of the auto-coding algorithm in replicating human annotations.

3.4 Humanlikeness Scoring

To quantify the humanlikeness of LLM responses, we used Jensen-Shannon (JS) divergence to compare the response distributions between human participants and LLMs. JS divergence, a symmetric measure of similarity between two probability distributions, is ideal for assessing how closely LLM responses mirror human behavior across linguistic levels. For each task, the auto-coding algorithm generated response distributions for both humans and LLMs. We computed **language human-likeness score (HS)** for each item as:

$$\begin{aligned} HS_{\text{item}} &= 1 - JS(P, Q) \\ &= 1 - \frac{1}{2}\left[KL(P \parallel M) + KL(Q \parallel M)\right] \end{aligned} \tag{2}$$

where P and Q are the human and LLM response distributions, and M is their average. For each experiment, we average the scores across all items. The overall language human-likeness score across all experiments is then computed as:

$$\begin{aligned} HS_{\text{Overall}} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\left(1 - \frac{1}{2}\left[\text{KL}(P_i \parallel M_i)\right.\right.\right. \\ \left.\left.\left. +\text{KL}(Q_i \parallel M_i)\right]\right)\right) \end{aligned} \tag{3}$$

The HS metric offers a significant advantage over traditional metrics by capturing not just average response differences but also the distributional similarities between human and model responses, making it more sensitive to subtle divergences in how LLMs mimic human behavior. See the case analysis section for a detailed explanation of how this metric captures distributional differences effectively.

4 Result

4.1 Overall Performance

We begin by providing an overview of the experimental results for the 20 LLMs and human participants across the 10 psycholinguistic tasks. Detailed experiment result data for each task are displayed in Figure 18 and plotted in Figure 8 to 17 in Appendix C. From these results, it was observed that Meta-Llama-3.1-70B-Instruct consistently exhibited minimal deviation from human results, both in terms of mean values and effect sizes (i.e., differences in mean values between conditions). In contrast, models such as Mistral-7B-Instruct-v0.3 displayed notable divergence from human results across several tasks. These results provide a foundational overview for the subsequent language human-likeness score (HS) calculations.

The language language human-likeness scores (HS) revealed significant variation in how well LLMs emulated human language use across the 10 psycholinguistic experiments (see Table 1; Table 2 in Appendix D for the complete leaderboard). Meta-Llama-3.1-70B-Instruct led both in overall humanlikeness and across several individual tasks, followed by Meta-Llama-3.1-8B-Instruct. On the other hand, Mistral-7B-Instruct-v0.2 scored lower among the models, with Zephyr-7B-beta receiving the lowest score.

4.2 Model Family Divergence in Humanlikeness

Statistical comparisons across model families reveal substantial and systematic differences in their alignment with human language behavior, as quan-

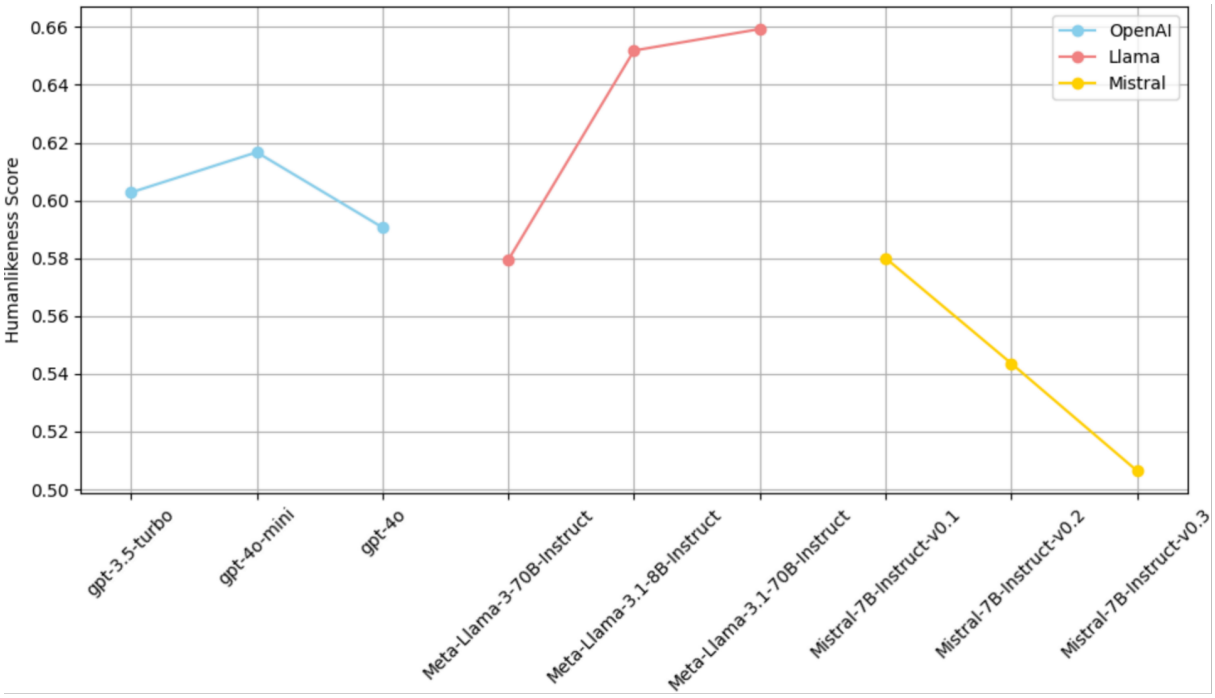


Figure 2: Language human-likeness scores of three LLM families

tified by language human-likeness scores (see Figure 3). Among the most striking findings, Llama models significantly outperformed Mistral models in humanlikeness (t = 10.44, p < .001), highlighting a pronounced divergence between these two architectural families. Although Llama models also surpassed OpenAI models (t = 3.13, p = .002), the magnitude of this difference was notably smaller, suggesting that the performance gap between Llama and Mistral models is particularly robust.

Intra-family comparisons further illuminate the developmental trajectories of these models. Within the Llama family, the transition from Meta-Llama-3-70B-Instruct to Meta-Llama-3.1-70B-Instruct yielded a significant improvement in humanlikeness (t = –4.85, p < .001), reflecting meaningful gains in language pattern alignment likely attributable to training data refinements or architectural tuning. By contrast, OpenAI's GPT-3.5-turbo and GPT-4o did not differ significantly in performance (t = –0.93, p = 0.352), indicating a plateau in their alignment with human responses across the tasks evaluated. Within the Mistral family, however, a significant performance decline was observed from Mistral-7B-Instruct-v0.1 to v0.3 (t = 5.45, p < .001), raising questions about the effect of model updates on cognitive fidelity.

