On the scaling relationship between cloze probabilities and language model next-token prediction

Cassandra L. Jacobs
Department of Linguistics
University at Buffalo
Buffalo, NY, USA
cxjacobs@buffalo.edu

Morgan Grobol
MoDyCo
Université Paris Nanterre
Paris, France
lgrobol@parisnanterre.fr

Abstract

Recent work has shown that larger language models have better predictive power for eye movement and reading time data. However, we know less about how model capacity relates to human production statistics in the cloze task, which are used to predict reading times as well. While even the best models under-allocate probability mass to human responses, larger models assign higher-quality estimates of next tokens and their likelihood of production in cloze data because they are less sensitive to lexical co-occurrence statistics while being better aligned semantically to human cloze responses. The results provide support for the claim that the greater memorization capacity of larger models helps them guess more semantically appropriate words, but makes them less sensitive to low-level information that is relevant for word recognition.

1 Introduction

Humans regularly engage in linguistic prediction, and can often guess what others are about to say. The predictability of a word in context is often quantified by computing its *cloze probability*: the proportion of human participants who provide a particular word as a response when completing a sentence in the cloze task (e.g., "He hated bees and feared encountering a ___"; Taylor, 1953). Cloze probabilities partly reflect a word's appropriateness in a given linguistic context or situation, and are determined in part by semantic fit to the sentence and knowledge of word and multiword statistics (Smith and Levy, 2011). In human experiments, words with higher cloze probabilities are typically read more quickly and/or trigger more attenuated neural activity (de Varda et al., 2024).

Understanding how language models align with cloze data is especially important because cloze captures dimensions of sentence predictability that language models may lack access to, such as real-world knowledge (de Varda et al., 2024; Szewczyk and Federmeier, 2022), but the cost of data collection for probability estimation with human subjects is high (Pivel-Villanueva et al., 2026). Additionally, recent works have largely turned to probabilities computed using large language models as a replacement for cloze probabilities (Shain et al., 2024; Nair and Oh, 2026).

It is understood that large language models do not perfectly align with human production preferences (Eisape et al., 2020; Jacobs and McCarthy, 2020; Nair and Oh, 2026; Pivel-Villanueva et al., 2026). Recent work suggests that model capacity and memorization ability influences psychometric predictive power (De Varda and Marelli, 2024; Futrell et al., 2020; Oh and Linzen, 2025; Shain et al., 2024; Szewczyk and Federmeier, 2022), but this is largely unexplored with cloze data. Some prior work analyzing the alignment between 5-gram models and human cloze productions showed that humans are biased toward guessing words that are higher-frequency and more semantically similar to other words in the preceding context (Smith and Levy, 2011). More recently, Michaelov et al. (2025) showed that larger models are more sensitive to semantics and less sensitive to n-gram statistics than smaller ones, suggesting less reliance on straightforward memorization. It seems reasonable to expect, then, that larger models will demonstrate improved capacity to model human language production biases (Pickering and Garrod, 2013; Smith and Levy, 2011).

We first characterize the scaling relationship between models and the lexical alignment between model generations and cloze productions (Experiment 1). The results show that larger models provide a *better* match to cloze data at the lexical level. Experiment 2 evaluates the contribution of memorization to fit to human cloze data. We compare n-gram statistics to both NLM predictions and cloze probabilities. Further analyses show that model size, training budget, and data deduplica-

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 544–554
July 3-4, 2026 ©2026 Association for Computational Linguistics

Table 1: Top 5 Pythia next-subword continuations in different Pythia models for the sentence fragment, "He hated bees and feared encountering a ____". $*$ "was" possibly corresponds to the first subword of "wasp" and "h" to "hive".

Pythia-70M	-160M	-2.8B	Human
lot	swarm	swarm	hive
swarm	bee	bee	swarm
bee	new	h*	bee
new	pest	was*	nest
threat	h*	colony	wasp

tion all influence fit to cloze probabilities. Finally, Experiment 3 probes the correspondence in semantic similarity structure between human and NLM predictions. The results provide corroborating evidence that larger models better align with the semantic structure of human cloze responses. We finally review how the results for cloze alignment contrast with similar investigations of reading time data (Nair and Oh, 2026).

