Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned

Anonymous ACL submission

Abstract

In this work we compare the generative behavior at the next token prediction level in several language models by comparing them to human productions in the cloze task. We find that while large models trained for longer are typically better estimators of human productions, but they reliably under-estimate the probabilities of human responses, over-rank rare responses, under-rank top responses, and produce highly distinct semantic spaces. Altogether, this work demonstrates in a tractable, interpretable domain that LM generations can not be used as replacements of or models of the cloze task.

1 Introduction

Many language models are trained to a cloze-like objective (Taylor, 1953), in which they perform next-word prediction (NWP) or masked language modeling (MLM), producing a distribution of probabilities of these tokens given the surrounding context. Effective language models (LMs) assign the greatest probability to words using the surrounding context. Humans similarly can perform this cloze task, such as when we try to guess what others are about to say. By aggregating across many human responses to produce a proportion, one can also obtain a probability distribution over words that fit to a given linguistic context or situation. Despite major advances in word prediction in context, however, it is generally understood that token prediction tasks cannot fully capture all of the factors that affect human responses, such as real-world knowledge, or syntactic relationships, even when fine tuned (Bender and Koller, 2020). Cloze productions, while instrumental in understanding language processing, are nevertheless poorly understood as a means of evaluating LM outputs.

Given the continued gap in understanding what makes generated "language" natural and human-like, this work aims to better characterize the linguistic-distributional and semantic properties that distinguish LMs from humans in cloze-like tasks. By better understanding the properties of human productions from a probabilistic perspective, we believe it is possible to bridge the gap between LMs and humans.

Here[1] we compare machine-generated predictions for sentence-final completions against a large-scale human cloze response dataset (Peelle et al., 2020). Such a cloze dataset allows us to specifically target *single-token generations*, which allows us to sidestep the high degree of uncertainty in multiword utterance production, and the many algorithms used to solve these problems (Giulianelli et al., 2023). That is, not only is multiword utterance production a computationally intractable problem, but different models use different parameters and decoding strategies for the production of longer texts – effectively dynamical systems – which makes them difficult to compare and analyze. Furthermore, selecting high-quality multiword utterances depends on accurately estimating the probabilities of words given the prior context, leading to the explosion of error (Valmari, 1998). Systems that cannot perform single word production cannot be trusted to generate longer, more complex language more accurately. We therefore present analyses of single-word and single-token response data to provide an estimate of the best case scenario for language modeling.

2 Human cloze and language modeling

The present work examines single-word production in cloze tasks to compare human and LM productions on an identical task. Note that the term "cloze" is used in psycholinguistics and natural language processing with close but distinct meanings. It always consists of a task setting where an agent (either a human subject or a computer model) is given a context, typically a sentence where a word has been removed, and is asked to complete it with their best guess as to what the original was. A probability distribution over

[1] All the code used to produce this analysis is available at https://example.com/anonymized/for/review.

Pythia-70M	Pythia-160M	Pythia-2.8B	RoBERTa	GPT-2	Human
lot	swarm	swarm	swarm	bee	hive
swarm	bee	bee	bee	swarm	swarm
bee	new	h	hive	hive	bee
new	pest	was	sting	black	nest
threat	h	colony	spider	bad	wasp

Table 1: Top 5 completions for "He hated bees and feared encountering a ____".

the possible completions can optionally be deduced from this process. However, while these probabilities are meant to estimate some kind of likelihood of occurrence $\mathrm{P}(w \mid c)$ of a word w in context, they are obtained in significantly different ways.

For the *human* cloze task in psycholinguistics, every human subject in a pool will normally provide one completion w for every context c (though see Roland et al., 2012 for an example with multiple responses). $\mathrm{P}(w \mid c)$ is estimated as the relative frequency of w among all the completions observed for c. For the purposes of simplicity, cloze probabilities are computed under the strong but usually implicit assumption that the choice of a completion by a subject is a non-deterministic process that is identical and identically parameterized across all subjects, so that $\mathrm{P}(w \mid c)$ can be estimated simply as the prevalence of w among all responses to c (Smith and Levy, 2011).

For *language models*, and in particular for recent *neural* ones (NLM), the cloze task is cast as classification task, where the inputs are the contexts, the classes are the words in a vocabulary, and the classification is done by assigning a score to each possible word using a neural network and choosing the one with the maximal score. A probability distribution over words can be obtained by applying the $\mathrm{softmax}$ function to the output layer of this neural network, which is usually trained by using a variant of the stochastic gradient descent algorithm to minimize the negative log-likelihood of the removed word for every context in a large training corpus. While there is no guarantee of it[2], it is often[3] assumed that this probability distribution is learned to fit a latent likelihood of occurrence of a word in context.

