What Goes Into a LM Acceptability Judgment? Rethinking the Impact of Frequency and Length

Lindia Tjuatja[1], Graham Neubig[1], Tal Linzen[2], Sophie Hao[2]

[1]Carnegie Mellon University, [2]New York University

Correspondence: lindiat@andrew.cmu.edu

Abstract

When comparing the linguistic capabilities of language models (LMs) with humans using LM probabilities, factors such as the length of the sequence and the unigram frequency of lexical items have a significant effect on LM probabilities in ways that humans are largely robust to. Prior works in comparing LM and human acceptability judgments treat these effects uniformly across models, making a strong assumption that models require the same degree of adjustment to control for length and unigram frequency effects. We propose *MORCELA*, a new linking theory between LM scores and acceptability judgments where the optimal level of adjustment for these effects is estimated from data via learned parameters for length and unigram frequency. We first show that MORCELA outperforms a commonly used linking theory for acceptability—SLOR (Pauls and Klein, 2012; Lau et al., 2017)—across two families of transformer LMs (Pythia and OPT). Furthermore, we demonstrate that the assumed degrees of adjustment in SLOR for length and unigram frequency overcorrect for these confounds, and that larger models require a lower relative degree of adjustment for unigram frequency, though a significant amount of adjustment is still necessary for all models. Finally, our subsequent analysis shows that larger LMs' lower susceptibility to frequency effects can be explained by an ability to better predict rarer words in context.[1]

1 Introduction

Are the probabilities provided by language models (LMs) compatible with theories of linguistics and human language processing? This is a fundamental question that has implications in fields from psycholinguistics to natural language processing applications, and requires understanding of how to relate LM probabilities with quantities associated with human language processing. In this work, we consider the relationship between LM probabilities and human judgments of *linguistic acceptability*, and investigate how LM probabilities should be treated when comparing them to human acceptability judgments.

(1) **Acceptable** (Score: 1.19)
It is silly for one to sing in the shower.

(2) **Borderline** (Score: 0.00)
Tanya danced with as handsome a boy as her father.

(3) **Unacceptable** (Score: −1.11)
It seems a cat to be in the tree.

Figure 1: English sentences with linguistic acceptability scores reported by Sprouse et al. (2013). Participants where asked to rate sentences on a scale from 1 (least acceptable) to 7 (most acceptable), whose scores were then normalized by participant to a mean of 0 and variance of 1. Scores shown are averaged across participants.

Acceptability judgments are speakers' reported perceptions about the well-formedness of utterances, which are often elicited by asking questions such as "How natural/acceptable/grammatical is this utterance?" (Sprouse, 2013). These judgments are typically reported in binary or numerical form (Sorace and Keller, 2005; Sprouse, 2007, 2015; Lau et al., 2017), and they are collected through a variety of annotation tasks, such as binary classification, Likert scale scoring, or ranking (Schütze, 2016; Sprouse et al., 2013). Examples of acceptability judgments, from Sprouse et al. (2013), are provided in Figure 1. Judgments such as these play a central role in linguistics, where they are used to motivate and evaluate theories of natural language syntax (Chomsky, 1957).

In order to relate LM probabilities with any human behavioral measure, we need a *linking theory* between them to make the two quantities

[1]Our code is available at https://github.com/lindiatjuatja/morcela.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2173–2186

comparable. Although the existence of a relationship between probability and acceptability has been subject to debate (Quine, 1960; Chomsky, 1969; Pereira, 2000; Norvig, 2017), an influential proposal by Lau et al. (2017) hypothesizes that sentence-level LM probabilities largely reflect linguistic acceptability, but are influenced by word frequency and sentence length in ways that humans are largely robust to. Thus, a linking theory between LM probabilities and human acceptability scores should somehow control for these factors. Out of the various functions they used to control for length and frequency, Lau et al. (2017) find that the *syntactic log-odds ratio* (SLOR, Pauls and Klein, 2012) served as the best linking theory between acceptability and probabilities from n-gram LMs and simple recurrent LMs (Elman, 1990). SLOR controls for unigram frequency and length in a uniform manner across LMs by dividing the probability of the sentence under the LM by the joint unigram frequency, then averaging over all tokens to control for length. However, it is not clear *a priori* that these model-agnostic transformations are the appropriate ones to link LM probabilities and acceptability judgments, nor that these transformations should be held constant across different LMs.

In this work, we first show that the model-agnostic transformations in SLOR may severely underestimate LM probability correlations with human acceptability judgments. We propose a new linking theory, *Magnitude-Optimized Regression for Controlling Effects on Linguistic Acceptability* (MORCELA), a parameterized linking theory where the effect sizes of length and unigram frequency are automatically estimated from human acceptability judgment data. Our experiments first show that MORCELA significantly outperforms SLOR in predicting human acceptability judgments from probabilities calculated by Transformer LMs (Vaswani et al., 2017) from the Pythia (Biderman et al., 2023) and OPT (Zhang et al., 2022) families. Our results show a relationship with scale, where larger models exhibit greater correlation with human judgments compared to smaller models in the same family, using the same linking theory. Examining the estimated optimal parameter values of MORCELA reveals that larger models are more robust to length and unigram frequency effects, and thus their probabilities require a lower degree of adjustment when comparing them to human acceptability judgments. We show in particular that larger models' lower reliance on unigram frequency is driven by their improved ability to predict rare words given appropriate context. These results demonstrate that when comparing probability-based LM acceptability judgments to those of humans, controls for factors like length and unigram frequency should be made on a per-model basis.

2 MORCELA: Acceptability Judgments from LM Probabilities

To evaluate LMs according to their ability to predict human judgments of linguistic acceptability, we need a *linking function* that takes as input the probability of the sentence under a LM and outputs an acceptability score, which we then correlate with human judgments. This linking function should account for the effects of length and unigram frequency, as noted by Lau et al. (2017), which impact LM probabilities in predictable ways that may cause them to deviate from human judgments. Specifically, longer sentences will be assigned a lower probability than to any strictly smaller prefix of the sentence and a sentence containing a rare token will likely have a lower probability compared to one containing a more frequent one, all else being equal.