2 Cloze data

We conduct our experiments on the Peelle et al. (2020) completion norms dataset. It consists of 3085 English sentence stems created by experimental psycholinguists for which human participants were asked to type the next word, with each stem receiving at least 100 manually validated responses to produce reliable cloze probability estimates. The stimuli were constructed to vary in the degree of final-word predictability and a search for these materials suggests the sentences are not included in the Pile (Xu et al., 2025), in contrast to naturalistic materials indexed on the web (Luke and Christianson, 2018; Oh et al., 2025). Of the 3085 sentences, only 16 sentences were shorter than 7 or longer than 10 words. Despite the short contexts, NLMs typically guessed appropriate words, and statistical tests confirmed that sentence length did not have a significant influence on Spearman rank correlation to human cloze probabilities ($p = .27$).

For analyses of language model probabilities, we choose to use the first subword of human cloze responses. While a human response might consist of several subword tokens, considering only word-initial subwords provides enough precision for our purpose (Nair and Oh, 2026). This also greatly simplifies probability estimation (Holtzman et al., 2021; Giulianelli et al., 2023; Oh and Schuler, 2024; Nair and Resnik, 2023; Pimentel and Meister, 2024; Pivel-Villanueva et al., 2026). For example, estimating a probability distribution on a complete full-word vocabulary would require computing probabilities for every sequence of subwords in that vocabulary — which is computationally prohibitive and not linguistically motivated. This decision is also justified by the distribution of human responses: approximately 50.4 % of human response tokens are very common words that correspond to a single token in the models' vocabularies. On average, each unique response type was broken into (1.64 ± 0.67) subwords. For analyses of semantic spaces, we mean-pool subword embeddings of multitoken continuations (Giulianelli et al., 2024; Nair and Oh, 2026).

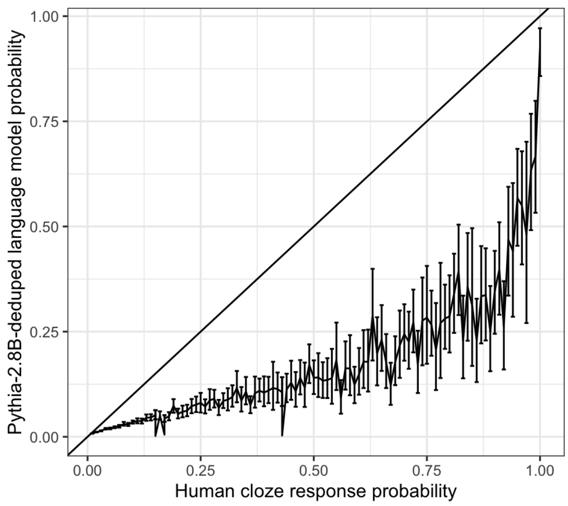
3 Experiment 1: Comparing production probabilities

Our experiments compare cloze probabilities reported in the Peelle et al. (2020) dataset against probabilities from NLMs trained to perform next-token prediction (Table 1). Inspired by similar work assessing the role of model capacity on fit to reading time data by Oh and Schuler (2023a), we leverage the Pythia suite of language models (Biderman et al., 2023) to understand whether and when language models capture human production variability in the cloze task. Pythia models are next token predictors trained on the same dataset (the Pile; Gao et al. 2021) and provide a way to test the influence of model hyperparameters on fit to cloze probabilities.

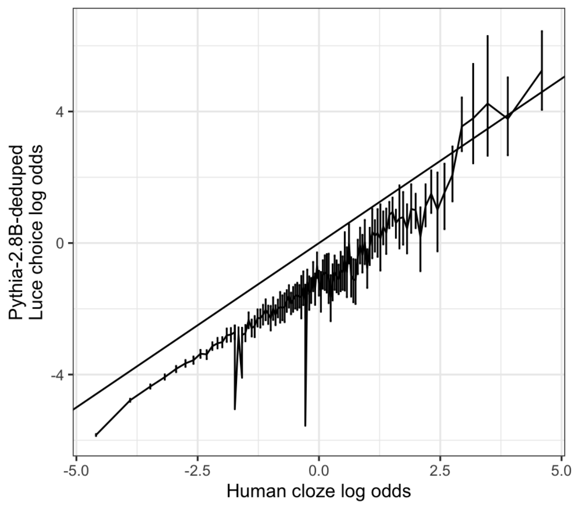
This analysis characterizes the best-case scenario for probabilistic alignment among the Pythia models by quantifying the correlation between language model-allocated probability mass and human cloze probabilities. `pythia-2.8B-deduped` obtained the overall highest correlation to the human responses. We show that the performance of this model in matching human production preferences can be impressive with some deficiencies in the allocation of probability mass across several transformations of language model probabilities.