Cloze data are significantly more sparse than language model predictions. Even the largest cloze datasets collect 40 to 100 human responses per context only and the number of *unique* responses is often much lower (Lowder et al., 2018; Luke and Christianson, 2018; Peelle et al., 2020). NLMs, on the other hand, typically generate predictions for over tens of thousands of tokens at once. It is precisely this estimation of the "long tail" of uncommon or improbable words that has been argued to account for the majority of their strength over human cloze data (Szewczyk and Federmeier, 2022; Shain et al., 2024; Oh and Schuler, 2023a).

Under these assumptions, the probability distributions obtained from a NLM (provided that a suitable one exists) should be highly correlated to those obtained from human cloze experiments and easier to obtain, leading to their use as a proxy in psycholinguistics works (Luke and Christianson, 2018; Staub et al., 2015; de Varda et al., 2024). Conversely, human cloze datasets can serve as benchmarks for NLMs: if it does indeed learn human-like probability distributions, a good NLM should provide distributions that are highly correlated to those obtained from human cloze data.

3 Data

We experiment on Peelle et al.'s (2020) completion norms dataset. It consists of 3085 English sentences for which human participants were asked to provide a final word, with each sentence receiving at least 100 manually validated responses in order produce reliable cloze probability estimates. The stimuli varied between eight and ten words in length, and varied in the degree of final-word predictability. The sentences are importantly long enough that LMs should be able to reliably guess appropriate words for the context.

We compare the cloze probabilities reported by Peelle et al. (2020) to probabilities extracted from several neural language models: GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019) in its "base" configuration, and the Pythia suite of language models (Biderman et al., 2023). All of these models use a subword vocabulary size of approximately 50 000 and thus have the same capacity for long tail

[2]Neither theoretically nor empirically, especially without explicitly optimizing for it — see e.g. Kong et al. (2020) or Ulmer et al. (2024)

[3]This is for instance the assumption made by Radford et al. (2019) and Devlin et al. (2019).

effects, though they differ in their training datasets and subword vocabularies.

In contrast to GPT-2 and RoBERTa, the Pythia models provide a way of exploring the influence of several hyperparameters on their behavior, namely model size, training time, and data deduplication. Pythia models are trained to perform NWP on the Pile (Gao et al., 2020) using a training procedure similar to GPT-2 and have been used extensively to evaluate LM fit to human reading data (Oh et al., 2024; Oh and Schuler, 2023b). These models are also interesting because their embedding (first) and projection (last) matrices are untied, in contrast to other models such as GPT-2, which could in principle produce different kinds of predictions. We note that due to their more recent release and because they were not designed for downstream task performance such as natural language inference, Pythia models could provide a poor baseline that will not generalize well to non-language modeling tasks (although the evaluations provided by Biderman et al. (2023) show that it is in fact not the case). In Experiments 1 and 2, we include only Pythia-160M because it contains a similar number of parameters as GPT-2 and RoBERTa and because the highest-performing, largest models are impractical for quick inspection of their behavior. However, we assess the role of model capacity in Experiment 3, which leverages a wider range of Pythia models to study the influence of model size and training budget.

We also assess GPT-2 (a decoder-only NWP model) and RoBERTa (an encoder-only MLM), which produce representations that have improved performance on many downstream NLP tasks and are perhaps the best-known examples of their respective flavor of Transformer-based LMs. Examining these earliest Transformer LMs ensures that the effects we observe are not tied to a possible low quality of the models in general: it has been empirically shown that they are good at what they have been designed for.

In all our experiments on model generations, we only consider predictions of a single word-initial subword token, which can either be a "word" or a fragment. While this can be somewhat limiting in theory because sometimes a human response might consist of several subword tokens[4], this decision makes it possible to directly estimate ranks and probabilities without having to resort to a sampling procedure (Holtzman et al., 2021; Giulianelli et al., 2023), which would otherwise make our results less interpretable. Directly measuring probabilities across the full sequence of generations is already a critical component of these sampling algorithms, but the relative computational complexity of generation of single tokens is much lower, enabling further experiments.

[4]In practice, we also found that most human responses were common words that are only a single token in the vocabularies of the models we study, so the distortion to the results we observe is generally negligible.