We propose MORCELA (**M**agnitude-**O**ptimized **R**egression for **C**ontrolling **E**ffects on **L**inguistic **A**cceptability), a parameterized linking function given by

$$\text{acceptability} \propto \frac{p - \beta u + \gamma}{\ell} \tag{1}$$

where ℓ is the length of a sentence, p is the sentence's LM log probability, u is the sentence's unigram log probability, and β and γ are learnable parameters. The values of β and γ can be estimated from human acceptability judgment data by fitting a linear regression model

$$\text{acceptability} \approx a\frac{p}{\ell} + b\frac{u}{\ell} + c\frac{1}{\ell} + d$$

and taking $\beta = -b/a$ and $\gamma = c/a$. MORCELA improves upon the *syntactic log-odds ratio* (SLOR, Pauls and Klein, 2012), widely used as a linking function for predicting acceptability judgments (Lau et al., 2017, 2020; Sprouse et al., 2018; Kann et al., 2018; Kumar et al., 2020; Misra and Mahowald, 2024; Lu et al., 2024), by allowing for arbitrary linear relationships between the variables p and u via the parameters β and γ. We argue here that MORCELA mitigates overcorrections for

length and frequency effects from SLOR, and our main experiment shows that MORCELA scores are significantly more correlated with z-normalized human acceptability ratings than SLOR scores.

2.1 Prior Work: SLOR

SLOR was proposed as a linking function by Lau et al. (2017), who compare it against several other linking functions on acceptability judgments from multiple sources. SLOR predicts the acceptability rating of a sentence to be given by

$$\text{acceptability} \propto \frac{p - u}{\ell} \tag{2}$$

where p, u, and ℓ are defined as above. Intuitively, the SLOR score of a sentence is the average log probability assigned to its tokens, adjusted for frequency. It uses p as an initial estimate of acceptability, and incorporates ℓ and u under the assumption that long sentences and sentences with rare words have lower LM probabilities, but not lower human acceptability judgments, than short sentences or sentences without rare words.

2.2 MORCELA

The normalizations involved in SLOR are based on specific assumptions about the impact of length and frequency on LM probabilities, namely that LM probabilities and unigram frequencies should have equal importance on the resulting acceptability score, and that taking the geometric mean of each token's probability under the LM largely eliminates the impact of sentence length. However, it is unclear *a priori* whether these assumptions hold, and furthermore whether they hold uniformly across models.

MORCELA relaxes these assumptions by allowing for arbitrary linear relationships between LM probabilities and unigram frequencies, expressed via the parameters β and γ. These parameters can be understood as mitigating overcorrections for frequency and length effects by SLOR, respectively. To see this, let us rewrite equations (1) and (2) as follows:

$$\text{MORCELA} = \text{SLOR} + \underbrace{(1 - \beta)\frac{u}{\ell}}_{\text{frequency}} + \underbrace{\gamma\frac{1}{\ell}}_{\text{length}}$$

The "frequency" term, controlled by β, adjusts SLOR according to the average unigram probability of the sentence's tokens, while the "length" term, controlled by γ, provides an adjustment to SLOR that is inversely proportional to the sentence's length.

3 Main Experiment

How much does optimizing the relative effect of length and unigram frequency via MORCELA impact fit of LM acceptability scores to human judgments? To investigate this, we correlate the LM acceptability scores from MORCELA to gradient human judgments across LMs of varying sizes and compare the resulting correlation to two baseline linking functions: log probabilities and SLOR.

3.1 Models

We evaluate models of varying sizes from the Pythia Scaling Suite (Biderman et al., 2023) and Open Pre-Trained Transformers (OPT, Zhang et al. 2022) families. Both families of models are decoder-only autoregressive transformer LMs. We test all eight sizes of Pythia models (70M–12B parameters) and all but the two largest OPT models (125M–30B parameters). Models within each family were trained on the same pretraining corpus: Pythia models were trained on The Pile (Gao et al., 2020), whereas the OPT models were trained on a concatenation of data from subsets of the RoBERTa training corpus (Zhuang et al., 2021), The Pile (Gao et al., 2020), and PushShift.io Reddit (Baumgartner et al., 2020; Roller et al., 2021). Both families of models saw $\approx$ 300B tokens during training. We list additional model hyperparameter details in Appendix A.

3.2 Unigram Frequency Estimation

As input to the various linking functions, we need to calculate the LM probability p, the unigram probability u, and length ℓ of the sentence in tokens.[2] To calculate u, we need to measure the frequency of tokens as they appear in the training corpus of the LM. This is easily done for the Pythia models as the training corpus (The Pile, Gao et al. 2020) is publicly available. However, since this is not the case for the OPT models, we instead look at text generated from the largest OPT model we test

[2]Technically, for the OPT models this involves calculating the probability conditioned on only the beginning of sequence (BOS) token. Pythia models were trained without a BOS token, so to calculate p we do not append an additional BOS token to the input sequence, and instead exclude the first token's probability when calculating p and u, though subsequent tokens in p are calculated with the first token provided in the context. We also exclude the first token when calculating the length of the sentence ℓ.

(OPT-30B) as a proxy for the training corpus, with the intuition that the distribution of generated text is largely similar to that of its pretraining corpus. We estimate token unigram frequency by aggregating the probability of a token being generated in each position of a sequence of arbitrary length, then averaging this value over a large number of generated sequences ($n = 100000$). We provide additional details for this estimation process in Appendix B.

3.3 Dataset

We evaluate acceptability predictions and fit MORCELA parameters using data from Sprouse et al. (2013), which contain acceptability judgments for example sentences from the *Linguistic Inquiry*, a leading theoretical linguistics journal. We use judgments reported on a 1–7 Likert scale by native English speakers in the United States, z-normalized by annotator.[3] To ensure balance, we limit our dataset to minimal pairs of acceptable and unacceptable sentences that Sprouse et al. (2013) have determined to have equal semantic plausibility. After filtering out unpaired sentences as well as sentences with missing data, we obtain a final dataset of acceptability judgments for 1450 sentences.

3.4 Fitting and Evaluating Linking Functions

For each linking function we examine, we calculate the correlation (Pearson's r) between the LM acceptability scores generated by the function and z-normalized Likert scale human judgments. For functions with learned parameters, we train and evaluate linear regression models using 5-fold cross validation (with shuffling), and report the average correlation over each test fold. To calculate an upper bound for correlation, we randomly split judgments per sentence into two groups, which yields an inter-group correlation of $r = 0.860$.

4 Results

We first compare MORCELA to two baseline linking functions (raw log probabilities and SLOR), then assess the impact that parameterizing either unigram frequency or length has on correlation with human judgments.

[3]For example, the acceptability score of 1.19 for sentence (1) in Figure 1 means that this sentence was judged to be 1.19 standard deviations more acceptable than the mean acceptability of sentences in the dataset.

4.1 MORCELA vs. SLOR

Figure 2 shows correlation of acceptability scores using log probabilites, SLOR, and MORCELA across varying sizes of Pythia and OPT models. There is a general increasing monotonic trend with size, though to a lesser degree with the OPT models as smaller OPT models have a higher correlation with humans compared to similarly sized Pythia models. Nevertheless, overall trends regarding the relative performances of the different linking functions, and how they change with scale, are similar.