3.1 Cloze probabilities

Pythia probabilities and Peelle cloze probabilities are correlated with each other (Pearson's $\rho = 0.492$; Spearman's $\rho = 0.485$), but this correlation is complex (Figure 1a). Firstly, many cloze completions are assigned probabilities approximately



(a) Relationship between language model probabilities and cloze probabilities. Error bars represent bootstrapped confidence intervals. Solid line represents 1:1 relationship.



(b) Relationship between log-odds of human cloze probabilities and NLM probabilities renormalized over human responses.

Figure 1: Human-NLM cloze correlations for Pythia-2.8B-deduped showing major deficiencies in assigning probability mass to human completions. All models at all sizes showed similar patterns and are not pictured here.

half of their empirical probability in the corpus. For instance, glancing at Figure 1a, it is clear that words at 0.75 cloze probability – representing on average 3 of every 4 responses – were assigned around language model probabilities of 0.25. Nevertheless, on average, words at the endpoints of the cloze probability scale ($p = .01$ and $p = 1$, respectively) are reasonably well captured; the middle range of probabilities is, however, systematically under-estimated by language model probabilities. This seems to be a case of miscalibration, a common issue in neural network classifiers (Guo et al., 2017).

3.2 Cloze log-odds

Comparing raw probabilities, as in Section 3.1, is made inconvenient by the skewness of the probability distributions observed. To alleviate this, we complement them by analyses of log-odds, i.e., after applying a logit transform to all probabilities p with a small smoothing parameter α:

$$\text{logit}(p) = \log \frac{p+\alpha}{1-(p+\alpha)}, \alpha = 1 \times 10^{-6}$$

Unlike raw probabilities, log-odds are unbounded, which helps to make low-probability tails less skewed. The results of a linear mixed effects regression with random intercepts by sentence stem and cloze response reveal a correlation between language model log-odds and cloze probabilities that was significantly positive ($\beta = 0.244$, $t(42410) = 133.9$, $p < .001$). The large negative estimate of the intercept term reveals that language model probability mass is assigned more uniformly ($\beta = -2.275$, $t(23930) = -167.6$, $p < .001$).

For a cleaner comparison to human responses, which sum to 1, we also renormalize the NLM probabilities for each preamble using Luce's choice rule (Luce, 1977), instead of taking the raw value after softmax over the NLM's whole vocabulary. The equivalent analysis with Luce choice log-odds analysis revealed that while it is possible to produce a stronger correlation to cloze probabilities ($\beta = 0.252$, $t(50440) = 139.7$, $p < .001$), Pythia still mostly under-allocates probability mass as seen in the intercept term ($\beta = -2.572$, $t(18290) = -226.7$, $p < .001$; left side of Figure 1b). At the very highest levels of cloze probability (on the right side of the figure), Luce-choice language model log-odds may slightly over-allocate probability mass. These results provide converging evidence that accounting for differences in the response space does not eliminate the tendency of Pythia-2.8B-deduped to under-allocate probabilities.

3.3 Rank correlations

Both human cloze and NLM next-token probabilities are typically skewed, with a long tail of rare responses. In this analysis, we follow Sinha et al. (2023) and compare the NLM *rank* of a human response's first subword probability against the em-

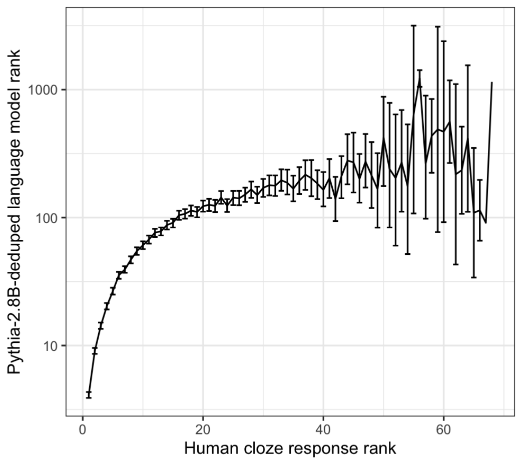


Figure 2: Relationship between language model ranks and cloze ranks. Error bars represent bootstrapped confidence intervals.

pirical rank of the response in the cloze dataset. This allows us to constrain the correlation to the same numerical range and determine if the language models are able to accurately order human responses, as in a multiple-choice test (Holtzman et al., 2020). This transformation partly addresses the skew in human probabilities and ensures an even distribution of values even when the vocabulary sizes are quite different.

We measure the alignment between human and LM rank distributions with Spearman's (1904) rank correlation coefficient ρ. Consistent with the previous analyses, the human cloze and language model probabilities are strongly correlated ($\rho = 0.48$, $p < .001$). However, one need not be optimistic about ranks: while there is a clear correspondence between NLM and human cloze probabilities, the model consistently under-ranks probable human responses (Figure 2).