4 Experiment 1

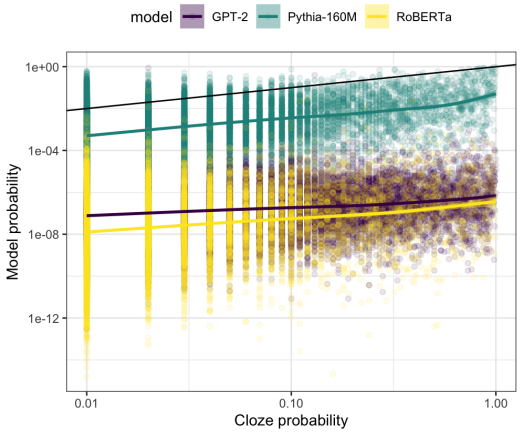


Figure 1: Correlation between human cloze probabilities and language model probabilities showing clear non-linearity in correlation and massive under-estimation of next-word probabilities.

This experiment compares GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), and Pythia (Biderman et al., 2023) to understand the probabilities of next-word generations by comparing human- and LM-derived estimates. Qualitatively, the generations of RoBERTa, GPT-2, and Pythia-160M vary widely in the content of their predictions. We present one example in Table 1.

In order to be used as models of human productions, the top LM predictions and/or their estimated probabilities should at the very least be consistent with observed human responses to the same cloze challenge. However, Figure 1 illustrates that LMs most often assign too little probability to human generations (Holtzman et al., 2021), with most models under-estimating cloze probabilities by several orders of magnitude.

These results also reveal that the item-level correspondence between LM and human cloze probabilities is poor, with considerable variance around the trendline and evidence of non-normality (Franke et al., 2024). Taken together, these findings suggest that LMs are generally "right for the wrong reasons" with respect to questions of predictive power (Giulianelli et al., 2024; Shain et al., 2024).

We note that the probabilities and correlations

extracted from GPT-2 and RoBERTa are much lower than previously reported in other works (Frisch and Giulianelli, 2024; Oh and Schuler, 2024). One possibility is that the reliance on web text is a poor estimator of human cloze probabilities (Smith and Levy, 2011). For instance, the subword vocabularies learned by RoBERTa and GPT-2 may be skewed in favor of decomposing responses into multiple subwords in human cloze data. By contrast, the Pile (Gao et al., 2020), which was used to train Pythia, contains much more diverse texts that are more representative of everyday language use, and may produce estimates of next word probabilities that are closer to psycholinguistic stimuli. Another possibility is that the differences in stimuli used by different studies (e.g., de Varda et al., 2024) produces different distributions of estimates. Finally, ongoing work suggests that subword aggregation plays a significant role (Nair and Resnik, 2023; Oh and Schuler, 2024) in the estimation of next-word probabilities.

5 Experiment 2

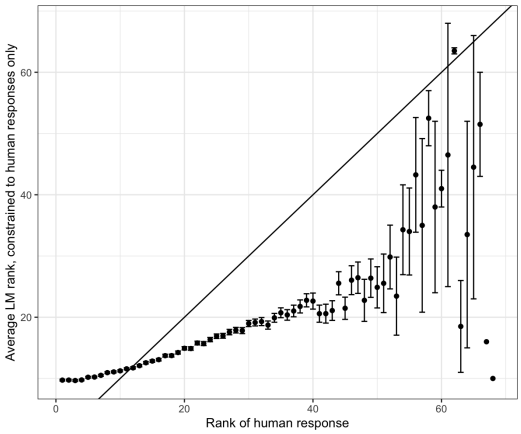


Figure 2: Rank correlation between Pythia-160M and human responses. Language models over-rank rare human responses (above solid line) and under-rank probable responses (below solid line).

Experiment 1 identified significant problems in the use of LM-based probabilities, but the selection of next words is largely one of ranking different candidates. One could reason that relying only on probabilities is unfair to language models, with generally reasonable rankings but poor estimation of the probability of next words due to data or training objective differents. For simplicity, we examined whether the Pythia-160M model (most similar in size to the state-of-the-art GPT-2; Biderman et al. 2023; Shain et al. 2024) was also miscalibrated in its rankings of human responses by comparing the rank of the probability associated with the first subword of a human response versus the empirical rank in the cloze dataset. This allows us to constrain the correlation to only valid human responses and determine if the language models are able to accurately order them, as in a multiple-choice test.

Figure 2 shows that one need not be optimistic about ranks here either. The rank correspondence between language model probability and human cloze probability suggests that Pythia-160M consistently under-estimates the ranks of probable responses (left), and over-estimates the ranks of rarer responses (right).