This cross-family analysis represents one of the most consequential findings of the present study. It underscores that not all LLMs progress uniformly in approximating human-like language use and that family-level design choices, training objectives, and versioning strategies can exert a marked influence on cognitive alignment. The superior performance of the Llama models across both inter- and intra-family comparisons highlights them as the current frontier in capturing fine-grained human linguistic patterns, while also reinforcing the utility of language human-likeness scores as a diagnostic tool for comparative evaluation.

4.3 Challenges in Aligning Semantic Interpretation Patterns

A closer examination of individual experiments further underscores the nuanced challenges LLMs face in replicating human-like language behavior. Among the ten tasks, Experiment 4, which assessed word meaning priming, yielded the most pronounced divergence between model and human responses, as indicated by a substantial statistical difference (t = –116.32, p < .001).

This experiment was designed to evaluate whether models, like humans, show sensitivity to recent contextual exposure when interpreting an ambiguous word such as post. Specifically, participants were presented with sentences that either primed the target meaning directly (e.g., post used

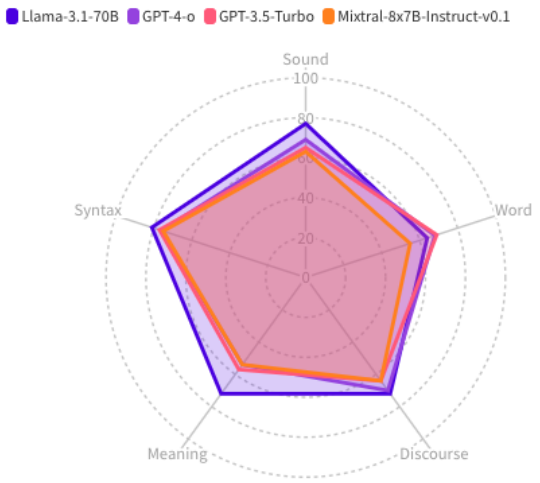


Figure 3: Language human-likeness scores of three LLM families

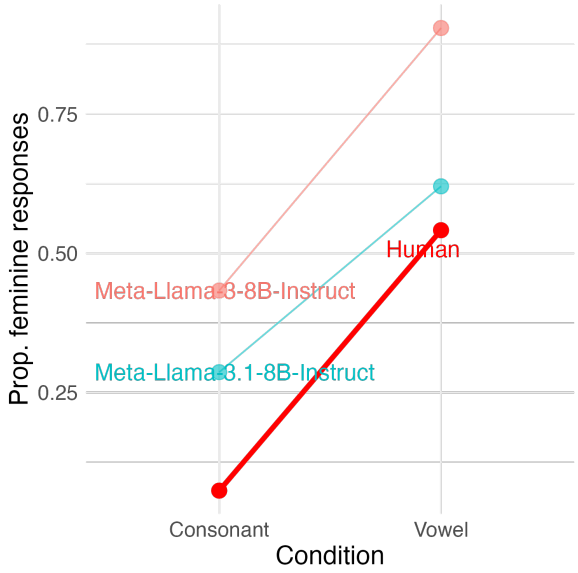


Figure 4: Experiment results of Llama-3-8B-instruct, Llama-3.1-8B-instruct and humans in E2 (sound-gender association)

as a job title) or through a synonymous context. Human participants demonstrated a relatively modest priming effect: 20% interpreted post with its job-related meaning following a word-meaning prime, compared to 18% following a synonym prime. These small differences reflect the graded and flexible nature of human semantic activation.

In contrast, Meta-Llama-3.1-70B exhibited an exaggerated priming response, with 52% and 38% selecting the job-related meaning in the respective conditions. While directionally consistent with the human pattern, the model's magnitude of response was disproportionately large, suggesting a rigidity or over-sensitivity in semantic activation that departs from human interpretive subtlety.

This case highlights a core challenge for LLMs: aligning with the probabilistic and contextually modulated nature of human semantic processing, especially in the face of lexical ambiguity. It further demonstrates that even when models approximate human behavior at a coarse level (e.g., showing a priming effect in the correct direction), the granularity and distribution of responses often reveal critical mismatches. Such discrepancies underscore the importance of evaluating not just outcome alignment but also the underlying response dynamics that shape human-like interpretation.

4.4 Distributional Divergence Beyond Effect Size

The E2 sound-gender association task provides a clear case study for illustrating the diagnostic value of the language human-likeness score, particularly in capturing divergences that are obscured when comparing only mean-level effects. Specifically, the contrast between Meta-Llama 3-8B-Instruct and Meta-Llama 3.1-8B-Instruct (hereafter Llama 3-8B and Llama 3.1-8B) reveals how similar effect sizes can mask substantial differences in response distributions—differences that are critical for assessing alignment with human linguistic behavior.

As shown in Figure 4, both models—along with human participants—exhibit a higher probability of classifying vowel-ending pseudonames as female, consistent with documented phonological biases. The difference in mean response between vowel and consonant conditions (i.e., effect size) is actually more similar between Llama 3-8B (difference = 0.47) and humans (0.41) than between Llama 3.1-8B (0.33) and humans. Nevertheless, Llama 3-8B received a substantially lower language human-likeness score (53) than Llama 3.1-8B (65). This counterintuitive result stems from distributional divergence: although Llama 3-8B matches human effect size more closely, its response distribution pattern deviates more markedly from human behavior than that of Llama 3.1-8B. The language human-likeness score, computed via JS divergence, captures these subtler discrepancies in response structure, not just aggregate values. By evaluating full response distribution, the score provides a more precise and meaningful assessment of how closely language models approximate human-like language behavior.

5 Discussion

The results of this benchmark study highlight notable differences in how LLMs approximate human language use across various linguistic levels. The Llama family of models, particularly Meta-Llama-3.1-70B-Instruct, consistently outperformed both the OpenAI and Mistral models in terms of language human-likeness score. This finding suggests that recent advancements in the Llama models have led to more humanlike language behaviors, especially in terms of semantic and discourse processing. The OpenAI models, including GPT-4o and GPT-3.5-turbo, showed relatively stable performance across tasks, with no significant differences between the models. This stability may reflect a plateau in the improvement of humanlikeness in these models, as compared to the more recent gains observed in the Llama family. On the other hand, the Mistral models demonstrated a decrease in language human-likeness scores, particularly in the transition to Mistral-7B-Instruct-v0.3. This suggests that certain training methods and data quality in Mistral may have reduced their alignment with human language patterns. One of the key insights from this study is that models differ not only in their overall language human-likeness scores but also in how they handle specific linguistic phenomena. For instance, in Experiment 4 (word meaning priming), we observed a significant divergence in resposnes between humans and LLMs, with the latter showing a much larger priming effect. This over-priming suggests that while LLMs may excel in certain aspects of language generation, they often lack the subtle flexibility that humans display when processing ambiguous or context-dependent language. A major strength of this study is its use of psycholinguistic experiments to evaluate LLMs, which goes beyond traditional NLP benchmarks that focus on task accuracy. By systematically probing various linguistic levels—sound, word, syntax, semantics, and discourse—this benchmark provides a more comprehensive understanding of how LLMs process and generate language.