4 Experiment 2: Model capacity and human cloze alignment

The previous experiment established that models systematically under-allocate the probability mass of human responses in the cloze task. This experiment builds on findings that language model next-token predictions are strongly correlated with unigram probabilities, n-gram probabilities, and (contextual) semantic similarity (Michaelov et al., 2025). Thus, we further examine how the correlations between NLM outputs from Pythia models of all sizes (14 M, 70 M, 160 M, 410 M, 1 B, 1.4 B, and 2.8 B), trained on the Pile and the deduplicated Pile compare to cloze data and simple n-gram statistics.

4.1 Pythia correlation to n-gram baseline

This experiment probes potential sources of the correlation between language model probabilities and human cloze responses in Section 3, while laying the groundwork for our comparison of NLMs of varying sizes and training durations in Section 4.2. Here we examine the correlation between the ranks of n-gram-based next-token prediction scores to the ranks of language model and human cloze probabilities in the Peelle et al. (2020) norms.

Following Michaelov et al. (2025), we estimated scores for a 5-gram model smoothed with Brants et al.'s (2007) Stupid Backoff heuristic ($n = 5$, $\alpha = 0.4$) on Wikipedia[1] and found strong correlations with NLM next-token probabilities. Interestingly, model capacity and data deduplication modulated the correlation between language model probabilities and 5-gram scores. We find that the Pythia-70M model produced the strongest correlation to 5-gram estimates ($\rho = 0.57$), and this relationship decreased in magnitude as models get larger; Pythia-2.8B model showed a much weaker correlation to 5-gram statistics ($\rho = 0.44$).

Data deduplication in the Pile negatively impacts the correlation between Wikipedia 5-gram scores and language model probabilities. Additionally, consistent with prior findings showing that NLMs rely less on n-gram statistics as they become larger, we find that the effect of deduplication on correlation to 5-gram scores becomes smaller as models get larger (Figure 3). These results suggest that smaller NLMs are more prone to memorization of n-gram statistics, and data deduplication reduces the bias toward these repeated sequences.

NLM probabilities were only weakly correlated with unigram scores, which were themselves strongly correlated with 5-gram scores. Human cloze responses showed a weak correlation with 5-gram scores and very weak correlation to unigram scores. We summarize these correlations in Table 2. Of particular note is the size of the correlation between NLM probabilities and 5-gram scores, which is similar in magnitude to the correlation between the best NLM and human probabilities.

4.2 Pythia correlation to cloze completions

This experiment assesses the contribution of training in NLMs to fit to human data. The analyses in

[1] Precisely, the 2023-11-01 dump of the English version of Wikipedia, tokenized with the GPT-2 tokenizer.

Table 2: Spearman correlation coefficient estimates between different language models (5-gram vs. Pythia NLM) and human cloze completions. All estimates significant.

	Human	NLM	5-gram
NLM	0.375		
5-gram	0.177	0.483	
Unigram	0.007	0.187	0.445

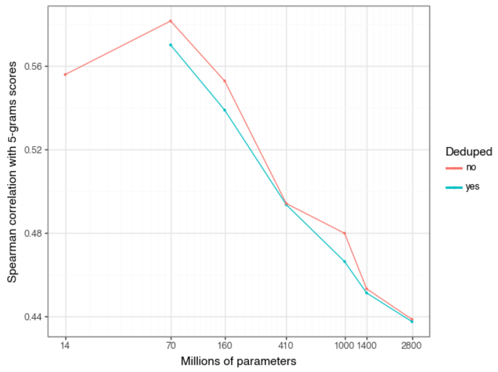
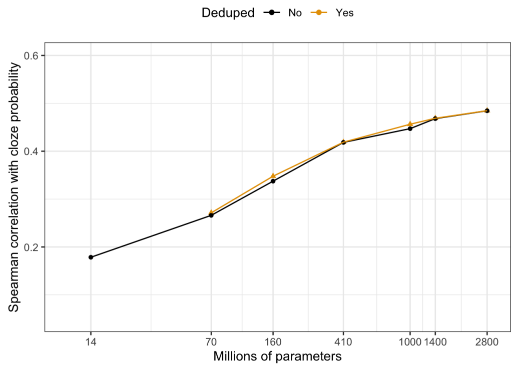
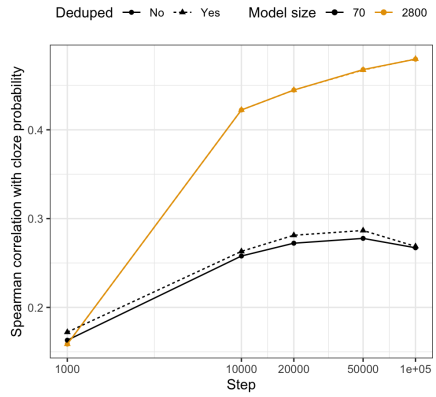


Figure 3: Correlation strength between Wikipedia 5-gram scores and Pythia next-token probabilities by model size and data deduplication. X axis in log scale. Note: 14M model does not have a deduplicated implementation.