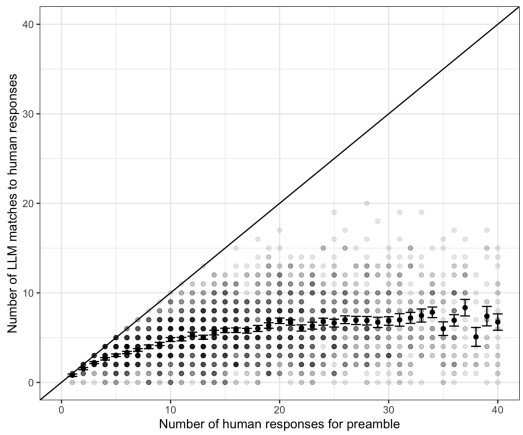


Figure 3: Pythia-160M next word prediction recovery compared to human productions.

An additional perspective on this ranking problem requires inspection of precision and recovery of cloze responses. We present in Figure 3 the recovery of human responses that appear in the top 40 predictions of Pythia-160M. Generally speaking, when the LM is constrained to produce approximately the same number of unique responses as humans could provide in the cloze task, it is unable to recover more than 10 of the human responses in the top 40 responses. Indeed, per this Figure, it is clear that the proportion of responses would fall and a larger threshold would be necessary to recover rare human responses.

However, it is possible that top human responses are simply poorly ordered within the models' top responses. We examine whether a more lenient criterion would improve fit to human cloze production by allowing ranks to include any higher ranks. Figure 4 shows the performance of this model on recovering words at the correct ranking or higher. It is clear that only very probable responses (i.e. sentences that produce only one or two unique completions) are recovered by this model, and that human and LM

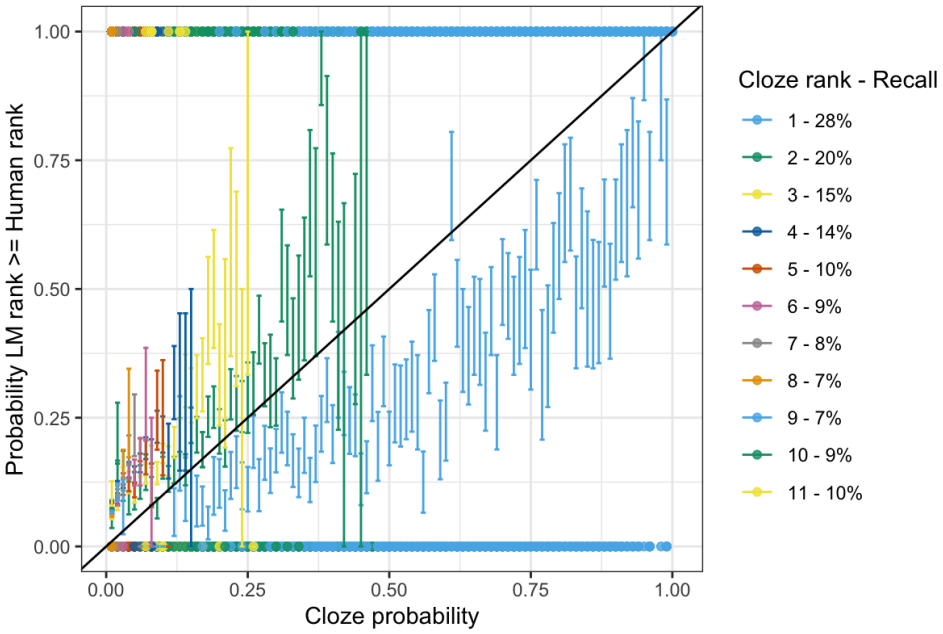


Figure 4: Pythia-160M recovery by cloze response rank and proportion of recoveries. Note that no responses with rank 11 are recovered at an equal or higher rank.

predictions diverge rapidly in more uncertain contexts. Even then, top responses are retrieved in first place by the models only 28 % of the time, while rarer responses in the top 10 human responses are selected at that rank or higher only 9 % of the time. Strikingly, this is an over-estimate, such that second-best and lower answers answers in the human data appear to be commonly ranked as the top human response. Indeed, the ranks of top human responses seem to be cannibalized by the higher-ranking behavior of lower-probability human responses.

