To Words and Beyond: Probing Large Language Models for Sentence-Level Psycholinguistic Norms of Memorability and Reading Times

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have recently been shown to produce estimates of psycholinguistic norms, such as valence, arousal, or concreteness, for words and multiword expressions, that correlate with human judgments. These estimates are obtained by prompting an LLM, in zero-shot fashion, with a question similar to those used in human studies. Meanwhile, for other norms such as lexical decision time or age of acquisition, LLMs require supervised fine-tuning to obtain results that align with ground-truth values. In this paper, we extend this approach to the previously unstudied features of sentence memorability and reading times, which involve the relationship between multiple words in a sentence-level context. Our results show that via fine-tuning, models can provide estimates that correlate with human-derived norms and exceed the predictive power of interpretable baseline predictors, demonstrating that LLMs contain useful information about sentence-level features. At the same time, our results show very mixed zero-shot and few-shot performance, providing further evidence that care is needed when using LLM-prompting as a proxy for human cognitive measures.

1 Introduction

How much useful information do Large Language Models (LLMs) contain about human psycholinguistic features? Prior work indicates that LLMs are able to predict word-level features such as concreteness or valence (Trott, 2024), age of acquisition (Sendín et al., 2025), and lexical decision times (Martínez et al., 2025) as well as for multi-word expressions (Martínez et al., 2024), (Brysbaert et al., 2024). One method of testing for the presence of useful psycholinguistically relevant information within LLMs is to simply prompt LLMs by asking a psycholinguistic query directly in zero-shot fashion. However, some evidence points to a lack of **introspection** in LLMs — their responses to prompts are not necessarily consistent with their latent knowledge that can be accessed in other ways, such as by inspecting token log-probabilities rather than directly prompting (Song et al., 2025; Hu and Levy, 2023). Fine-tuning language models based on small amounts of supervised data provides a way to better capitalize upon the rich learned model representations of pre-trained models. For example, Conde et al. (2025b) fine-tuned Llama 3 models to predict English familiarity ratings, achieving Pearson's correlation improvements of up to 0.3 over zero-shot baselines.

The approach of prompting an LLM to provide psycholinguistic norms differs from approaches which directly predict psycholinguistic features based on theoretically motivated, interpretable features. For example, surprisal theory states that human processing difficulty during reading is related to the predictability of words in context (Hale, 2001; Smith and Levy, 2013; Levy, 2008). For the domain of word memorability, Tuckute et al. (2025) used the theoretically motivated predictors of number of meanings and number of synonyms to predict the memorability of words, while Clark et al. (2026) used the distinctiveness of a sentence's Sentence-BERT semantic embedding (Reimers and Gurevych, 2019), as a theoretically motivated, zero-shot, sentence-level predictor of sentence memorability (among other features such as average word memorability and average word frequency). In our view, these methods address different questions, and thereby form a complementary set of approaches that collectively shed light on the functioning of LLMs and their alignment with human cognition.

Probing for psycholinguistic features exists as part of a broader paradigm of using artificial intelligence models to study and simulate human cognition (Frank and Goodman, 2025; Hagendorff et al., 2024; Binz et al., 2025). The use of LLM-prompting in cognitive science has also raised ques-

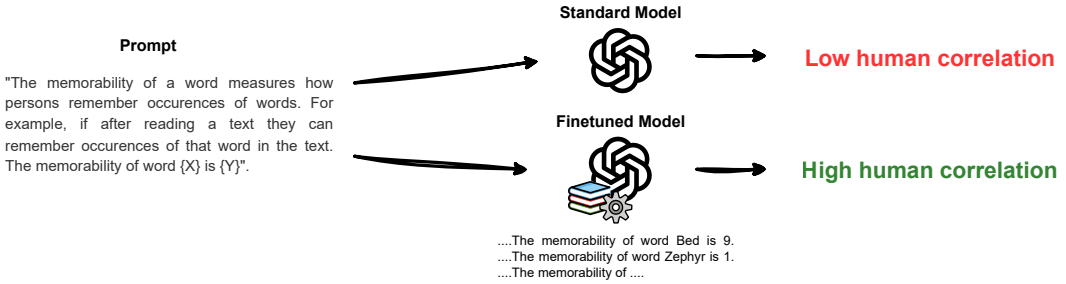

Figure 1: Overview of our approach, contrasting zero-shot prompting with fine-tuning on small supervised datasets for predicting psycholinguistic norms.

tions both about the best methodologies for eliciting predictions, as well as calls for caution regarding using LLM outputs as a substitute for human data (Gao et al., 2025; Dentella et al., 2023). Within this landscape, it is crucial to understand how well LLM outputs actually align with varied psycholinguistic features, whether and how this alignment can be improved, and whether the paradigm of LLM-prompting can be extended to sentence-level, and not just word-level, features.

