arXiv:2411.02528

What Goes Into a LM Acceptability Judgment?

Utrecht University

22 January 2026

Paper contribution

- Grammaticality is normally tested based on acceptability of utterances. *I is presenting*
- When we use LM probabilities as "acceptability judgments", they are confounded by**(i)** sentence length and **(ii)** lexical frequency.
- Normalization factors need to be model-aware, depending on how well the model encodes grammaticality.
- This paper proposes a **parameterized linking theory** that *learns* how much to correct for confounds **per size/architecture** from human judgment data.

Reading question for us: is this a real modeling insight, or mainly a better-calibrated score?

Setup: acceptability vs. probability

- Data: human acceptability judgments (Likert, z-normalized).
- We want a **linking function** that maps LM quantities to an acceptability score.
- Two predictable confounds in LM sentence log-probability p:
  - **Length:** longer sequences have lower total log-probability.
  - **Frequency:** rare words contribute large negative log-probability.
- Humans are less sensitive to these two factors (at least in this evaluation framing).

Baseline linking function: SLOR

- SLOR (Pauls & Klein; Lau et al. 2017):

$$\text{SLOR} \propto \frac{p - u}{\ell}$$

- p: LM sentence log-probability (sum of token log-probs)
- u: unigram log-probability of the token sequence (frequency baseline)
- ℓ: sentence length in tokens
- No adjustment for model-specific tokenization / lexical competence.

MORCELA: make the corrections learnable

- MORCELA linking function:

$$\text{MORCELA} \propto \frac{p - \beta u + \gamma}{\ell}$$

- β: how strongly to correct for unigram frequency.
- γ: a length-normalized intercept (acts like a length correction beyond dividing by ℓ).
- Estimation: fit linear regression on human acceptability

$$\text{acc} \approx a\frac{p}{\ell} + b\frac{u}{\ell} + c\frac{1}{\ell} + d, \qquad \beta = -b/a,\ \gamma = c/a.$$

Interpretation: "SLOR + two knobs"

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

- If $\beta < 1$: SLOR **over-corrects** for frequency; MORCELA adds some u/ℓ back.
- If $\gamma > 0$: dividing by length **over-corrects** for length; MORCELA compensates.
- Key claim: **optimal** corrections differ **by model** (and vary with scale).

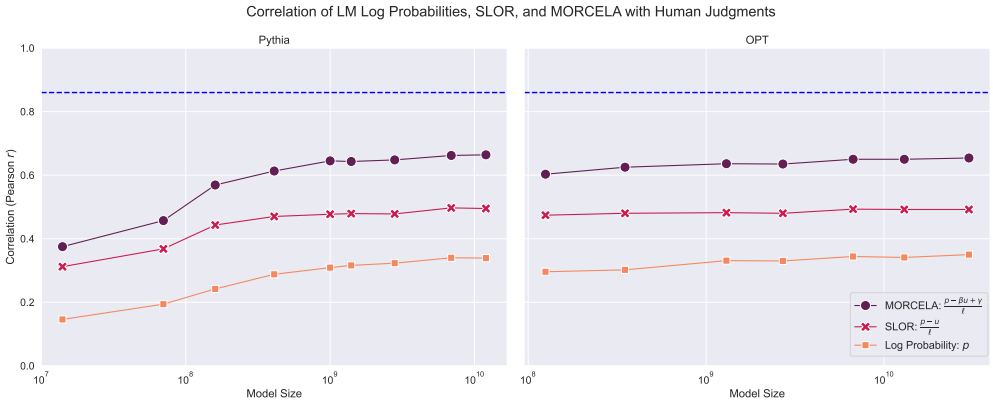
Data and evaluation

- Human judgments: Sprouse et al. (2013), sentential examples from *Linguistic Inquiry*.
  - Likert 1–7, z-normalized per annotator
  - filtered to minimal pairs; final: **1450 sentences**
- Metric: Pearson correlation between LM score and human rating.
- 5-fold CV (train β, γ; evaluate correlation on held-out fold).
- Upper bound: split-human correlation reported as $r = 0.860$.