(a) Effect of model size in millions of parameters on rank correlations with human responses. X axis in log scale. Note: 14M model does not have a deduplicated implementation.



(b) Effect of training budget (number of steps) on rank correlations with human responses. Deduplication plays a larger role in fit with smaller models. X axis in log scale.

Figure 4: Effect of model capacity and size on cloze correlation.

the previous section assessed the fit between NLMs and cloze data overall, but did not closely examine the effect of model size, training duration, or data deduplication. Better understanding the contribution of these factors is important because larger models or models trained for longer appear to produce less predictive next-word probabilities than smaller NLMs (Oh and Schuler, 2023a,b; Shain et al., 2024) but this relationship is not necessarily expected for cloze data. We again standardize the correlation between human cloze and NLM probabilities and only consider the relative ranks of each. For comparisons based on model size, we use all the models up to and including Pythia-2.8B, in standard and deduplicated versions. For comparisons based on training progress, we use checkpoints of Pythia-70M and Pythia-2.8B models with and without deduplication.

Here performance generally improved with model size (Figure 4a), such that greater model sizes had higher correlations with human responses. As evidenced by the non-linear slope plotted in Figure 4b, we observe diminishing returns over the course of multiple epochs, suggesting that longer pre-training does not necessarily lead to better estimation of human predictions. The declining fit evident at the longest training times replicates some prior work suggesting a phase change in model behavior indicating overtraining (Biderman et al., 2023). Models trained on deduplicated text typically produced higher-quality estimates of cloze rankings, even for models of the largest size ($\Delta\text{LL} = 19$; $\chi^2(0) = 37.154$; $p < 0.001$).

Interim summary

Experiment 1 showed modest-to-high correlations between human- and machine-generated responses in a cloze task for approximating rankings of human responses. Those analyses replicated previous findings demonstrating deficiencies in next-word prediction even among the largest pretrained neural language models (Holtzman et al., 2020; Vaidya et al., 2023; Botch and Finn, 2024; Gruteke Klein et al., 2024; Oh et al., 2024). Experiment 2 demonstrated a novel result that data deduplication has an important effect on the degree of correspondence between human cloze and NLM predictions, such that deduplication appears to improve the probability estimates of some combinations. This finding coupled with the reverse relationship between Wikipedia 5-gram scores and deduplicated NLM probabilities provides additional support for the claim that memorization of n-grams is pervasive in smaller language models (Michaelov et al., 2025).

5 Experiment 3: Human-LM semantic representational similarity analysis

Michaelov et al. (2025) argue that larger models can encode "more complex relationships" between words in text, such as anticipating (distributionally) semantically similar words. We therefore expect that greater model capacity will promote higher semantic alignment to human cloze responses, which should be reflected in the structure of their semantic spaces. We wish to specifically quantify comparisons such as those presented in Table 3. We explore this by comparing the distributional semantic spaces from human and language model-generated sources using a procedure known as Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008). For all model capacities, we first assess semantic spaces derived from symbolic co-occurrence information in Section 5.2 and then move to language model embedding-based similarity spaces in Section 5.3.

5.1 Analysis procedure

NLM and human participant completions only partially overlap. So, we implement a slight modification to RSA that allows us to compare the macro structure of language model next-word predictions and human responses independently of their overlaps for any given preamble by quantifying the changes in the structures of semantic neighborhoods.

In order to compare semantic spaces, we represent each of them as a matrix of similarities over W, the set of words they have in common. In practice, the semantic spaces we consider are instantiated as sets of vector representations, and we quantify the similarity $s^{\alpha}_{i,j}$ in a given space α between two words $w_i \in W$ and $w_j \in W$ as the cosine similarity between their vector representations v^{α}_i and v^{α}_j:

$$s^{\alpha}_{i,j} = \frac{\langle v^{\alpha}_i \mid v^{\alpha}_j \rangle}{\|v^{\alpha}_i\| \|v^{\alpha}_j\|}$$

The alignment between two semantic spaces α and β is then computed as a function of their respective matrix representations S_α and S_β.