In this paper, we go beyond existing work studying word-level psycholinguistic features in LLMs, and explore whether LLMs can generate predictions for norms which require processing at the sentence level. Specifically, we systematically compare zero-shot vs. fine-tuned predictions from LLMs to psycholinguistic behavioral norms, across the domains of memorability and reading times.

Our contributions are as follows:

1. We demonstrate that for memorability (both word and sentence-level), zero-shot prompting yields predictions that are uncorrelated with empirical norms. Meanwhile, for reading times, zero-shot prompting yields predictions that do correlate with empirical norms — modestly for self-paced reading and more strongly for eye-tracking data.

2. We demonstrate that, for both memorability and reading times, fine-tuning on a few hundred examples yields predictions that are strongly correlated with empirical norms — exceeding the predictive power of simple, interpretable baselines. This shows that via fine-tuning, the prompting of LLMs for psycholinguistic norms can be extended to sentence-level norms, not just word-level features.

The remainder of the paper is organized as follows. Section 2 discusses the data, models, and procedures used in the evaluation. Section 3 presents the results, which are then discussed in Section 4. We conclude by discussing implications for the intersection of psycholinguistics and NLP, and the limitations and caveats of this approach.

2 Methods

This section describes the data, models, and procedures used to generate and evaluate LLM predictions of memorability and reading times, across zero-shot and fine-tuned settings. A general overview of the procedure is illustrated in Figure 1.

2.1 Data

Here we introduce the datasets which we use in our model evaluations. We focus on two different but important psycholinguistic domains — memorability and reading times — which have not yet been the target of comparison with LLM predictions. In contrast to past work (e.g., Conde et al., 2025a; Trott, 2024), which has considered word-level features such as the Glasgow word norms (Scott et al., 2019) or the Lancaster sensorimotor word norms (Lynott et al., 2020), the domains of sentence memorability and reading times pose a distinct challenge for estimating psycholinguistic norms via LLM prompting, because they are *sentence-level* features. For instance, the memorability of a sentence is a property of the entire sentence, influenced by its compositional meaning (Clark et al., 2026). Likewise, while reading times are known to depend on certain stable properties of a word, such as its length, frequency, or age of acquisition (Smith and Levy, 2013; Brothers and Kuperberg, 2021; Luke and Christianson, 2018; Demberg and Keller, 2008; Kennedy et al., 2013), they also depend on its contextual surprisal, which varies from sentence to sentence. Any model that successfully predicts reading time variation across different instances

of the same wordform will need to take into account the relationship between multiple words in a sentence.

We also include word memorability as a comparison for sentence memorability. For both types of memorability data, norms are collected not by asking human raters for their judgments, but rather by conducting a repeat detection experiment; this makes these norms inherently more scarce and expensive to acquire. Another challenge of memorability data is its dissociation from human subjective judgments, as reported in prior literature. Isola et al. (2014) report that subjective human judgments of the memorability of images actually correlate *negatively* with empirically measured memorability, while Clark et al. (2026) report that subjective memorability judgments for sentences have a weak correlation of 0.24 with empirical sentence memorability. Therefore, for all of the following datasets — word memorability, sentence memorability, and reading times — we have reason to believe that LLM predictions may not align as well with the ground truth compared to results from the prior literature.

2.1.1 Word Memorability

We use the word memorability data of Tuckute et al. (2025)[1], which contains 2109 English words with empirical memorability scores collected via a behavioral experiment with native English speakers. This experimental paradigm builds on a body of work related to measuring the intrinsic memorability of various classes of stimuli, such as images (Isola et al., 2011) and faces (Bainbridge, 2017). A word's memorability is quantified as the average accuracy of responses across participants within a repeat detection (recognition memory) paradigm, yielding values between 0 and 1. In this paradigm, reporting a novel stimulus as familiar and failing to report a repeat stimulus as familiar are the two sources of incorrect responses.

2.1.2 Sentence Memorability

We use the sentence memorability data of Clark et al. (2026)[2], which contains 2500 English sentences with empirical memorability scores collected via a behavioral experiment with native English speakers, following the same experimental paradigm as Tuckute et al. (2025), thereby resulting in scores between 0 and 1. Table 1 shows examples of high- and low-memorability sentences from the dataset.