Models and unigram frequency estimation

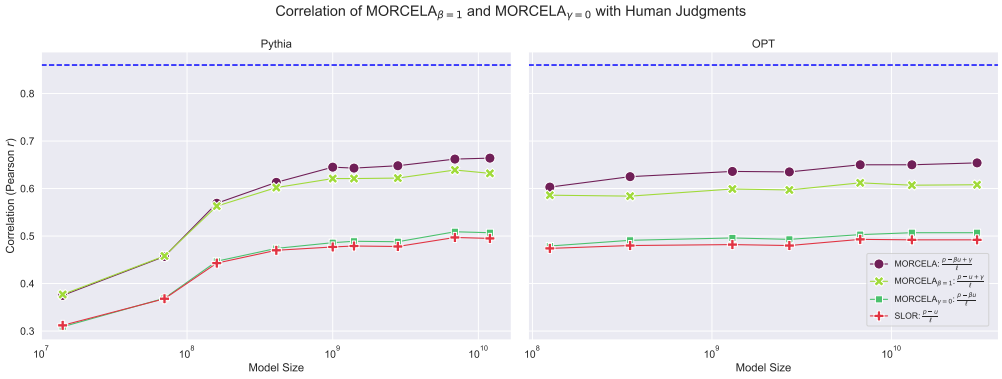
- Two decoder-only transformer families:
  - **Pythia** (70M–12B; trained on The Pile)
  - **OPT** (125M–30B tested)
- Need unigram probabilities u for each model/tokenizer.
  - Pythia: estimate from The Pile directly.
  - OPT: training data not public $\rightarrow$ estimate from OPT-30B generated text (100k sequences) as a proxy.

Main result: better correlation with H. ratings



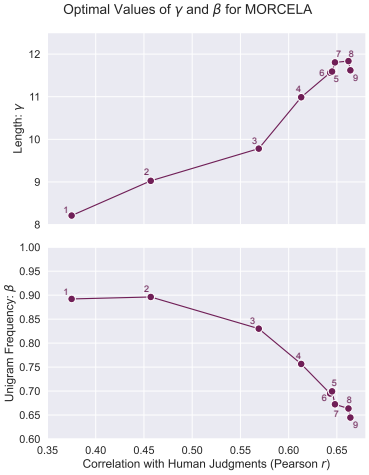
MORCELA consistently ends up closer to human judgements; reported gains up to $+0.17$ over SLOR.

Ablating the trained parameters



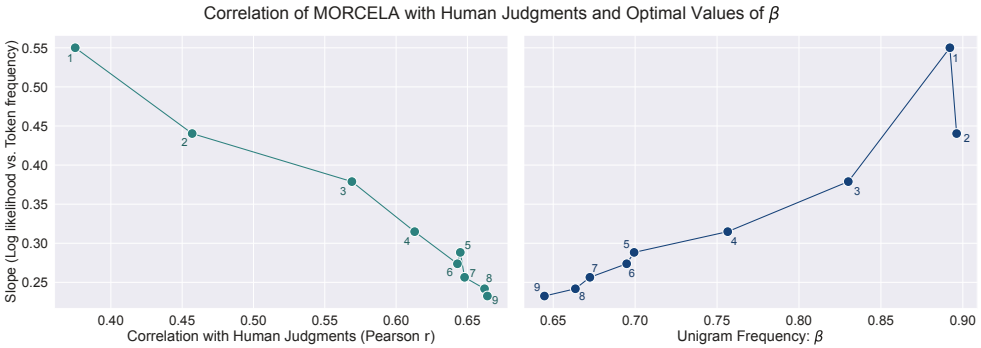
Correlations for fixed γ and fixed β. Fixing γ (length intercept factor) draws the correlation down; fixing β makes less impact; more noticeably, when models are large.

Estimated optimal corrections (qualitative takeaway)



- Empirically: $\gamma > 0$ and grows with model quality/size.
- Empirically: $\beta < 1$ (SLOR over-corrects frequency) and tends to *decrease* as models scale.
- Interpretation: larger LMs are **less sensitive to unigram frequency** when predicting acceptability.

Mechanistic story: rare-word prediction in context



Models that are better at predicting rare tokens (**log likelihood** less associated with **unigram frequency**) show higher acceptability correlation, and lower β.

Where to be cautious (potential issues)

- **Supervised calibration:** MORCELA uses human judgments to tune β, γ.
  - Raises the question: are we measuring LM "linguistic competence" or *fit to this dataset*?
  - Cross-validation helps, but generalization to other acceptability datasets is the key test.
- **Unigram estimation:**
  - generally unclear how to estimate u well enough.
  - using OPT-30B generations as a proxy for training-token frequencies may bias u.

Take-home points

- SLOR's fixed frequency and length corrections are not optimal for modern LMs.
- Learned parameters suggest: larger LMs require less frequency correction (consistent with better rare-word prediction).
- Value for our field: testable assumptions about how LM probabilities map to linguistic judgments (grammatical and more).
- Disentanglingn grammatical and other kinds of acceptability.

Code and details are available in the paper's repository.