6 Experiment 3

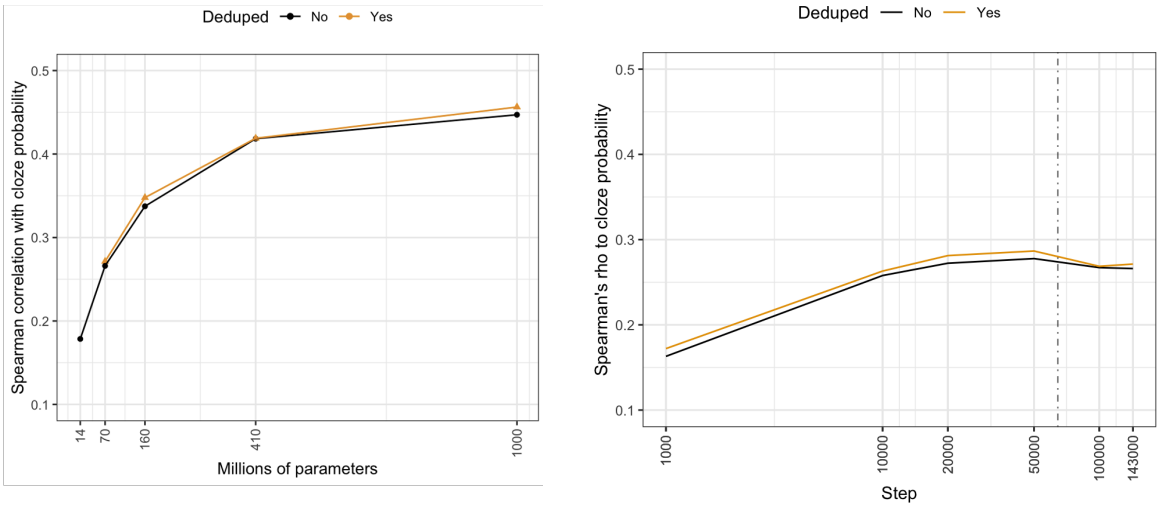
This experiment assesses the contribution of learning in LMs to fit to human data. To assess the influence of model size and training progress on the correlation between NLM and cloze rank distributions, we experiment with the Pythia suite (Biderman et al., 2023), a series of Transformer-based NLM spanning a wide range of model sizes for a single architecture, trained on the same data and with publicly available intermediate checkpoints. For comparisons based on model size, we use all the models up to and including Pythia-2.8B[5], in standard and "deduped" versions both. For comparisons based on training progress, we use checkpoints of Pythia-160M[6] at five training steps from 1000 to 143 000.

We measure the similarity between human and LM rank distributions with Spearman's (1904) rank correlation coefficient ρ. For LMs, we only consider the relative ranks they give to human responses and ignore all preceding and intervening responses. In other words, these correlations only answer the question "Did the LM rank human responses in the same order humans did?". As in Experiment 1, this allows us to obviate any discrepancy due to the long tail of LM vs. human responses, but it may present an optimistic estimate of the fit of the LM to human production data.

LM performance generally improves logarithmically with model size, such that greater model sizes have higher correlations with human responses, but shows clear diminishing returns in terms of computational budget. Additionally, deduplicated text typically produces higher-quality estimates of cloze productions, which suggests that highly repetitive data sources such as web text will consistently under-estimate human language productions (Oh and Schuler, 2023a; Smith and Levy, 2011). Surprisingly, the performance of the five checkpoints we examined reveals diminishing returns over the course of multiple epochs as well, suggesting that more training data does not necessarily lead to better estimation of

[5]The lower bound for the "phase change" in model behavior observed during training by Biderman et al. (2023).

[6]According to Figure 5a, the 2.8B version would have had a slightly better correlation overall, but to reduce the cost of our experiments, and because given the results reported by Biderman et al. (2023) we expect the trends relatively to training budget to be similar enough, we only experiment with the 160M checkpoints for Figure 5b.



(a) Effect of model size on rank correlations with human responses.

(b) Effect of training budget on rank correlations with human responses.

Figure 5: Spearman rank correlation between LM (Pythia 160m-deduped) and human responses as functions of model size and training budget. Both show positive correlation plateauing after a certain point, staying at $\rho < 0.5$ out of a maximum of 1.

human linguistic creativity.

Figure 5a shows a trend of steady increasing of correlation between ranks as model size increases. However, even in this lenient setting, the largest models stay at $\rho < 0.5$, an at best moderate correlation, and the trend suggests rapidly diminishing returns over 410 million parameters. As for training steps, Figure 5b shows that the correlation between human and LM ranks — while overall mediocre — also increase during training, but plateaus around 50 000 steps.