6 Conclusion

In this paper, we introduced a novel benchmark for evaluating the humanlikeness of LLMs in language use based on psycholinguistic experiments. Our study evaluated 20 LLMs, including OpenAI's GPT family, Meta's Llama family, the Mistral family and others, across 10 experiments that spanned key linguistic aspects such as sound, word, syntax, semantics, and discourse. Using responses from over 2,000 human participants as a baseline, the results revealed significant differences in model performance, with Llama models consistently outperforming both OpenAI and Mistral models in terms of language use humanlikeness. These findings underscore the potential of psycholinguistic benchmarks to capture aspects of language that are often missed by traditional NLP evaluations.

This benchmark provides a framework for future research on LLMs, offering a more meaningful and comprehensive way to evaluate their performance in real-world language use. It also highlights areas where current LLMs diverge from human language patterns, particularly in tasks involving semantic priming and ambiguity resolution. By identifying these gaps, this study offers critical insights for the next generation of LLM development, paving the way for models that more closely mirror the intricacies of human communication.

7 Limitation

However, there are several limitations to this study. First, while the benchmark covers a wide range of linguistic tasks, it may not encompass the full complexity of human language use. Some linguistic phenomena, such as pragmatic reasoning, were not explored in this study. Second, we did not manipulate models' parameters, particularly the temperature or top k, to control the diversity of the generated responses. While using default parameters, particularly temperature, may seem limiting, this choice ensures that we evaluate models in their most typical and practical configurations. Default settings reflect how these models are commonly used in real-world applications, offering a fair and standardized comparison. Tuning parameters like temperature could introduce bias and variability across models, making it difficult to ensure consistent evaluation. By using default settings, we eliminate these concerns, allowing for a more reliable assessment of humanlikeness. Finally, while the study includes a large sample of human participants, the specific demographic characteristics (e.g., native English speakers from the UK and US) may not fully represent global language use patterns. Compared to previous benchmarks that focus on task-based performance, this study offers a more in-depth analysis of language models' alignment with human linguistic behavior. Similar studies,

such as Ettinger (2020), have used psycholinguistic principles to probe LLMs, but our study stands out by incorporating a broader range of linguistic levels and by using a large-scale dataset of human responses for direct comparison. The significant differences found between model families, such as the higher humanlikeness of Llama models, provide valuable insights for the ongoing development and fine-tuning of LLMs.

References

Mistral AI. 2024. Mistral-7b-instruct-v0.3: An advanced instruction-based language model. Hugging Face Model Card. Released on May 22, 2024. Available at: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120. Publisher: Proceedings of the National Academy of Sciences.

Zhenguang G. Cai, Xufeng Duan, David A. Haslett, Shuqi Wang, and Martin J. Pickering. 2024. Do large language models resemble humans in language use? *arXiv preprint*. ArXiv:2303.08014 [cs].

Kimberly Wright Cassidy, Michael H. Kelly, and Lee'at J. Sharoni. 1999. Inferring gender from name phonology. *Journal of Experimental Psychology: General*, 128(3):362–381.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks. *arXiv preprint*. ArXiv:2207.07051 [cs].

R Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2024. Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*, page pgae400.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701. Publisher: Nature Publishing Group.

TD Erickson and ME Mattson. 1981. From words to meaning: A semantic illusion. j verbal learn verbal behav 20 (5): 540–551.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

R Futrell. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguistic inquiry*, 5(3):459–464.

Edward Gibson, Leon Bergen, and Steven T Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Thilo Hagendorff. 2023. Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv preprint*. ArXiv:2303.13988 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. *arXiv preprint*. ArXiv:2212.10403 [cs].

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2023. Estimating the Personality of White-Box Language Models. *arXiv preprint*. ArXiv:2204.12000 [cs].

Wolfgang Köhler. 1967. Gestalt psychology. *Psychologische forschung*, 31(1):XVIII–XXX.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving Quantitative Reasoning Problems with Language Models. *arXiv preprint*. ArXiv:2206.14858 [cs].

Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Publisher: Proceedings of the National Academy of Sciences.

James A. Michaelov and Benjamin K. Bergen. 2023. Emergent inabilities? Inverse scaling over the course of pretraining. *arXiv preprint*. ArXiv:2305.14681 [cs].

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? An Exploration of Personality, Values and Demographics. *arXiv preprint*. ArXiv:2209.14338 [cs].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A Multilingual Code Generation Benchmark for Cross-lingual Natural Language Generalization. *arXiv preprint*. ArXiv:2402.16694 [cs].

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with Language Model Prompting: A Survey. *arXiv preprint*. ArXiv:2212.09597 [cs].

Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. 2023. Pragmatic Implicature Processing in ChatGPT.

Qualtrics. 2024. Qualtrics and all other qualtrics product or service names are registered trademarks or trademarks of qualtrics. Provo, UT, USA.

Jennifer M Rodd, Belen Lopez Cutrin, Hannah Kirsch, Alessandra Millar, and Matthew H Davis. 2013. Long-term priming of the meanings of ambiguous words. *Journal of Memory and Language*, 68(2):180–198.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2023. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. *arXiv preprint*. ArXiv:2210.13312 [cs].

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Murray Singer and Jackie Spear. 2015. Phantom recollection of bridging and elaborative inferences. *Discourse Processes*, 52(5-6):356–375.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do Large Language Models know what humans know? *arXiv preprint*. ArXiv:2209.01515 [cs].

Yuka Tsubota and Yoshinobu Kano. 2024. Text Generation Indistinguishable from Target Person by Prompting Few Examples Using LLM. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 13–20, Tokyo, Japan. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv preprint*. ArXiv:1905.07830 [cs].

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification. *arXiv preprint*. ArXiv:2308.07921 [cs] version: 1.

A Details of the Ten Psycholinguistic Experiments

This section introduces the ten psycholinguistic experiments used to evaluate the humanlikeness of LLMs across multiple linguistic levels. Each experiment was designed to test a specific linguistic phenomenon and compare the performance of LLMs to human participants.

Sounds: sound-shape association People often associate specific sounds with certain shapes, a phenomenon known as sound symbolism. We tested whether LLMs, like humans, tend to link spiky-sounding words (words consist of /i/, /eɪ/, /ə/ as vowels and /p/, /t/, /k/ as consonants, such as *takete* or *kiki*) with spiky objects and round-sounding words (words consist of /o/, /u/ as vowels and /b/, /g/, /l/, /m/, /n/, /w/ as consonants, like *maluma* or *baamoo*) with round objects.

Sounds: sound-gender association People can often guess if an unfamiliar name is male or female based on its sound. In English, women's names more frequently end in vowels compared to men's names. In this task, we asked participants to complete a preamble containing either a consonant-ending name (e.g., *Pelcrad* in 1a) or a vowel-ending novel name (e.g., *Pelcra* in 1b).