Across all models, the addition of the two learned parameters in MORCELA leads to a significant gain in correlation with human judgments. We observe up to $+\Delta 0.33$ increase from raw log probabilities and $+\Delta 0.17$ from SLOR with Pythia-6.9B and 12B, which amounts to a 46% relative error reduction from SLOR with respect to the inter-group correlation upper bound. As models get larger (and correspondingly, generally better at predicting human judgments), we also observe greater differences in correlation between SLOR and MORCELA. This suggests that larger models with higher baseline correlation with humans (as demonstrated by higher raw log probability correlation) reap greater benefits from the additional parameterization.

4.2 Parameter Ablation Study

The performance gap between SLOR and MORCELA clearly shows that the assumed values ($\beta = 1, \gamma = 0$) in SLOR are non-optimal across all models, and especially so for larger ones. But how important is the optimization of either parameter in improving fit to human judgments? To answer this, we perform ablations to MORCELA, where either length or unigram normalization are set to their default values ($\beta = 1, \gamma = 0$) and the other is allowed to vary.

The results of these ablations are shown in Figure 3, where MORCELA$_{\beta=1}$ optimizes the value of the length-normalized intercept γ given the default weight for unigram frequency, and vice versa for MORCELA$_{\gamma=0}$. We find that optimizing the unigram coefficient β without a length intercept (MORCELA$_{\gamma=0}$) leads to little to no gain in performance. In contrast, adding the length-normalized intercept γ while keeping the unigram coefficient β fixed (MORCELA$_{\beta=1}$) can—for the smallest Pythia models—reach the performance of MORCELA, though the difference between

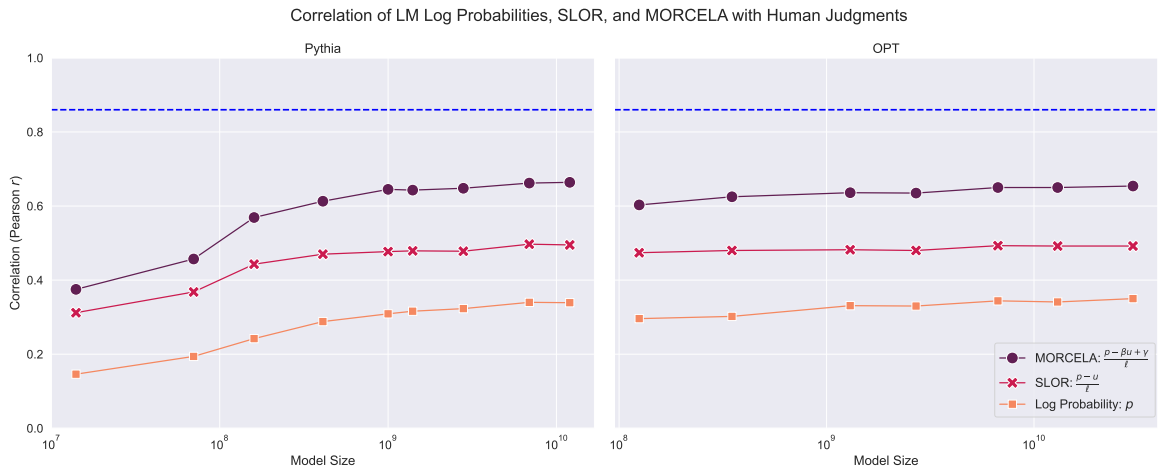

Figure 2: Correlation of LM acceptability scores with human judgments using raw log probabilities, SLOR, and MORCELA. The blue dashed line indicates inter-group correlation between randomly partitioned participant ratings ($r = 0.860$). MORCELA consistently outperforms SLOR, with up to $+\Delta 0.17$ gain in correlation.

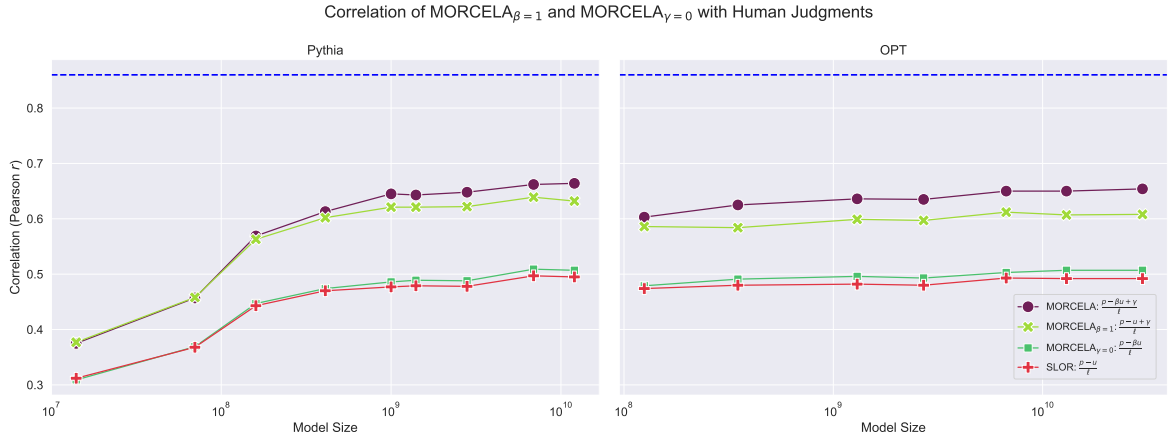

Figure 3: Comparison of SLOR and MORCELA with linking functions where either length or unigram normalization are set to their default values ($\beta = 1$, $\gamma = 0$) and the other is allowed to vary. Allowing the length-normalized intercept γ to vary on its own (MORCELA$_{\beta=1}$) leads to similar performance to MORCELA, though models still benefit from additionally varying the unigram coefficient β.

MORCELA$_{\beta=1}$ and MORCELA tends to grow as models become larger.

A natural question that follows is whether this gain in correlation is significant enough to warrant the extra degree of freedom that comes with additionally varying β. Using two model selection criteria—Akaike information criterion (AIC, Akaike 1974) and Bayes information criterion (BIC, Schwarz 1978), which take into account both model fit and number of predictors—we find that it is: MORCELA is preferred over MORCELA$_{\beta=1}$ for all but one LM (Pythia-14M).[4] Thus, while the addition of the length-normalized intercept γ on its own can significantly correlation with human judgments, adding the unigram coefficient β in conjunction with γ is still preferred.

4.3 Trends in Length and Unigram Frequency Effects Across Models

The above results tell us that the coefficients used by SLOR to control for length and unigram frequency—namely an equal weighting of LM log probabilities and unigram log probabilities and the lack of a length-normalized intercept—are non-optimal, and that the impact of turning these controls into tuned parameters impacts models to varying degrees. However, looking at correlation alone does not tell us about *how* these controls are non-optimal. We inspect the learned optimal values of the unigram coefficient β and length-normalized

[4]We include details for calculation and values of AIC and BIC for each linking function per LM in Appendix C.2.