Biderman et al. (2023) report that model performance is positively correlated with both model size and training time. Our results show the same trend for correlation with human responses. The drop in performance at longer training budgets replicates some prior work suggesting that the number of tokens can influence predictive power (Oh and Schuler, 2023a). However, the generally low correlations at all sizes and all training times suggest that even if these trends persisted, it would be impractical to rely solely on them to obtain good models of human productions.

7 Experiment 4

The results of Experiments 1 and 2 showed modest correlations between human- and machine-generated responses in a cloze task for approximating the probabilities and ranks of human responses, replicating previous work showing deficiencies in next-word prediction (Holtzman et al., 2020; Vaidya et al., 2023; Botch and Finn, 2024; Klein et al., 2024). Moreover, the improvements in fit to cloze probabilities from greater

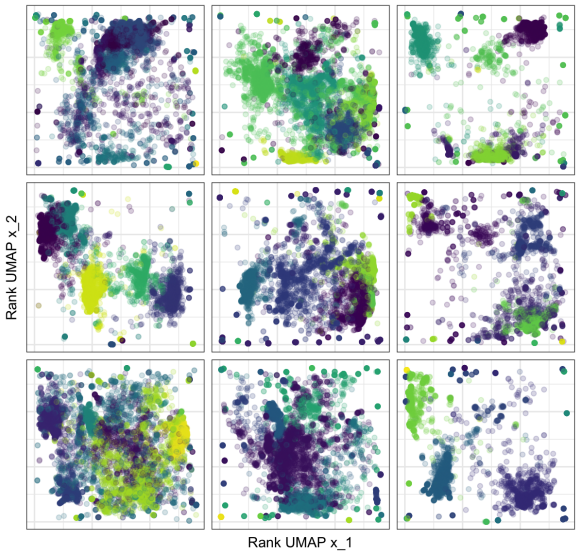


Figure 6: UMAP demonstrates high degrees of cluster separability in human-generated productions.

model size and longer model training seem modest, suggesting that better language model memorization comes with poor creativity gains. That is, language models show clear effects of over-fitting to the training data (Shain et al., 2024; Oh et al., 2024). While this and previous results underscore how clearly language models and humans prefer to produce responses in a different order from each other, it is possible that human and language model responses are more closely linked on a semantic level due to the distributional training regime (Baroni et al., 2014). Furthermore, individual responses may effectively express the

same meaning while being written slightly differently (Holtzman et al., 2021). Thus, this experiment examines the distributional semantic properties of human and language model responses and assesses the degree to which they inhabit the same geometric space.

To better understand the semantic structure of generations by humans and language models, we perform clustering experiments on the middle layer (layer 7) from the Pythia-160m-deduped model. We selected this layer due to the apparent greater compositional capacity of the middle layers of language models (Tenney et al., 2019; Jawahar et al., 2019; Antonello and Huth, 2024), which are less tuned toward next-word prediction and better reflect a word's role in a sentence. To cluster the responses, we use the scikit-learn (Pedregosa et al., 2011) implementation of the Bayesian Gaussian mixture model (BGMM), which uses a Bayesian Dirichlet prior (Blei et al., 2001) to approximate the distribution of n-dimensional embeddings of human and/or model responses as samples from a mixture of n-dimensional multivariate normal distributions.

We are inspired by recent work that has applied BGMMs to learn coherent cluster representations with relatively stable parameters for cloze data (Jacobs et al., 2024), which we apply here. A visual inspection of the human responses, which we plot in Figure 6 using UMAP over pairwise Euclidean distances between the embeddings (McInnes et al., 2018), shows a high degree of local clustering. Data points, which are colored according to the Bayesian Gaussian mixture model's cluster assignments after training, show considerable separation between the different clusters, demonstrating the utility of this clustering algorithm for interpretation.

If human and LM generations have different semantic network topologies, then we expect the degree of human-versus-LM content in a dataset to influence the clusterings that are obtained, both qualitatively and quantitatively. For this experiment we construct four BGMM models trained on embeddings belonging to different categories of generation. We ultimately wish to assign cluster labels to three response-based categories, which are human-only responses (HUMAN; 39 995 tokens), LM-only responses (LM; 106 253 tokens), and shared responses (BOTH; 17 147 tokens). To minimize the effect of dataset size on clustering performance, we constructed clustering models that sampled the input to the mixture model to have the same number of data points as the human-only dataset.