5.2 Positive pointwise mutual information to assess shared response biases

The first analysis focuses on symbolic, co-occurrence based definitions of similarity. Consider for example cases where a model guesses both *dog* and *cat* for a sentence stem, and humans guess these same words for a different sentence. According to the distributional hypothesis, so long as each data source broadly produces the same co-occurrence statistics, the model knows that it should assign high probability to words with similar meanings in the same contexts. To probe this, we use the principal component factorization (PCA, Wold et al., 1987) of a positive pointwise mutual information (PPMI) matrix derived from co-occurrence counts, which encodes distributional semantic similarities between words (Levy and Goldberg, 2014).

To this end, we consider the Peelle et al. (2020) stimulus and count two words w_i and w_j as co-occurring whenever they both appear in the top-k responses[2] for a particular preamble given a particular data source. We then compute a PPMI matrix by normalizing counts into log probabilities, subtracting the marginal log probabilities and keeping only positive values:

$$\mathrm{PPMI}_{i,j} = [\log(p(w_i, w_j)) - \log(p(w_i)p(w_j))]^+$$

We factorize this matrix using PCA — after a row-wise normalization to produce centered and scaled values — and keep only the d dominant components, yielding d-dimensional vector representations of words.

We compute (cosine) similarity matrices from these vector representations (see 5.1) then estimate their similarities as the Spearman correlation between their upper triangles $(s^{\alpha}_{i,j})_{i<j}$ and $(s^{\beta}_{i,j})_{i<j}$. We expected that this value would be high *a priori* due to the degree of overlap between the hu-

[2]We set $k = 40$ because very few Peelle sentences elicited more than 40 unique completions and visual inspection of the human-NLM rank correlation (Figure 2) suggests that the correspondence to Peelle completions degrades rapidly. Consequently, much higher k would runs the risk of corrupting the semantic spaces with irrelevant words.

Table 3: Top five most similar words to the word *bee* in human and NLM cloze completion semantic spaces.

Count-based PPMI + PCA		Averaging Contextual Embeddings		
Pythia-160M	Human	Pythia-160M	Pythia-2.8B	Human
mosquito	fly	moth	moth	moth
sting	ant	insect	mosquito	mosquito
snake	mosquito	bug	insect	spider
spider	insect	spider	bug	bug
bug	butterfly	mouse	spider	bird

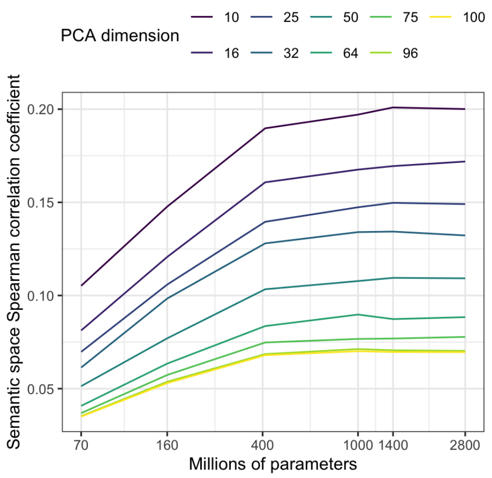


Figure 5: Correlation between model capacity, PCA dimensionality, and correlation between language model and human semantic spaces.

man and language model responses. All models were strongly and significantly correlated with human cloze semantic spaces, with some variation in fit based on model capacity, which generally improved as model capacity improved. In particular, we observe that higher d for PCA produced more differentiated similarity spaces, such that the size of the correlation coefficient between human and LLM completions monotonically decreased with increasing d across models of all sizes (Figure 5).

5.3 Semantic neighborhood structure of contextual embedding representations

One shortcoming of the previous analysis is that human responses can align with language model completions even in the absence of lexical overlap (Holtzman et al., 2021). This experiment compares the neighborhoods constructed from contextual embeddings of each NLM-generated and human-produced cloze response. To put human and NLM-generated data sources of all sizes on equal representational footing, we use Pythia-2.8B to produce embeddings, which we extract from the last layer. This ensures that differences in embedding quality do not detrimentally affect the comparison of semantic spaces, particularly for smaller models. If NLMs are largely producing words that are reasonable substitutes for the same words that humans would have provided, then embeddings of those predictions should be highly similar to humans' completions. Additionally, if larger NLMs better capture semantic structure, then we expect semantic correspondence to improve with capacity.