1a. Consonant-ending name: *Although Pelcrad was sick...*

1b. Vowel-ending name: *Although Pelcra was sick...*

Words: word length and predictivity Shorter words are suggested to make communication more efficient by carrying less information. If both humans and LLMs are sensitive to the relationship between word length and informativity, they should prefer shorter words over longer ones with nearly identical meanings when completing sentence preambles that predicted the meaning of the word (making it less informative; e.g., 2a), compared to neutral sentence preambles (e.g., 2b)

2a. Predictive context: *Susan was very bad at algebra, so she hated... 1. math 2. mathematics*

2b. Neutral context: *Susan introduced herself to me as someone who loved... 1. math 2. mathematics*

Words: word meaning priming Many words have multiple meanings; for instance, *post* can refer to mail or a job. People update an ambiguous word's meaning based on recent exposure. We tested whether humans and LLMs similarly demonstrate word meaning priming phennomenon: Participants associated post with its job-related meaning more frequently after reading sentences using that context rather than synonyms' contexts (3a vs.3b).

3a. Word-meaning prime: *The man accepted the post in the accountancy firm.*

3b. Synonym prime: *The man accepted the job in the accountancy firm.*

Syntax: structural priming In structural priming, people tend to repeat syntactic structures they've recently encountered. We had participants complete prime preambles designed for either PO (prepositional-object dative structure, e.g., *The racing driver gave helpful mechanic wrench* to complete 4a) or DO (double-object dative structure, e.g., *The racing driver gave torn overall his mechanic* to complete 4b). Participants then completed target preamble which could be continued as either DO/PO. If structural priming is demonstrated, participants replicate structure of the prime preamble.

4a. DO-inducing prime preamble: *The racing driver showed the helpful mechanic ...*

4b. PO-inducing prime preamble: *The racing driver showed the torn overall ...*

4c. Target preamble: *The patient showed ...*

Syntax: syntactic ambiguity resolution The way people parse words into syntactic structures has garnered significant attention in psycholinguistics. For instance, in VP/NP ambiguity (e.g., *The ranger killed the poacher with the rifle*), people usually interpret the ambiguous prepositional phrase (PP, *with the rifle*) as modifying the verb phrase (VP, *killed the poacher*) rather than the noun phrase (NP, *the poacher*). However, contextual information can modulate this resolution: People are more likely to interpret ambiguous PPs as modifying NPs when there are multiple possible referents (e.g., 5b) compared to when there is only a single referent (e.g., 5a). We examine how effectively LLMs use contextual information to resolve syntactic ambiguities and exhibit such modulation patterns.

5a. Single referent: *There was a hunter and a poacher. The hunter killed the dangerous poacher with a rifle not long after sunset. Who had a rifle, the hunter or the poacher?*

5b. Multiple referents: *There was a hunter and two poachers. The hunter killed the dangerous poacher with a rifle not long after sunset. Who had a rifle, the hunter or the poacher?*

Meaning: implausible sentence interpretation Listeners often need to recover intended messages from noise-corrupted input. Errors in production

or comprehension can make a plausible sentence implausible by omitting (e.g., *to* omitted, 6a) or inserting words (e.g., *to* inserted, 6b). People may interpret an implausible sentence nonliterally if they believe it is noise-corrupted. who found that people more frequently reinterpret implausible DO sentences than PO sentences due to the likelihood of omissions over insertions. We tested whether people and LLMs similarly assume that implausible sentences result from noise corruption, with omissions being more likely than insertions.

6a. Implausible DO: *The mother gave the candle the daughter.*

6b. Implausible PO: *The mother gave the daughter to the candle.*

6c. Question: *Did the daughter receive something/someone?*

Meaning: semantic illusions People often overlook obvious errors in sentences. For instance, when asked (7a), many fail to notice that the question should refer to *Noah* instead of *Moses*. Such semantic illusions suggest that processing sentence meanings involves partial matches in semantic memory. We tested whether LLMs and people alike produce semantic illusions and are more likely to catch a weak imposter (e.g., *Adam*, less similar to *Noah*, 7b) than a strong imposter (e.g. *Morse*, more similar to *Noah*, 7a).

7a. Strong: *During the Biblical flood, how many animals of each kind did Moses take on the ark?*

7b. Weak: *During the Biblical flood, how many animals of each kind did Adam take on the ark?*

Discourse: implicit causality Certain verbs prompt people to associate causality with either the subject or the object within a sentence. For instance, stimulus-experiencer verbs like scare typically lead people to attribute causality to the subject (e.g., completing 8a as *Gary scared Anna because he was violent*), whereas experiencer-stimulus verbs like fear generally lead people to attribute causality to the object (e.g., completing 8b as *Gary feared Anna because she was violent*). We assessed whether LLMs, like humans, show similar patterns of causal attribution based on verb type.

8a. Stimulus-experiencer verb: *Gary scared Anna because...*

8b. Experiencer-stimulus verb: *Gary feared Anna because...*

Discourse: drawing inferences People make bridging inferences more frequently than elaborative inferences. Bridging inferences connect two pieces of information (after reading 9a, people infer that Sharon cut her foot) while elaborative inferences extrapolate from a single piece of information (people are less likely to make this inference after reading 9b). We examined how well an LLM aligns with human patterns of inference by comparing the bridging and elaborative conditions.

9a. Bridging: *While swimming in the shallow water near the rocks, Sharon stepped on a piece of glass. She called desperately for help, but there was no one around to hear her.*

9b. Elaborative: *While swimming in the shallow water near the rocks, Sharon stepped on a piece of glass. She had been looking for the watch that she misplaced while sitting on the rocks.*

Question: *Did she cut her foot?*

B Results of Human Participants Across Psycholinguistic Experiments

Across 10 classic psycholinguistic and cognitive experiments, human participants consistently replicated well-established effects, as indicated by mixed-effects logistic regression models (see also Figures 5, 6, 7.)

In Experiment 1, participants were more likely to match round-sounding pseudowords to round shapes than to spiky shapes (Round: 0.83 vs. Sharp: 0.37, β = 2.28, *SE* = 0.28, z = 8.02, $p < .001$), demonstrating robust sound-shape symbolism.

In Experiment 2, vowel-ending names were more likely to be interpreted as female than consonant-ending names (Vowel: 0.54 vs. Consonant: 0.07, β = 3.39, *SE* = 0.64, z = 5.32, $p < .001$), replicating the classic sound-gender association.

In Experiment 3, participants showed a preference for using shorter words in predictive contexts than in neutral ones (Predictive: 0.40 vs. Neutral: 0.31, β = 0.57, *SE* = 0.18, z = 3.14, p = .002), consistent with the principle of communicative efficiency.