The results of the clustering analysis provide evidence that responses that are shared by both the Pythia model and human participants are unique from other types of responses. In terms of clustering model likelihoods assigned to out-of-sample data, the human-only and LM-only clustering models have poor clustering performance on responses produced by both generative processes (Table 2, but a model that is designed to cluster those responses shows a high degree of coherence relative to all other models, and is in fact the only clustering model with a positive mean likelihood. Additionally, it is clear that human responses are also poorly explained by language model representations, with substantially lower likelihoods than LM-based likelihoods. One way to interpret these data is that LM-based clusters are of sufficiently high quality to assign LM-generated next token predictions, but the resulting representations for human productions are of lower quality due to being out-of-distribution.

	HUMAN	LM	BOTH
HUMAN	**-1.79**	-3.95	-6.60
LM	-1.95	**-0.93**	-2.18
BOTH	-8.55	-7.21	**0.10**

Table 2: Mean data likelihood scores per observation for different data sources (rows) and mixture models (columns).

We next compare the similarity of the cluster assignments across different clustering datasets. Specifically, we compute a measure known as Clustering Mutual Information (CMI), a special case of the general Clustering Agreement Index (CAI; Rabbany and Zaïane, 2017) that estimates the degree of overlap between multiple clustering models (i.e., across different LM implementations) for the same data points (productions). We review the method in depth in the Appendix. Table 3 shows the clustering agreement across each of the Gaussian mixture models that we trained. The human-derived clustering model provided the highest degree of agreement with responses produced by both the human and Pythia-160M-deduped model, though the LM-only clustering model proved similarly effective. On the other hand, the LM-only model and the model trained on shared human and LM responses had relatively poor performance on human data. Note that all models have low degrees of agreement in general relative to the ideal scenario.

8 Related work

Our approach is inspired in part by Smith and Levy (2011), who explored the differences between n-gram language models and human productions in a next-word cloze task. Their study found that relative to 5-gram probabilities drawn from web

	HUMAN	LM	BOTH
HUMAN	-	*.478*	*.479*
LM	.505	-	**.519**
BOTH	**.600**	.596	-

Table 3: Clustering agreement indices for different data sources (rows) and mixture models (columns).

text, human responses in the cloze task were more concrete, familiar, diverse, and semantically related to the preceding context. Their work suggests that cloze productions deviate substantially from web text in ways that should be evident in our experiments.

Relative to human predictions for sentence-final completion data, NWP models such as GPT-{2, 3, J} appear to produce better estimates of the predictability of a word, with evidence mostly coming from correlational studies of large-scale reading time datasets (Shain et al., 2024). However, recent studies have repeatedly shown that linguistic predictability quantified by language models lags behind other sources of linguistic knowledge. Moreover, while the predictive power of modern LMs is a significant improvement beyond n-gram language models (Goodkind and Bicknell, 2018), modern language models are commonly miscalibrated to most tasks, such as natural language inference.

The reasons for the lack of predictive power of human cloze data are still largely unknown, but are generally attributed to the expense of probing enough humans in enough linguistic contexts to accurately measure next-word probability. However, it seems clear that at present, LLMs cannot function as a drop-in replacement for cloze estimates. These issues present a challenge for computational psycholinguistics about the claim that LM predictions are fundamentally better for capturing human predictions (Shain et al., 2024). off-the-shelf language models must still be calibrated toward human language prediction tasks.

The model capacity experiments we present build on some previous work (**?**), which reported a peak in fit to reading time data at approximately step 1000 in training the same class of Pythia models, which is earlier than we observe in our analyses. The effect that is seen in this literature may arise due to differences in sensitivity to word frequency depending on the psycholinguistic task (Oh et al., 2024; Klein et al., 2024).

Adjusting the temperature of the $\mathrm{softmax}$ function when computing probabilities from LLMs' outputs has been found to improve predictive power for reading times (Liu et al., 2023). We note that temper does not affect the ranks of responses, unless temperature settings are so high as to collapse all low-probability responses to the same value, and so temperature scaling would not affect massive rank disparities relative to human cloze probabilities.

Previous clustering approaches other than Gaussian mixture models have been applied to improve intepretability of LM outputs (Thompson and Mimno, 2020; Grootendorst, 2022) and have demonstrated that some LMs cluster contextual embeddings more effectively than others. However, this work has not examined human or LM productions together.

9 Discussion

Given the differences for participant responses against human ranks in both LM ranks and LM probabilities, it is clear that LMs' ability to directly model human production behavior is quite limited, even at the single-token level. These observations are consistent across models with different training data (Pythia vs GPT-2) and training objective (RoBERTa's MLM vs. Pythia and GPT-2 NWP). The fact that neither ranks nor top-k prediction sets match properly suggests further that these divergences are not due to probability distortions due to long tail/truncation effects.