Sentence	Memorability
Does olive oil work for tanning?	0.98
Scott cried, pursing his pink lips.	0.89
The weather was warm and dry.	0.79
I can't get hold of him.	0.72
We want to make it better.	0.56

Table 1: Example sentences from the dataset of Clark et al. (2026), with human-derived memorability scores.

2.1.3 Self-Paced Reading Times

We use the reading time data from the Natural Stories Corpus (Futrell et al., 2021)[3], which used self-paced reading (Aaronson and Scarborough, 1976; Mitchell and Green, 1978) to gather average reading durations for words in naturalistic stories. In this paradigm, speakers are presented with one word at a time on a screen, and advance to the next word by pressing a key. Reading duration (in milliseconds) is the time between key presses (while the target word is on screen). The dataset consists of 433 sentences comprising 10,256 words.

2.1.4 Eye-Tracking Reading Times

We use the reading time data from the OneStop Corpus, which used eye-tracking to gather reading times for words in a variety of news articles (Berzak et al., 2025)[4]. In this paradigm, speakers read naturalistically while their eye movements are tracked, yielding reading measures with high spatial and temporal resolution. Multiple reading measures are available, including first fixation duration, gaze duration, and total duration. For our main reading measure in this study, we use gaze duration (the sum, in milliseconds, of the durations of all fixations that land on a word during the first pass, before the gaze leaves the word), as this has an intuitive interpretation as the amount of time used to read a word when it is first encountered. In a different dataset (MECO; Siegelman et al., 2022), gaze duration was shown to be largely predictable based on word length, frequency, and surprisal, with R^2 values of 0.6-0.8 (Opedal et al., 2024). The dataset consists of 1,213 sentences comprising 36,120 words.

[1]Experiment 1, available at https://github.com/gretatuckute/memorable_words under MIT license.

[2]Available at https://github.com/thomashikaru/sentence_memorability_share under MIT license.

[3]Available at https://github.com/languageMIT/naturalstories under CC BY-NC-SA 4.0 license.

[4]Available at https://osf.io/2prdq/ under CC BY 4.0 license.

2.2 Model Evaluation

Here we describe in detail our procedure for extracting predictions from models in three settings: zero-shot, few-shot and fine-tuning. In line with the majority of previous work that has relied on models from the GPT-4 family, we used the `GPT-4o-mini-2024-07-28` model (Conde et al., 2025b). Additionally, we test models from the Llama, Gemma, and Qwen families to provide a comparison of model performance. The selected models are Llama-3.1-8b-Instruct, Gemma-3-27b-it, and Qwen3-32b. Due to hardware limitation during fine-tuning, Qwen3-32b and Gemma-3-27b-it were loaded in a 8 bit resolution rather that the full 16 bit resolution

2.2.1 Zero-Shot

For the zero-shot evaluation, we use OpenAI's batch API to submit all requests. For all requests in both zero-shot and fine-tuned model evaluation, the temperature is set to 0; this forces the model to use greedy sampling, where it always selects the token with the highest probability. This generates nearly deterministic (He and Lab, 2025) outputs, which facilitates reproducibility.

We use the following zero-shot prompt for the word-memorability dataset:

```
You are an expert in psycholinguistics. Your
task is to estimate the memorability of English
words. You will give each word a rating from 0
to 1 with two decimal digits. A rating of 1
indicates that the word is maximally memorable,
meaning that people who see the word always
remember having seen it later, and never confuse
 it with a different word (even a similar one).
A rating of 0 indicates that the word is not
memorable, meaning that people who see it forget
 it or may confuse it with another word. Please
limit your answer to a number with two decimal
digits. The word is {word}
```

For each entry in the dataset, `{word}` was replaced by the corresponding word of the dataset.

For sentence memorability, a similar prompt was used, simply replacing references to "words" with references to "sentences".