In Experiment 4, participants selected meanings congruent with primed word meanings more often than with semantic associations alone (Word Meaning: 0.27 vs. Semantic: 0.24, β = 0.28, *SE* = 0.14, z = 1.97, p = .049), reflecting a subtle effect of lexical priming.

In Experiment 5, a structural priming effect was observed: participants were more likely to produce prepositional object (PO) constructions after PO primes than after double object (DO) primes (PO: 0.65 vs. DO: 0.48, β = 0.73, *SE* = 0.15, z = 4.95, $p < .001$).

In Experiment 6, participants more often interpreted ambiguous phrases as noun phrases (NP) following plural contexts compared to singular ones (Plural: 0.16 vs. Single: 0.11, β = 0.70, *SE* = 0.17, z = 4.20, $p < .001$), supporting prior findings in syntactic ambiguity resolution.

In Experiment 7, no significant difference was observed in interpretation of implausible sentences across syntactic structures (PO: 0.53 vs. DO: 0.50, β = –0.12, *SE* = 0.36, z = –0.33, p = .741), suggesting limited sensitivity to structure in this context.

In Experiment 8, participants more frequently overlooked semantic inconsistencies when the keyword was weak versus strong (Weak: 0.74 vs. Strong: 0.61, β = 0.70, *SE* = 0.13, z = 5.21, $p < .001$), demonstrating the classic semantic illusion effect.

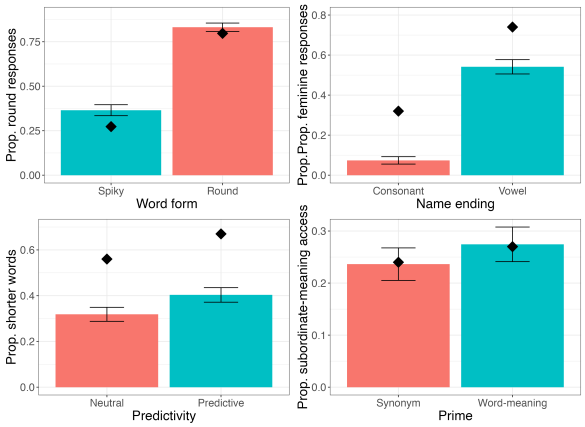


Figure 5: Results of Experiments 1–4. Bar plots show the proportion of target-consistent responses in each condition across participants, with error bars representing ±1 *SE*. (a) Participants more often matched round-sounding pseudowords to round shapes than to spiky shapes (Experiment 1). (b) Vowel-ending names were more frequently judged to be feminine than consonant-ending names (Experiment 2). (c) Shorter words were more likely in predictive contexts than in neutral ones (Experiment 3). (d) Word-meaning primes increased access to subordinate word meanings compared to semantic primes (Experiment 4). Black diamond markers indicate the proportion reported in the original studies. Our replication used a one-trial-per-run design, which deviates from the original multiple-trials-per-run setup. This adjustment minimizes potential context effects and offers a more LLM-compatible benchmarking format.

In Experiment 9, subject vs. object pronoun continuation strongly depended on verb type: participants overwhelmingly chose object continuations following stimulus-experiencer (SE) verbs and subject continuations following experiencer-stimulus (ES) verbs (ES: 0.93 vs. SE: 0.14, β = 25.50, *SE* = 1.31, z = 19.47, $p < .001$), consistent with implicit causality patterns.

In Experiment 10, bridging inferences were more likely to be made in bridging contexts than in elaborative ones (Bridging: 0.48 vs. Elaborative: 0.28, β = 0.98, *SE* = 0.19, z = 5.22, $p < .001$), confirming sensitivity to discourse structure in inference making.

C Experiment Results of Models and Human

Figures 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17 showed the results of the 20 models in 10 experiments, together with human results for comparison.

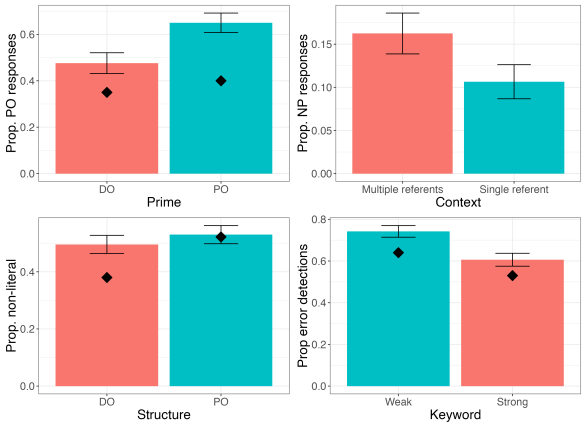


Figure 6: Results of Experiments 5–8.Bar plots show the proportion of structure- or meaning-consistent responses in each condition, with error bars representing ±1 *SE*. (a) Structural priming: Participants more often produced PO structures following PO primes than DO primes (Experiment 5). (b) Participants showed a preference for NP interpretation following plural noun contexts (Experiment 6). (c) Sentence interpretation was not significantly influenced by syntactic structure in implausible constructions (Experiment 7). (d) Semantic illusions occurred more frequently when critical words were weakly associated (Experiment 8).Black diamond markers reflect the originally reported values. As in all experiments, the one-trial-per-run format was used to reduce context sensitivity and align with LLM evaluation conditions.

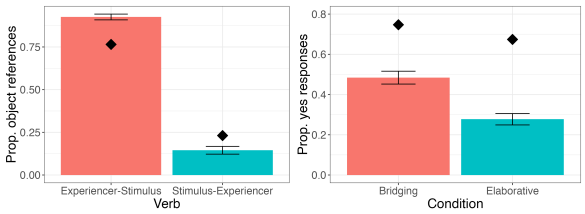


Figure 7: Results of Experiments 9–10. (a) Participants showed strong implicit causality biases, preferring object continuations for stimulus-experiencer verbs and subject continuations for experiencer-stimulus verbs (Experiment 9). (b) Bridging inferences were more likely endorsed than elaborative inferences, consistent with original discourse inference patterns (Experiment 10). Bars reflect mean response proportions across participants; error bars indicate ±1 *SE*. Black diamond markers denote the original experimental outcomes. The present one-trial-per-run setup removes potential carryover or adaptation effects and enhances the design's suitability for benchmarking LLMs.

Figure 18 showed the mean values.