In addition to poor probabilistic and ranking behavior, the semantic spaces generated by the model that we examined are not semantically aligned with human productions. Human beings' responses are poorly clustered by mixture models trained on non-human generations, and the types of responses that humans and LMs produce similarly poorly capture the semantic diversity of human responses that LMs do not predict.

The similarity between human cloze results and LMs are dissimilar enough that LMs cannot be used as proxies for the human production process. We believe that these results present a cautionary tale, in which poor fit of cloze probabilities to reading time data relative to language model probabilities arises not from sparsity but the inability to retrieve human-like predictions, as evidenced by weak correlations and substantial under-estimation of human responses. This suggests that the use of LM probability estimates may not rely on the same types of predictive processes at all, counter to much recent work (Shain et al., 2024; Blank), and that researchers should be very careful in general when using LMs as proxies for any predictive modeling.

10 Limitations

In this work, we evaluate only a single dataset to understand the correspondence between cloze probabilities and language model estimates, which may limit the ability to draw conclusions about the nature of the relationship for all cloze tasks and datasets. Some of the cloze datasets that we did not examine include word-by-word cloze completions that allow researchers to assess the probabilities of words at each stage in sentence processing. Additionally, different subword tokenization strategies can have an impact on estimates, and our approach to select only word-initial subwords may inflate estimates of next-word probabilities (Oh and Schuler, 2024). Finally, it is common in the literature to report Δ log likelihood instead of correlation coefficients; we do not report this here because we do not do model comparison, as we are concerned entirely with unbiased generations. However, future work should compare and contrast these approaches directly.

11 Ethical Considerations

The Peelle et al. (2020) data were gathered using crowdsourcing on Mechanical Turk. The original researchers obtained IRB approval for their research study. Some of the human responses, and many of the LM predictions, contain profanity, sexually explicit content, or other offensive content. Furthermore, LMs in general are likely to produce sexist, ableist, and racist responses even when guardrails are implemented. Researchers seeking to evaluate LM outputs using human judgments should be aware of the potential for harmful material in these outputs.

References

Richard Antonello and Alexander Huth. 2024. Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Idan A Blank. What are large language models supposed to model? *Trends in Cognitive Sciences*.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. *Advances in neural information processing systems*, 14.

Thomas L Botch and Emily S Finn. 2024. Humans diverge from language models when predicting spoken language. In *ICLR 2024 Workshop on Representational Alignment*.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

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.

Michael Franke, Polina Tsvilodub, and Fausto Carcassi. 2024. Bayesian statistical modeling with predictors from llms. *arXiv preprint arXiv:2406.09012*.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. Generalized measures of anticipation and responsivity in online language processing. *arXiv preprint arXiv:2409.10728*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *International Conference on Learning Representations*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

Cassandra Jacobs, Ryan Hubbard, Kara D Federmeier, and Loïc Grobol. 2024. Uncovering patterns of semantic predictability in sentence processing.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. The effect of surprisal on reading times in information seeking and repeated reading. *arXiv preprint arXiv:2410.08162*.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2023. Improving fit to human reading times via temperature-scaled surprisal. *arXiv preprint arXiv:2311.09325*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Matthew W. Lowder, Wonil Choi, Fernanda Ferreira, and John M. Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, 42(S4):1166–1183.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11251–11260, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023a. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary poses a confound for calculating word probabilities. *arXiv preprint arXiv:2406.10851*.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonathan E Peelle, Ryland L Miller, Chad S Rogers, Brent Spehar, Mitchell S Sommers, and Kristin J Van Engen. 2020. Completion norms for 3085 english sentence contexts. *Behavior Research Methods*, 52:1795–1799.

Reihaneh Rabbany and Osmar Zaïane. 2017. A general clustering agreement index: For comparing disjoint and overlapping clusters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Self-hosted preprint.

Douglas Roland, Hongoak Yun, Jean-Pierre Koenig, and Gail Mauner. 2012. Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122(3):267–279.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

Adrian Staub, Margaret Grant, Lori Astheimer, and Andrew Cohen. 2015. The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82:1–17.

Jakub M Szewczyk and Kara D Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of memory and language*, 123:104311.

Wilson L Taylor. 1953. "Cloze Procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626*.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. 2024. Calibrating large language models using their generations only. *Preprint*, arXiv:2403.05973.

Aditya Vaidya, Javier Turek, and Alexander Huth. 2023. Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 58–69.

Antti Valmari. 1998. *The state explosion problem*, pages 429–528. Springer Berlin Heidelberg, Berlin, Heidelberg.