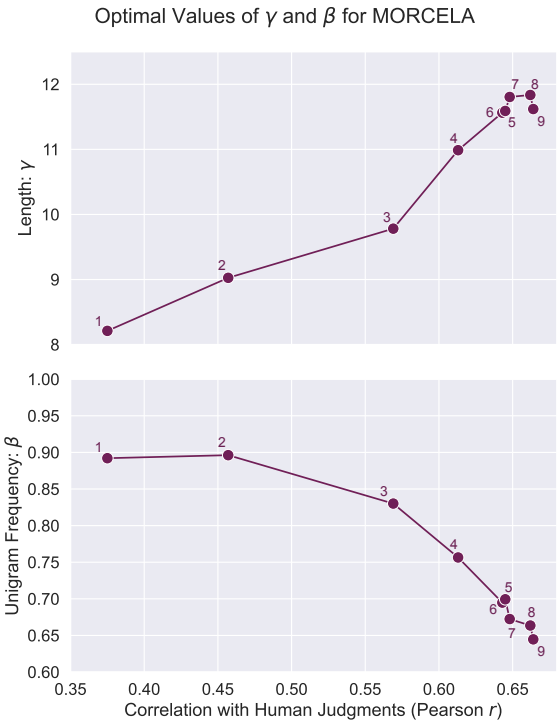

Figure 4: Optimal values of γ (top) and β (bottom) versus correlation with human judgments across for MORCELA, fit using all of the data. As models become better correlated with human judgments, γ increases and β decreases. Points shown are from Pythia models (numbered from smallest to largest), though these trends generally hold for the OPT models as well (see Appendix C).

intercept γ to see whether the assumed values in SLOR are under- or overestimating the impact of these confounds across models.

Figure 4 shows the optimal values of γ and β for MORCELA, fit using all the data. All values of γ are positive, and grow larger as models become better correlated with human judgments. Qualitatively, the observation that γ is positive indicates that naively normalizing by dividing by the length of the sentence is actually *overcorrecting* for length. A larger positive γ more dramatically increases acceptability scores of shorter sentences relative to longer ones. This counteracts the division by length, which on its own increases the scores of longer sentences.

Like γ, for the unigram coefficient β we see a trend with respect to correlation (and thus, to a large extent, model size), though in this case the value of β decreases as correlation increases. Notably, all values of β are less than the default value of 1. This, too, shows that the assumed impact of unigram frequency as used in SLOR is an overestimate, and that larger models tend to require less adjustment for unigram frequency. Additionally, we find that trends across both γ and β also hold among various other linking functions that parameterize β and/or γ, as shown in Figure 6 in Appendix C.

5 Ability to Predict Infrequent Tokens Explains Impact of Unigram Frequency

As we have just shown, as models get larger and better at predicting human acceptability judgments, the smaller the relative importance of unigram frequency becomes. One possible explanation for this is that models that are better predictors of acceptability are so because they are better at predicting more infrequent tokens in context, and as a result are more robust to the effect of unigram frequency. The intuition behind this is that while some tokens may be very rare within the distribution of the entire corpus (e.g. names of chemical compounds), they may be relatively frequent given a specific context (e.g. within a scientific article). Thus, if a model is better able to predict such cases of tokens by utilizing context, they should no longer need to be controlled as heavily for unigram frequency.

To test this hypothesis, we first need a way to quantify the ability of a LM to predict rarer tokens in context. We operationalize this by correlating the LM's conditional log likelihood over instances of tokens with the unigram log-probability of those tokens. As our corpus to calculate conditional log likelihood over, we use a portion ($\sim$ 100 million tokens) of the test set of The Pile (Gao et al., 2020), the training corpus of the Pythia models. For unigram log-probabilities, we use counts from the entire training split (as in our calculations of $p_U(S)$). We do this for all sizes of models in the Pythia suite. To calculate conditional log likelihood, we use a sliding window of the max sequence length of the Pythia models (2048 tokens) with a stride of 1024. As before, since Pythia was trained without a BOS token, the log likelihood of the first token in a document is not considered.

If our hypothesis—that LMs better at predicting rarer tokens in context are more robust to unigram frequency effects—holds, we would expect that models that are worse at predicting human acceptability judgments have *lower* conditional log likelihood for more infrequent tokens compared to models that are better fits to judgments, and vice versa. In other words, if we were to plot conditional

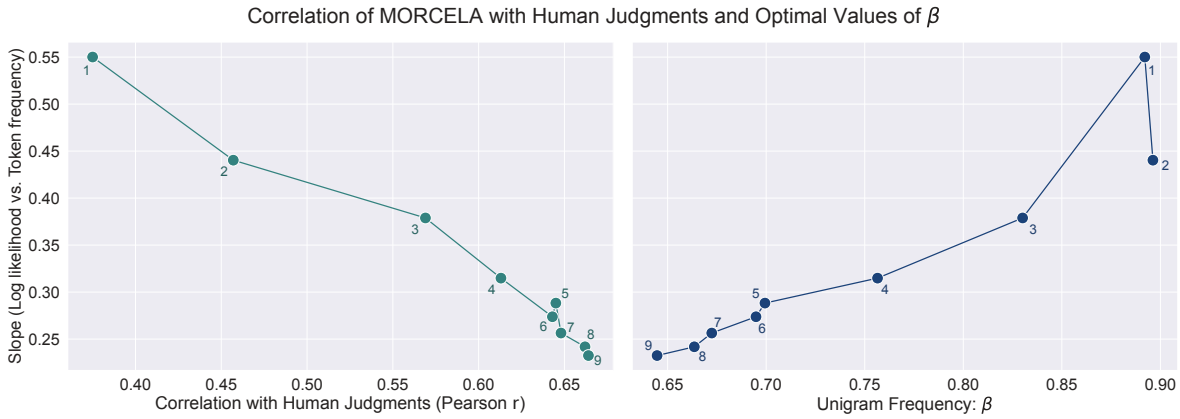

Figure 5: The slope of LM conditional log likelihood per token vs. token unigram frequency plotted against the correlation (average correlation using 5-fold CV) and optimal unigram coefficient β values of MORCELA (fit using all the data). Each point represents a single Pythia model, numbered from smallest to largest. Models that are better at predicting rarer tokens in context should have a *lower* slope value, which we find is correlated with higher correlation with human judgments and generally lower β.

log probability versus unigram log-probability, we should see that models with higher correlation with human judgments and lower values of β have a steeper positive slope.

We find that this prediction largely holds, as shown in Figure 5. In general, as models get better at predicting rarer tokens in context, i.e. a higher log likelihood on aggregate for rarer tokens and thus a less positive slope between log likelihood and token unigram frequency, they show greater correlation with human judgments as well as increasingly lower values of β.