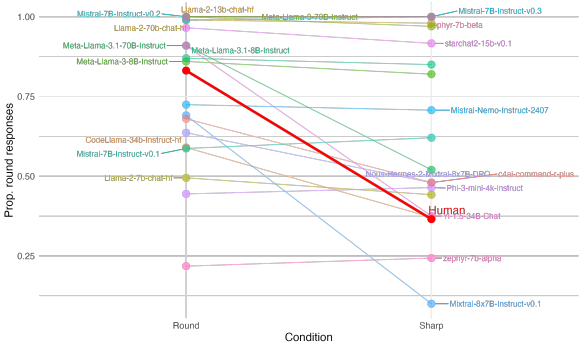


Figure 8: Experiment results of models and humans in Experiment 1

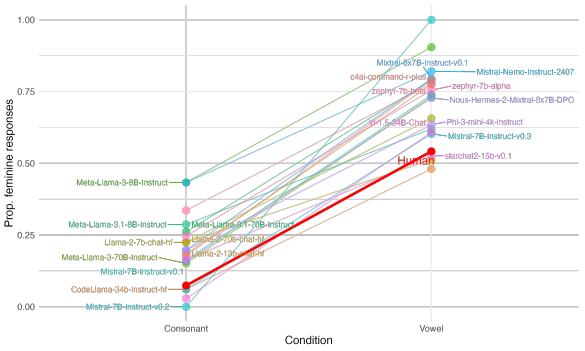


Figure 9: Experiment results of models and humans in Experiment 2

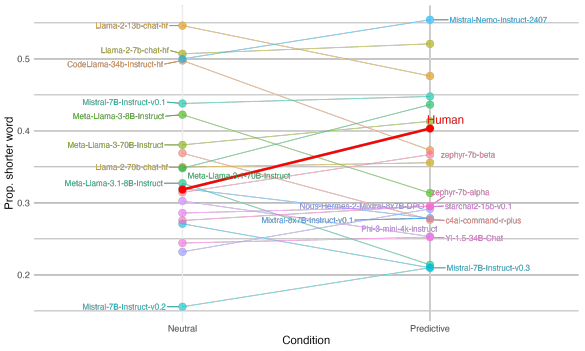


Figure 10: Experiment results of models and humans in Experiment 3

D Humanlikeness Scores for All 20 LLMs

As shown in Table 2, the language human-likeness scores for the 20 language models across 10 psycholinguistic experiments are summarized.