Here we model semantic spaces for human and LM completions using non-contextual word representations (v_i^α) obtained by averaging the Pythia 2.8B contextual embeddings of all their occurrences:

$$v_i^\alpha = \frac{1}{|D_i^\alpha|} \sum_{c \in D_i^\alpha} h_{2.8\text{B}}(w_i, c)$$

where D_i^α is the set of the preambles where word w_i appears as a completion for dataset α (either human completions of completions for a given LM) and $h_{2.8\text{B}}(w_i, c)$ is the Pythia 2.8B contextual embedding of w_i given preamble c.

We then define the alignment $A_k(\alpha, \beta)$ between two spaces by their *neighborhood overlap*: the degree to which each word ("pivot") has the same neighbors in both spaces, quantified as the average Jaccard similarity between the top k most similar words for a pivot in each space:

$$A_k(\alpha, \beta) = \frac{1}{|W|} \sum_{w \in W} \frac{|V_k^\alpha(w) \cap V_k^\alpha(w)|}{|V_k^\alpha(w) \cup V_k^\alpha(w)|}$$

where $V_k^\alpha(w)$ is the set of the k closest neighbors of word w in space α.

For $k = 20$, a generalized linear model predicting overlaps as a function of model size revealed

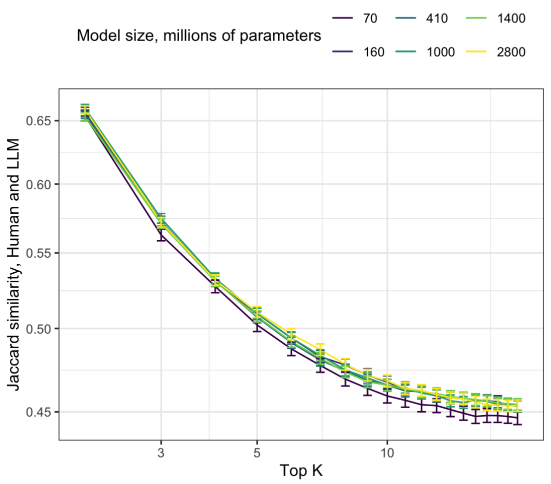


Figure 6: Neighborhood-based alignment between human and NLM completions as a function of model size and neighborhood size considered. X axis in log scale.

modest improvements in overlap with greater capacity ($\beta = 0.006041$, $t(24233) = 2.540$, $p < .05$). However, the overall alignment was very similar across models of all sizes; Pythia-70M averaged 48.7 % and Pythia-2.8B averaged 49.8 % neighborhood overlap. Further analysis applying stricter cutoffs for k (i.e., 2 to 19; Figure 6) revealed that while it is possible to produce greater semantic alignment to human data (up to 66 % in the case of the top neighbor), greater model capacity did not produce significantly higher degrees of overlap.

6 Discussion

Probabilities extracted from NLMs correlate strongly with scores from n-gram models, reading times, and cloze probababilities. We present strong evidence that larger models typically better explain cloze productions than smaller ones, both at the lexical and semantic levels of linguistic structure. The consequences of this finding for computational psycholinguistics are multifaceted. First, one possible explanation for the gap in accounting for the variety of cloze responses could be that people are not choosing words solely proportionate to their activation (Kumar and Hawkins, 2025; Meister et al., 2024). For example, some models of language production assume a type of semantic competition (Abdel Rahman and Melinger, 2009; Harley, 1993). New models could constrain or alter the next-word prediction process to explicitly account for semantic structure and better capture cloze task dynamics. Since larger models that are trained to predict language production output, highly accurate models should be better able to capture the complexity of factors that influence word choice (Antonello and Huth, 2024; Oh et al., 2024; Michaelov et al., 2025).

Second, our findings shed light on the purported weak correspondence between human cloze productions and reading times (Nair and Oh, 2026; Shain et al., 2024; Staub, 2025; Szewczyk and Federmeier, 2022). Since larger models correlate better with cloze probabilities both lexically and semantically, and both are less predictive of reading times, the cloze task may engage semantics and longer-range context to a greater extent than reading tasks (Staub et al., 2015), which may be more sensitive to narrow statistical context or perceptual features (Futrell et al., 2020; Milligan et al., 2023; Nair and Oh, 2026; Schotter et al., 2023). Language comprehension must be robust to unexpected language, and making very specific predictions may not always be optimal (Arora et al., 2022), so relying on lexical statistics more than semantic structure may be optimal for reading (Oh et al., 2024). Our findings complement those of Nair and Oh (2026), who focus specifically on the relationship between cloze probabilities, language model probabilities, and fit to reading times.