6 Discussion

Our method MORCELA lies within the context of a large body of work that examines how well (neural) language models reflect human-like language processing. We now discuss our work's relation to previous studies evaluating LMs as psycholinguistic subjects and the application of parameterized linking theories in this setting, and compare our results to related findings from comparisons of LM surprisal to reading times.

6.1 Methods for Evaluating LMs as Psycholinguistic Subjects

A common methodological setup in comparing the linguistic capabilities of language models to humans is the targeted syntactic evaluation paradigm (Linzen et al., 2016; Marvin and Linzen, 2018; Gulordava et al., 2018, inter alia). In this setup, probability-based LM judgments are considered consistent with those of humans if they assign a higher probability to acceptable sentences compared to their minimally different unacceptable counterparts. Datasets such as BLiMP (Warstadt et al., 2020), as well as evaluation frameworks like SyntaxGym (Gauthier et al., 2020), follow this paradigm. A notable feature of many of these works is the use of a forced choice, binary judgment setup, which comes with the assumption that one sentence is more acceptable than the other. Thus, in these evaluations the information about relative differences within and across pairs is not present, though the use of minimal pairs in itself does not require this. For example, work by Leong and Linzen (2023) correlate the difference in LM probabilities between the acceptable and unacceptable sentences within a minimal pair with the difference in gradient human judgments. Nevertheless, the use of minimal pairs makes strictly controlling for length and frequency effects possible at the data construction stage.

Our experimental setting is most similar to work such as Lau et al. (2017), Vázquez Martínez et al. (2023), Misra and Mahowald (2024), and Lu et al. (2024) which instead correlate (transformed) LM probabilities directly to gradient human judgments on individual sentences. However, while this setting more easily allows for greater granularity in judgments across a wider range of examples, it requires a linking function that either assumes or estimates the effects of various factors on LM probabilities. Commonly used linking functions in these

settings, such as SLOR, bake in assumptions about length and frequency effects, namely that length can be controlled for by dividing by the number of tokens in the sequence, and that unigram frequency and LM probabilities should have equal weight. Our work challenges these assumptions by instead estimating this effect directly from acceptability data via a parameterized linking function.

6.2 Linking Functions

More generally, a linking function between measured quantities from LMs and humans is a way to control for asymmetries in effects of factors external to the construct of interest. In the case of LM probabilities and acceptability judgments, we expect that LM probabilities over sentences are impacted by unigram frequency and the length in ways that humans are thought to be largely robust to (Lau et al., 2017; Goodall, 2021).

We can draw parallels to methods in comparing LM word-level surprisal with measures of incremental sentence processing (e.g. eye tracking, reading and reaction times), where instead there are external effects on the human side, such as the length or predicatbility (estimated using a statistical language model) of a word in reading time experiments (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2021; Meister et al., 2021). However, unlike other works comparing LMs to human acceptability judgments that assume the strength and quality of effects (specifically of length and unigram frequency), it is standard for the parameters associated with the covariates in these studies are learned and fit per participant. MORCELA can be viewed as following a similar methodology, where we instead fit parameters per model to correct for model-side effects.

Nevertheless, MORCELA, like SLOR, still makes the assumption that the form of the relationship between LM probabilities and acceptability is log-linear. Meister et al. (2021) find evidence for a super-logarithmic relationship between LM probabilities and reading times, as well as binary acceptability judgments, differing from our gradient judgment setting; future work could explore other forms to fit between probabilities and gradient judgments.

6.3 Impact of Scale on Similarity with Humans

MORCELA demonstrates that the strength of length and unigram frequency effects (1) are not uniform across models, (2) are overestimated by the default values in SLOR, and (3) show a trend with scale. As models become larger and generally better predictors of human acceptability judgments (up to a certain point), the less they need to be controlled for unigram frequency effects. In contrast, prior work by Oh and Schuler (2023) observe the opposite trend with respect to scale when comparing LM surprisal with reading times, with larger models serving as poorer fits to humans. In follow-up work they found that this trend can be explained by frequency, with the inverse correlation between model size and reading times being the strongest amongst the least frequent words (Oh et al., 2024). Similar to our analysis, they show that this is driven by the ability of larger LMs to more accurately predict rare words.

We hypothesize that the seeming paradox between more human-like judgments vs. less human-like reading time predictions may be a consequence of the role of predictability in offline and online language processing. In the case of reading times, it may be that frequency effects at the word level are important for humans, and thus models can be "too good" at predicting rare words relative to humans, whereas this is may not the case—at least, to the same extent—in predicting acceptability judgments.

7 Conclusion

In this work, we reexamine the assumptions made by commonly used linking theories such as SLOR in evaluating LMs' fit to human acceptability judgments. We introduce a new linking theory, MORCELA, which parameterizes controls for length and unigram frequency, and learns the optimal values for these controls from acceptability data via linear regression. By adding two simple, interpretable parameters, MORCELA drastically improves correlation with human judgments compared to SLOR, showing that SLOR greatly underestimates correlation between LM and human acceptability scores. An inspection of the optimal values of these parameters shows that the magnitude of correction for confounds in SLOR overestimate the impact of frequency and length in Transformer LMs, and that this overestimation is greater as models grow larger. Finally, we show that LMs' robustness to unigram frequency effects can be explained by their ability to predict rarer words in context.

Our findings suggest that evaluations of probability-based LM acceptability judgments should account for model-specific qualities with respect to factors like frequency and length, and that doing so reveals that LMs may be better correlated with human judgments than previously thought. However, there is still a sizable gap between the maximum correlation between LMs and human judgments and the correlation between annotators. Future work could investigate what additional factors/transformations could lead to closer correspondence between LMs and humans, and further integrate these insights into training more cognitively plausible models.

8 Limitations

Our evaluations are limited to two model families trained on predominantly English data on judgments of English sentences by English AMT workers with a US-based location (Sprouse et al., 2013). Thus, there is no guarantee that our results would hold for models or data in other languages/in a multilingual setting. The size of our data ($n = 1450$) is relatively small, so while we expect general trends with respect to scale to hold, the actual values of the optimized parameters may change with larger and more varied data.

9 Ethics Statement

This work uses publicly available models and data and does not release any new artifacts. For the human acceptability judgment data, we point readers to Sprouse et al. (2013) for details on data collection. We do not foresee any negative ethical consequences for our work.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. BCS-2114505 and a gift from Amazon AWS. This work was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. We would also like to thank Kanishka Misra, members of NeuLab and Caplab, and our reviewers for their helpful feedback on our work.

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

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.

Noam Chomsky. 1957. *Syntactic Structures*, 1 edition. Mouton, The Hague, Netherlands.