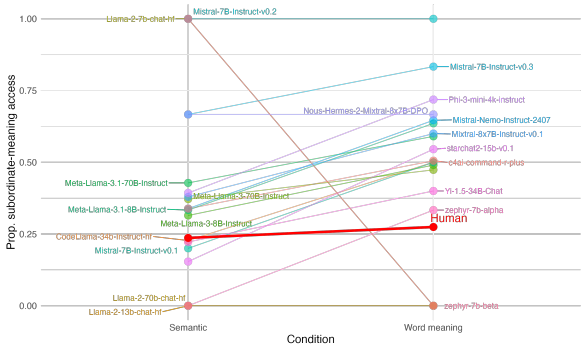


Figure 11: Experiment results of models and humans in Experiment 4

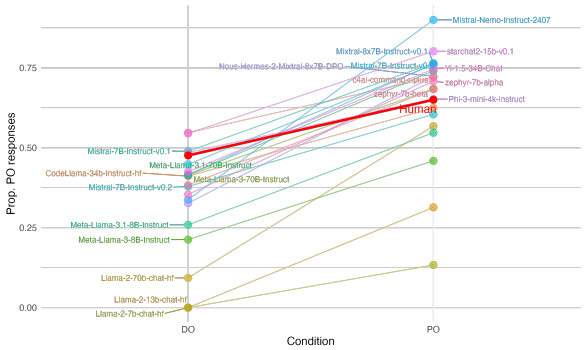


Figure 12: Experiment results of models and humans in Experiment 5

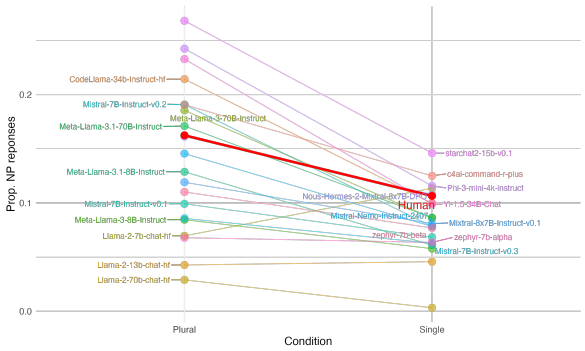


Figure 13: Experiment results of models and humans in Experiment 6

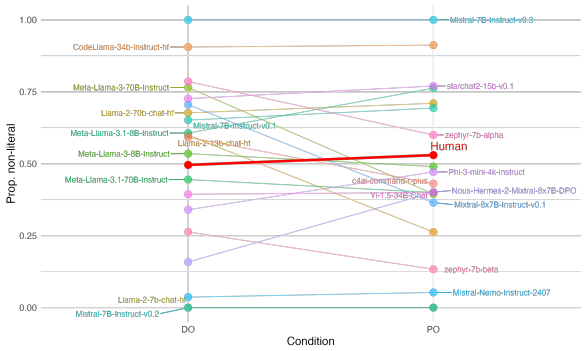


Figure 14: Experiment results of models and humans in Experiment 7

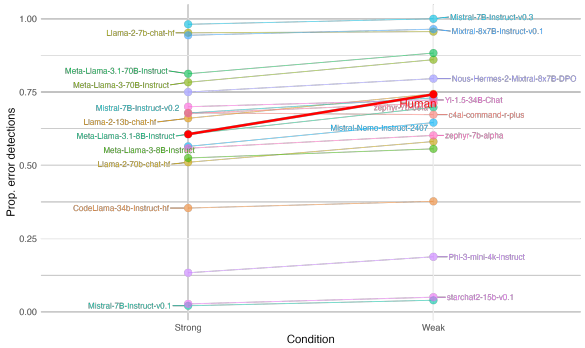


Figure 15: Experiment results of models and humans in Experiment 8

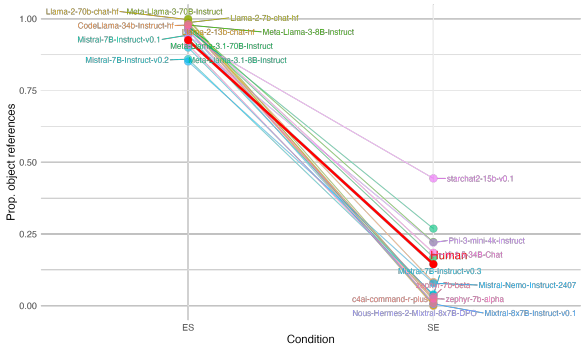


Figure 16: Experiment results of models and humans in Experiment 9

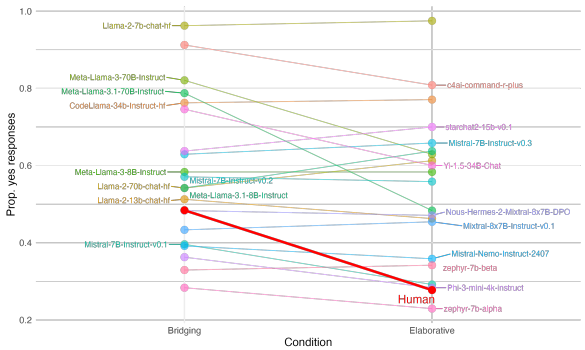
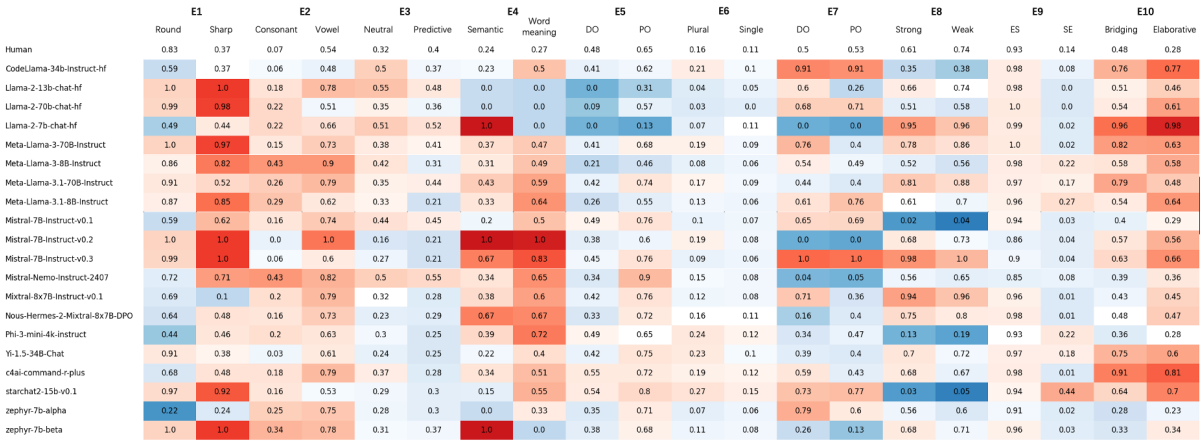


Figure 17: Experiment results of models and humans in Experiment 10



	E1		E2		E3		E4		E5		E6		E7		E8		E9		E10	
	Round	Sharp	Consonant	Vowel	Neutral	Predictive	Semantic	Word meaning	DO	PO	Plural	Single	DO	PO	Strong	Weak	ES	SE	Bridging	Elaborative
Human	0.83	0.37	0.07	0.54	0.32	0.4	0.24	0.27	0.48	0.65	0.16	0.11	0.5	0.53	0.61	0.74	0.93	0.14	0.48	0.28
CodeLlama-34b-Instruct-hf	0.59	0.37	0.06	0.48	0.5	0.37	0.23	0.5	0.41	0.62	0.21	0.1	0.91	0.91	0.35	0.38	0.98	0.08	0.76	0.77
Llama-2-13b-chat-hf	1.0	1.0	0.18	0.78	0.55	0.48	0.0	0.0	0.0	0.31	0.04	0.05	0.6	0.26	0.66	0.74	0.98	0.0	0.51	0.46
Llama-2-70b-chat-hf	0.99	0.98	0.22	0.51	0.35	0.36	0.0	0.0	0.09	0.57	0.03	0.0	0.68	0.71	0.51	0.58	1.0	0.0	0.54	0.61
Llama-2-7b-chat-hf	0.49	0.44	0.22	0.66	0.51	0.52	1.0	0.0	0.0	0.13	0.07	0.11	0.0	0.0	0.95	0.96	0.99	0.02	0.96	0.98
Meta-Llama-3-70B-Instruct	1.0	0.97	0.15	0.73	0.38	0.41	0.37	0.47	0.41	0.68	0.19	0.09	0.76	0.4	0.78	0.86	1.0	0.02	0.82	0.63
Meta-Llama-3-8B-Instruct	0.86	0.82	0.43	0.9	0.42	0.31	0.31	0.49	0.21	0.46	0.08	0.06	0.54	0.49	0.52	0.56	0.98	0.22	0.58	0.58
Meta-Llama-3.1-70B-Instruct	0.91	0.52	0.26	0.79	0.35	0.44	0.43	0.59	0.42	0.74	0.17	0.09	0.44	0.4	0.81	0.88	0.97	0.17	0.79	0.48
Meta-Llama-3.1-8B-Instruct	0.87	0.85	0.29	0.62	0.33	0.21	0.33	0.64	0.26	0.55	0.13	0.06	0.61	0.76	0.61	0.7	0.96	0.27	0.54	0.64
Mistral-7B-Instruct-v0.1	0.59	0.62	0.16	0.74	0.44	0.45	0.2	0.5	0.49	0.76	0.1	0.07	0.65	0.69	0.02	0.04	0.94	0.03	0.4	0.29
Mistral-7B-Instruct-v0.2	1.0	1.0	0.0	1.0	0.16	0.21	1.0	1.0	0.38	0.6	0.19	0.08	0.0	0.0	0.68	0.73	0.86	0.04	0.57	0.56
Mistral-7B-Instruct-v0.3	0.99	1.0	0.06	0.6	0.27	0.21	0.67	0.83	0.45	0.76	0.09	0.06	1.0	1.0	0.98	1.0	0.9	0.04	0.63	0.66
Mistral-Nemo-Instruct-2407	0.72	0.71	0.43	0.82	0.5	0.55	0.34	0.65	0.34	0.9	0.15	0.08	0.04	0.05	0.56	0.65	0.85	0.08	0.39	0.36
Mixtral-8x7B-Instruct-v0.1	0.69	0.1	0.2	0.79	0.32	0.28	0.38	0.6	0.42	0.76	0.12	0.08	0.71	0.36	0.94	0.96	0.96	0.01	0.43	0.45
Nous-Hermes-2-Mixtral-8x7B-DPO	0.64	0.48	0.16	0.73	0.23	0.29	0.67	0.67	0.33	0.72	0.16	0.11	0.16	0.4	0.75	0.8	0.98	0.01	0.48	0.47
Phi-3-mini-4k-instruct	0.44	0.46	0.2	0.63	0.3	0.25	0.39	0.72	0.49	0.65	0.24	0.12	0.34	0.47	0.13	0.19	0.93	0.22	0.36	0.28
Yi-1.5-34B-Chat	0.91	0.38	0.03	0.61	0.24	0.25	0.22	0.4	0.42	0.75	0.23	0.1	0.39	0.4	0.7	0.72	0.97	0.18	0.75	0.6
c4ai-command-r-plus	0.68	0.48	0.18	0.79	0.37	0.28	0.34	0.51	0.55	0.72	0.19	0.12	0.59	0.43	0.68	0.67	0.98	0.01	0.91	0.81
starchat2-15b-v0.1	0.97	0.92	0.16	0.53	0.29	0.3	0.15	0.55	0.54	0.8	0.27	0.15	0.73	0.77	0.03	0.05	0.94	0.44	0.64	0.7
zephyr-7b-alpha	0.22	0.24	0.25	0.75	0.28	0.3	0.0	0.33	0.35	0.71	0.07	0.06	0.79	0.6	0.56	0.6	0.91	0.02	0.28	0.23
zephyr-7b-beta	1.0	1.0	0.34	0.78	0.31	0.37	1.0	0.0	0.38	0.68	0.11	0.08	0.26	0.13	0.68	0.71	0.96	0.03	0.33	0.34

Figure 18: Average targeted response rates for 20 language models and human participants across conditions in 10 psycholinguistic tasks. Color gradients indicate relative performance: red cells reflect higher response rates than those of humans, while blue cells indicate lower rates.