The results of this study underscore the need to formally specify the cognitive and computational processes that are used to generate next-token predictions in both language models and humans. More empirical and modeling work is needed to determine whether and how different facets of language processing make use of generative procedures (Staub, 2025).

7 Limitations

The present studies did not assess a very large number of language models or model architectures, which limits the generalizability of the scaling relationships we present here, which may not hold for other types. Additionally, our results may not generalize to systems that rely on reinforcement learning or other methods that could be used to explicitly simulate the cloze task (Martínez et al., 2025). However, the work we present here suggests that work remains to be done in human-like text generation (Kendro et al., 2026).

Our results are likely limited cross-linguistically, as few languages have high-quality pre-trained models and large-scale cloze completion norms that can be used to study scaling relationships to cloze tasks. Languages with more complex morphology or different word orders will likely require different strategies for comparison from English.

Our analyses ultimately focus on words that are shared between data sources, and thus any biases of different models toward subword-scale predictions rather than English words may influence the semantic neighborhood structure of next-token predictions in ways that hinder human-model alignment. Future work should assess how these two factors interact.

8 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 NLM predictions, contain profanity, sexually explicit content, or other offensive content. Furthermore, NLMs in general are likely to produce sexist, ableist, and racist responses even when guardrails are implemented. Researchers seeking to evaluate NLM outputs using human judgments should be aware of the potential for harmful material in these outputs.

References

Rasha Abdel Rahman and Alissa Melinger. 2009. Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes*, 24(5):713–734.

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.

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. Estimating the entropy of linguistic distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Dublin, Ireland. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

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

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Andrea De Varda and Marco Marelli. 2024. Locally biased transformers better align with human reading times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–36, Bangkok, Thailand. Association for Computational Linguistics.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619, Online. Association for Computational Linguistics.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2021. The pile: An 800GB dataset of diverse text for language modeling. *ArXiv preprint*, abs/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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669, Miami, Florida, USA. 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. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney,* *NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Trevor A Harley. 1993. Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, 8(3):291–309.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cassandra L. Jacobs and Arya D. McCarthy. 2020. The human unlikeness of neural language models in next-word prediction. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, page 115, Seattle, USA. Association for Computational Linguistics.

Kelly Kendro, Jeffrey Maloney, and Scott Jarvis. 2026. Do large language models produce texts with “human-like” lexical diversity? evidence from four chatgpt models. *International Journal of Applied Linguistics*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Abhilasha A Kumar and Robert D Hawkins. 2025. Lexical search and social reasoning jointly explain communication in associative reference games. *Journal of Experimental Psychology: General*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185.

R Duncan Luce. 1977. The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3):215–233.

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

Gonzalo Martínez, Javier Conde, Pedro Reviriego, and Marc Brysbaert. 2025. Simulating lexical decision times with large language models to supplement megastudies and crowdsourcing. *Behavior Research Methods*, 57(10):294.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498.

James A Michaelov, Roger P Levy, and Ben Bergen. 2025. Language model behavioral phases are consistent across architecture, training data, and scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sara Milligan, Brian Nestor, Martín Antúnez, and Elizabeth R Schotter. 2023. Out of sight, out of mind: Foveal processing is necessary for semantic integration of words into sentence context. *Journal of Experimental Psychology: Human Perception and Performance*, 49(5):687.

Sathvik Nair and Byung-Doh Oh. 2026. Clozing the gap: Exploring why language model surprisal outperforms cloze surprisal.

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 Tal Linzen. 2025. To model human linguistic prediction, make llms less superhuman. *ArXiv preprint*, abs/2510.05141.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

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 pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

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.

Byung-Doh Oh, Hongao Zhu, and William Schuler. 2025. The inverse scaling effect of pre-trained language model surprisal is not due to data leakage. *ArXiv preprint*, abs/2506.01172.

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.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Estrella Pivel-Villanueva, Elisabeth Frederike Sterner, and Franziska Knolle. 2026. Can llms capture stable human-generated sentence entropy measures?

Elizabeth R Schotter, Sara Milligan, and Victoria M Estevez. 2023. Event-related potentials show that parafoveal vision is insufficient for semantic integration. *Psychophysiology*, 60(7):e14246.

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.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

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. 2025. Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, 11(1):17–34.

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.

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, Singapore. Association for Computational Linguistics.

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.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. Infini-gram mini: Exact n-gram search at the Internet scale with FM-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24944–24969, Suzhou, China. Association for Computational Linguistics.