Noam Chomsky. 1969. *Quine's Empirical Assumptions*, pages 53–68. Number 21 in Synthese Library. Springer Dordrecht, Dordrecht, Netherlands.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

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*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Grant Goodall. 2021. *Sentence Acceptability Experiments: What, How, and Why*, page 7–38. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

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, Salt Lake City, UT. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the*

22nd Conference on Computational Natural Language Learning, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. Association for Computational Linguistics.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Jiayi Lu, Jonathan Merchan, Lian Wang, and Judith Degen. 2024. Can syntactic log-odds ratio predict acceptability and satiation? In *Proceedings of the Society for Computation in Linguistics 2024*, pages 10–19, Irvine, CA. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *arXiv preprint arXiv:2403.19827*.

Peter Norvig. 2017. On Chomsky and the Two Cultures of Statistical Learning. In Wolfgang Pietsch, Jörg Wernecke, and Maximilian Ott, editors, *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data*, pages 61–83. Springer Fachmedien, Wiesbaden, Germany.

Byung-Doh Oh and William Schuler. 2023. 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, 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*, pages 2644–2663.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Fernando Pereira. 2000. Formal Grammar and Information Theory: Together Again? *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

Willard Van Orman Quine. 1960. *Word and Object*. MIT Press, Cambridge, MA.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Carson T Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Gideon Schwarz. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115(11):1497–1524. Data in Theoretical Linguistics.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*.

Jon Sprouse. 2013. Acceptability judgments.

Jon Sprouse. 2015. Three open questions in experimental syntax. *Linguistics Vanguard*, 1(1):89–100.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C. Berwick. 2018. Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.

Héctor Vázquez Martínez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Gotlieb Wilcox, Pranali Vani, and Roger P Levy. 2021. A targeted assessment of incremental processing in neural languagemodels and humans. *arXiv preprint arXiv:2106.03232*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Model Hyperparameters

Hyperparameters of the Pythia and OPT models examined in this work are shown in Table 1.

Model Variant	#L	#H	d_{model}	#Parameters
OPT 125M	12	12	768	∼125M
OPT 350M	24	16	1024	∼350M
OPT 1.3B	24	32	2048	∼1.3B
OPT 2.7B	32	32	2560	∼2.7B
OPT 6.7B	32	32	4096	∼6.7B
OPT 13B	40	40	5120	∼13B
OPT 30B	48	56	7168	∼30B
Pythia 14M	6	4	512	∼14M
Pythia 70M	6	8	512	∼70M
Pythia 160M	12	12	768	∼160M
Pythia 410M	24	16	1024	∼410M
Pythia 1B	16	8	2048	∼1B
Pythia 1.4B	24	16	2048	∼1.4B
Pythia 2.8B	32	32	2560	∼2.8B
Pythia 6.9B	32	32	4096	∼6.9B
Pythia 12B	36	40	5120	∼12B

Table 1: Hyperparameters of model variants examined in this work. #L, #H, and d_{model} respectively refer to number of layers, number of attention heads per layer, and embedding size.

B Unigrahams Estimator

This section proposes a novel technique, the *unigrahams estimator*, for estimating the unigram distribution of an LM without access to its training corpus. The unigrahams estimator estimates the unigram distribution for text generated by an LM of a given length ℓ, which we assume approximates the unigram distribution of the LM's training corpus. Given a prompt x, we generate a number n of responses to x, counting the number of occurrences of each token in each response. These frequency counts are weighted by LM probabilities in the following sense: if, during the generation process, the LM assigns a next-token probability of q to token w, then we assume that q instances of w have occurred in this position of the generated text. A full description of the unigrahams estimator is given in Algorithm 1, and its correctness is proven in Subsection B.1.

Algorithm 1 Unigrahams Estimator

Inputs: LM m, prompt x, response length ℓ
Parameters: Number of samples n, vocabulary $\mathbb{V}$
Output: Frequency estimates $f_\ell(\cdot \mid x) : \mathbb{V} \to \mathbb{R}$

for all $w \in \mathbb{V}$ **do**
  $f_\ell(w \mid x) \leftarrow 0$
repeat n **times**
  $y \leftarrow x$
  repeat
    for all $w \in \mathbb{V}$ **do**
      $f_\ell(w \mid x) \leftarrow f_\ell(w \mid x) + \frac{\mathbb{P}_m[w|y]}{n}$
    Sample a token $v \sim \mathbb{P}_m[\cdot \mid y]$
    $y \leftarrow yv$
  until $|y| = \ell + |x|$
return $f_\ell(\cdot \mid x)$

As mentioned in Subsection 3.2, we use the unigrahams estimator to estimate the unigram distribution of the OPT training corpus from the OPT-30B model. We use parameter values of $n = 10^6$ and $\ell = 34$, the latter being the length of the longest sentence in our dataset of acceptability judgments from Sprouse et al. (2013).

B.1 Theoretical Analysis

We now prove the correctness of the unigrahams estimator. Let m be a LM, and let $\mathbb{P}_m[y \mid x]$ denote the probability that m will generate response y to prompt x. Let $|y|$ denote the length of y (in tokens), let $y_{:i}$ denote the first i tokens of y, and for each token w in vocabulary $\mathbb{V}$, let $|y|_w$ denote the number of occurrences of w in y. Below we give a formal definition of a LM's unigram distribution.

Definition 1. The *length-ℓ unigram frequency* of a token $w \in \mathbb{V}$ with respect to prompt x and LM m is the expected number of times w occurs in responses to x of length ℓ:

$$f_\ell(w \mid x) := \mathbb{E}_{|y|=\ell}[|y|_w]$$

where $y \sim \mathbb{P}_m[\cdot \mid x]$.

Our goal is to show that the unigrahams estimator is an unbiased estimator of $f_\ell(w \mid x)$ for each $w \in \mathbb{V}$. This result is stated as follows.

Theorem 1. *For all $\ell \geq 1$ and $w \in \mathbb{V}$,*

$$f_\ell(w \mid x) = \mathbb{E}_{|y|=\ell-1}\left[\sum_{i=0}^{\ell-1} \mathbb{P}_m[w \mid xy_{:i}]\right]$$

To see why Theorem 1 is the desired result, observe that Algorithm 1 returns the mean of

$$\sum_{i=0}^{\ell-1} \mathbb{P}_m[w \mid xy_{:i}]$$

for a sample of ys of length $\ell - 1$ drawn from $\mathbb{P}_m[\cdot \mid x]$. This is an unbiased estimator of the right-hand side of Theorem 1, whence it follows that proving Theorem 1 suffices for verifying the correctness of the unigrahams estimator.

To that end, we note that Theorem 1 is straightforwardly derived from the following lemma.