Experiment	Overall	Sound		Word		Meaning		Syntax		Discourse	
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
Meta-Llama-3.1-70B-Instruct	**66.50**	**89**	62	61	6	81	77	**80**	67	**80**	63
Meta-Llama-3.1-8B-Instruct	65.89	73	65	60	12	84	**78**	79	**74**	79	56
Phi-3-mini-4k-instruct	64.61	61	68	59	19	**89**	71	76	48	**80**	76
Mistral-Nemo-Instruct-2407	63.69	74	63	56	10	84	77	52	75	79	68
Llama-2-13b-chat-hf	63.15	57	57	51	25	75	67	74	72	76	79
Mistral-7B-Instruct-v0.1	62.77	73	**70**	62	24	87	43	69	36	79	**84**
CodeLlama-34b-Instruct-hf	62.18	79	64	60	23	82	53	58	63	79	61
c4ai-command-r-plus	60.77	79	60	63	8	72	72	66	59	78	50
Meta-Llama-3-8B-Instruct	60.65	69	53	57	14	79	78	59	66	77	54
starchat2-15b-v0.1	60.57	58	**70**	58	25	87	73	62	36	75	62
gpt-4o	58.58	60	63	**68**	2	71	77	47	61	75	62
gpt-3.5-turbo	58.32	55	61	66	3	76	76	71	47	76	50
Yi-1.5-34B-Chat	58.20	67	54	55	13	72	70	61	65	78	48
Llama-2-7b-chat-hf	57.47	74	61	58	22	67	69	50	60	75	39
zephyr-7b-alpha	56.96	57	62	47	23	85	29	44	73	76	75
Meta-Llama-3-70B-Instruct	56.73	60	61	55	4	71	75	57	59	75	50
gpt-4o-mini	56.21	56	58	62	3	70	75	46	58	75	57
Mistral-8x7B-Instruct-v0.1	52.80	60	53	48	23	71	46	43	59	73	52
Mistral-7B-Instruct-v0.3	52.45	53	58	47	25	75	38	49	59	73	47
Mistral-7B-Instruct-v0.2	50.18	13	58	54	14	72	61	46	64	71	49
zephyr-7b-beta	47.85	28	53	48	**26**	71	7	38	73	75	60

Table 2: Language human-likeness scores for 20 language models across 10 psycholinguistic experiments. **Bold** values indicate the highest task-wise scores; underlined values denote the lowest.

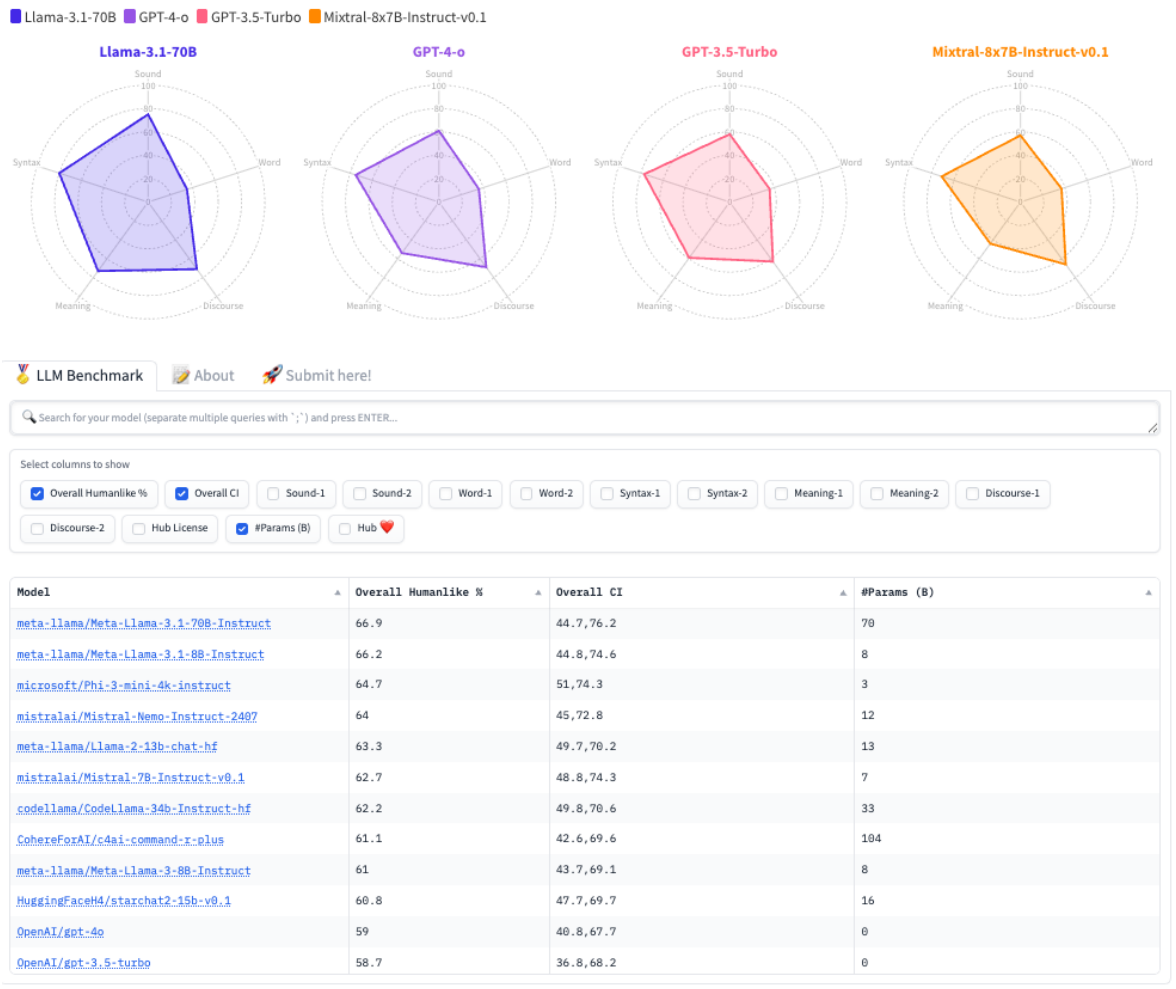


Figure 19: Leaderboard of language human-likeness scores across LLMs as assessed by the HLB