Lemma 1. *For all $\ell > 1$ and $w \in \mathbb{V}$,*

$$f_\ell(w \mid x) = f_{\ell-1}(w \mid x) + \mathbb{E}_{|y|=\ell-1}[\mathbb{P}_m[w \mid xy]]$$

Proof of Lemma 1. Observe:

$$\begin{aligned}
&f_\ell(w \mid x) \\
&= \sum_{|y|=\ell} |y|_w \mathbb{P}_m[y \mid x] \\
&= \sum_{|y|=\ell-1} \Bigg[(|y|_w + 1)\mathbb{P}_m[yw \mid x] + \\
&\qquad |y|_w \sum_{v \in \mathbb{V}\setminus\{w\}} \mathbb{P}_m[yv \mid x]\Bigg] \\
&= \sum_{|y|=\ell-1} (|y|_w + \mathbb{P}_m[w \mid xy])\mathbb{P}_m[y \mid x] \\
&= \mathbb{E}_{|y|=\ell-1}[|y|_w + \mathbb{P}_m[w \mid xy]] \\
&= f_\ell(w \mid x) + \mathbb{E}_{|y|=\ell-1}[\mathbb{P}_m[w \mid xy]].
\end{aligned}$$

□

Proof of Theorem 1. We induct on ℓ. To prove the $\ell = 1$ case, we simply observe that

$$f_1(w \mid x) = \mathbb{P}_m[w \mid x].$$

Now suppose Theorem 1 holds for some value of ℓ. Observe that

$$\begin{aligned}
&\mathbb{E}_{|y|=\ell-1}\left[\sum_{i=0}^{\ell-1} \mathbb{P}_m[w \mid xy_{:i}]\right] \\
&= \mathbb{E}_{|y|=\ell}\left[\sum_{i=0}^{\ell-1} \mathbb{P}_m[w \mid xy_{:i}]\right].
\end{aligned}$$

Thus, by Lemma 1 we have

$$\begin{aligned}
&f_{\ell+1}(w \mid x) \\
&= f_\ell(w \mid x) + \mathbb{E}_{|y|=\ell}[\mathbb{P}_m[w \mid xy]] \\
&= \mathbb{E}_{|y|=\ell}\left[\mathbb{P}_m[w \mid xy] + \sum_{i=0}^{\ell-1} \mathbb{P}_m[w \mid xy_{:i}]\right] \\
&= \mathbb{E}_{|y|=\ell}\left[\sum_{i=0}^{\ell} \mathbb{P}_m[w \mid xy_{:i}]\right]
\end{aligned}$$

as desired. □

C Additional Results

C.1 Optimal values of β and γ

Optimal values of MORCELA and the two ablations (MORCELA$_{\beta=1}$ and MORCELA$_{\gamma=0}$) for OPT and Pythia models are shown in Table 2 and visualized in Figure 6.

C.2 AIC and BIC Calculations

We calculate AIC and BIC for SLOR, MORCELA$_{\beta=1}$, MORCELA$_{\gamma=0}$, and MORCELA using the following formulas, where SSE is the sum of squared error and n and k are the sample size and number of predictor terms (including the intercept), respectively:

$$\text{AIC} = n * \ln \frac{\text{SSE}}{n} + 2k \tag{3}$$

$$\text{BIC} = n * \ln \frac{\text{SSE}}{n} + k * \ln n \quad (4)$$

AIC and BIC for each linking function for every model is reported in Table 3 and Table 4. Rows in each table are sorted in ascending order using BIC. For both AIC and BIC, a lower value is better.

Model Variant	β	γ	r
OPT 125M	0.743	13.410	0.606
OPT 350M	0.640	12.827	0.751
OPT 1.3B	0.657	14.103	0.637
OPT 2.7B	0.654	14.034	0.637
OPT 6.7B	0.649	14.267	0.648
OPT 13B	0.627	13.916	0.652
OPT 30B	0.611	14.246	0.653
Pythia 14M	0.892	8.211	0.375
Pythia 70M	0.896	9.026	0.457
Pythia 160M	0.830	9.782	0.569
Pythia 410M	0.756	10.987	0.613
Pythia 1B	0.699	11.590	0.645
Pythia 1.4B	0.695	11.563	0.643
Pythia 2.8B	0.672	11.805	0.648
Pythia 6.9B	0.664	11.863	0.662
Pythia 12B	0.645	11.619	0.664

Table 2: Optimal values of β and γ for each model, along with their correlation (Pearson r) with human judgments.

Size	Linking Function	AIC	BIC	SSE	Predictors
125M	MORCELA	-1384.3	-1363.2	555.1	4
	MORCELA$_{\beta=1}$	-1329.5	-1313.7	577.2	3
	MORCELA$_{\gamma=0}$	-1101.2	-1085.4	675.7	3
	SLOR	-1091.2	-1080.7	681.3	2
350M	MORCELA	-1448.8	-1427.7	530.9	4
	MORCELA$_{\beta=1}$	-1323.5	-1307.7	579.6	3
	MORCELA$_{\gamma=0}$	-1133.5	-1117.6	660.8	3
	SLOR	-1101.8	-1091.3	676.3	2
1.3B	MORCELA	-1472.9	-1451.8	522.2	4
	MORCELA$_{\beta=1}$	-1363.6	-1347.8	563.8	3
	MORCELA$_{\gamma=0}$	-1128.8	-1112.9	663.0	3
	SLOR	-1105.6	-1095.0	674.6	2
2.7B	MORCELA	-1472.6	-1451.4	522.3	4
	MORCELA$_{\beta=1}$	-1360.3	-1344.5	565.1	3
	MORCELA$_{\gamma=0}$	-1126.0	-1110.2	664.2	3
	SLOR	-1102.3	-1091.7	676.1	2
6.7B	MORCELA	-1520.8	-1499.7	505.2	4
	MORCELA$_{\beta=1}$	-1401.8	-1386.0	549.2	3
	MORCELA$_{\gamma=0}$	-1150.7	-1134.9	653.0	3
	SLOR	-1126.0	-1115.4	665.2	2
13B	MORCELA	-1524.5	-1503.4	503.9	4
	MORCELA$_{\beta=1}$	-1386.4	-1370.6	555.0	3
	MORCELA$_{\gamma=0}$	-1155.3	-1139.4	651.0	3
	SLOR	-1124.3	-1113.7	665.9	2
30B	MORCELA	-1535.6	-1514.4	500.1	4
	MORCELA$_{\beta=1}$	-1389.3	-1373.4	553.9	3
	MORCELA$_{\gamma=0}$	-1156.7	-1140.9	650.3	3
	SLOR	-1124.0	-1113.4	666.1	2

Table 3: AIC and BIC of various linking functions across OPT models.

Size	Linking Function	AIC	BIC	SSE	Predictors
14M	MORCELA$_{\beta=1}$	-934.8	-919.0	757.9	3
	MORCELA	-935.6	-914.5	756.4	4
	SLOR	-782.2	-771.6	843.1	2
	MORCELA$_{\gamma=0}$	-780.7	-764.8	842.8	3
70M	MORCELA	-1150.0	-1128.9	652.4	4
	MORCELA$_{\beta=1}$	-1141.2	-1125.3	657.3	3
	SLOR	-974.3	-963.8	738.5	2
	MORCELA$_{\gamma=0}$	-972.7	-956.9	738.3	3
160M	MORCELA	-1341.4	-1320.3	571.7	4
	MORCELA$_{\beta=1}$	-1307.7	-1291.8	586.0	3
	SLOR	-1044.9	-1034.3	703.4	2
	MORCELA$_{\gamma=0}$	-1045.6	-1029.7	702.1	3
410M	MORCELA	-1464.9	-1443.8	525.1	4
	MORCELA$_{\beta=1}$	-1395.6	-1379.8	551.5	3
	MORCELA$_{\gamma=0}$	-1105.2	-1089.3	673.8	3
	SLOR	-1099.5	-1089.0	677.4	2
1B	MORCELA	-1454.0	-1432.9	529.0	4
	MORCELA$_{\beta=1}$	-1352.5	-1336.7	568.2	3
	MORCELA$_{\gamma=0}$	-1090.2	-1074.3	680.9	3
	SLOR	-1078.0	-1067.4	687.5	2
1.4B	MORCELA	-1446.0	-1424.9	531.9	4
	MORCELA$_{\beta=1}$	-1353.0	-1337.1	568.0	3
	MORCELA$_{\gamma=0}$	-1082.8	-1067.0	684.3	3
	SLOR	-1072.6	-1062.1	690.1	2
2.8B	MORCELA	-1527.3	-1506.1	503.0	4
	MORCELA$_{\beta=1}$	-1417.1	-1401.3	543.4	3
	MORCELA$_{\gamma=0}$	-1113.4	-1097.6	670.0	3
	SLOR	-1101.8	-1091.2	676.3	2
6.9B	MORCELA	-1560.6	-1539.5	491.5	4
	MORCELA$_{\beta=1}$	-1451.9	-1436.1	530.5	3
	MORCELA$_{\gamma=0}$	-1128.0	-1112.1	663.3	3
	SLOR	-1117.4	-1106.8	669.1	2
12B	MORCELA	-1539.3	-1518.1	498.8	4
	MORCELA$_{\beta=1}$	-1413.2	-1397.3	544.9	3
	MORCELA$_{\gamma=0}$	-1113.5	-1097.6	670.0	3
	SLOR	-1099.2	-1088.6	677.6	2

Table 4: AIC and BIC of various linking functions across Pythia models.

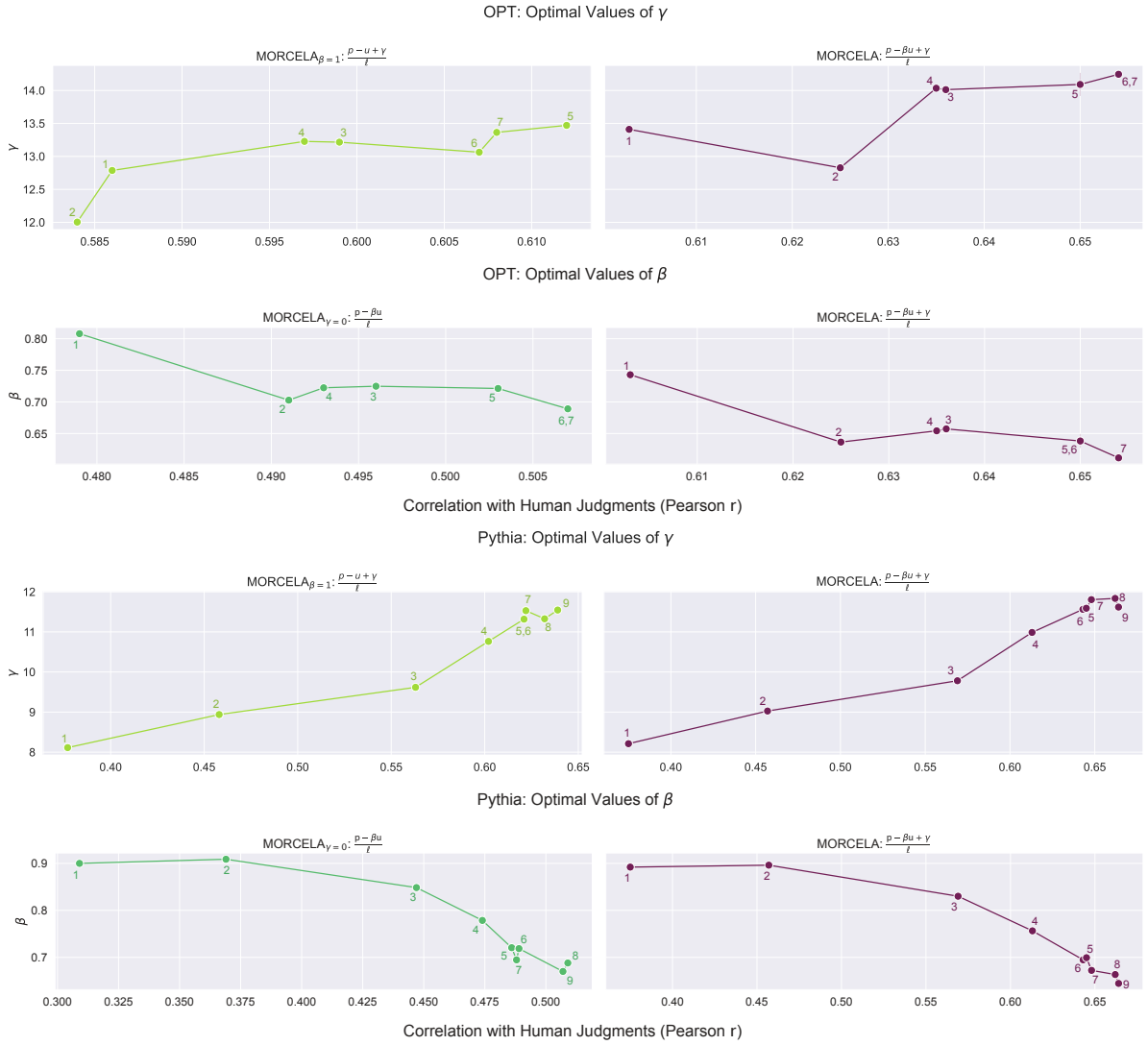

Figure 6: Optimal values of γ and β versus correlation with human judgments across linking theories. Top set of plots are for the OPT models, bottom are for Pythia. Points are numbered in order of model size (smallest to largest).