For both self-paced and eye-tracking reading times, the requested output was a JSON-like structure containing the estimates for each word of the sentence.

```
You are an expert in psycholinguistics. Your
task is to estimate how long, in milliseconds,
an average reader will take to read each word of
 an English sentence. Take into accounts factors
 such as the difficulty of reading the word and
the context of the word within the sentence.
Output a JSON-like data structure containing
word-duration pairs. For example, for the
sentence ''I like cats'' and the reading time
estimates 100ms, 200ms, 200ms, the output must
be {'I':200, 'like': 200, 'cats': 200}. Include
duplicate keys if there are duplicate words even
 if the result is not strictly valid JSON. The
order of keys should be identical to the order
of words in the sentence. Do not add any other
information. The sentence is: {sentence}
```

For sentence-level data, we align the model's output with the input while accounting for the possibility of missing or inserted words using a dynamic programming approach, implemented via the `jiwer` Python library.

2.2.2 Few-Shot

Few-shot evaluation proceeds identically to zero-shot evaluation, with the difference that three supervised examples are provided as part of the prompt.

2.2.3 Fine-Tuning

We performed Supervised Fine-Tuning (SFT) in which we provided the model with both the prompt and the expected output based on the ground truth. For all fine-tuning processes, 25% of the dataset was used for training and 75% for evaluation. For the word and sentence memorability datasets, the expected output was the raw estimate (a single number), whereas for the reading time datasets, the expected output was a JSON-like structure (as described above). During the training process, the model weights are adjusted to adapt the estimate to the expected value.

To train GPT-4o, we used the default fine-tuning API settings, with a fixed learning rate of 1.8. Once the training is completed, the fine-tuned version is stored and can be queried later through the batch API. The other models were fine-tuned locally using LoRA (Low-Rank Adaptation), a technique to adapt large models to specific tasks by training only a small number of new parameters (low-rank adapters) instead of the entire model. The details of the fine-tuning process are presented in Table 2.

The prompts used for fine-tuning were the same ones used for the zero-shot evaluation. Once a fine-tuning job is created, the training process begins until the performance converges. For all evaluations, the final model checkpoint (after 3 epochs of fine-tuning) was used.

Once the tuning process was completed, we evaluated its performance with the test dataset using the same methodology as in the zero-shot evaluation.

Dataset	# Examples	Epochs	Batch Size
Word mem.	527	3	1
Sentence mem.	625		
Self-paced	108		
Eye-tracking	303		

Table 2: Fine-tuning parameters used for each dataset. Note that for the reading time datasets, "# Examples" denotes the number of sentences.

2.3 Correlation Analysis

For each set of model predictions, we evaluate the correlation between human behavioral measures and the model-generated values. We quantify the correlation using the Pearson correlation coefficient and R^2 values.

2.4 Baselines

In order to compare the performance of model predictions for the norms of memorability and reading times, we establish baselines using linear regressions trained on a) interpretable features from the literature, argued to strongly correlate with each psycholinguistic feature of interest, and b) word and sentence embeddings. Using 100 random 0.75/0.25 train/test splits of the data, we fit regressions using each baseline predictor on the training data, and then evaluate the R^2 value on the held-out test data. This procedure yields a distribution over R^2 values, from which we report the mean.

For word memorability, we use the number of meanings, number of synonyms, and word frequency, using values provided in the dataset of Tuckute et al. (2025). In that study, the values for number of synonyms and number of meanings were collected via a human norming study, and word frequency was computed using the Subtlex corpus (Brysbaert and New, 2009). Additionally, we include a regression based on all three scalar predictors, as well as the 300-dimensional GloVe embedding for each word. For sentence memorability, we use Sentence-BERT embedding distinctiveness, average word memorability, and average word frequency, using values provided in the dataset of Clark et al. (2026), since these were identified as predictors of sentence memorability. In that study, Sentence-BERT distinctiveness was computed as the mean cosine distance of a sentence's Sentence-BERT representation to all other sentences in a large and diverse sample of sentences, while average word memorability was estimated using the human norms of Tuckute et al. (2025) and word frequency was taken from the Subtlex corpus and averaged across words in a sentence. Additionally, we include a regression based on all three scalar predictors, as well as the 384-dimensional Sentence-BERT embedding for each sentence.

For the reading time norms, we establish baselines using the interpretable features of word length, frequency, and contextual surprisal, following past work (Demberg and Keller, 2008; Brothers and Kuperberg, 2021; Smith and Levy, 2013; Wilcox et al., 2023; Opedal et al., 2024, inter alia). Word length was defined simply as the number of characters in the orthographic representation of the word. Frequency was computed using the `wordfreq` package (Speer, 2022) for Python. Surprisal was computed using the `GPT-2` language model (Radford et al., 2019) and the `wordsprobability` Python package (Pimentel and Meister, 2024). Additionally, we include linear regressions based on all three scalar predictors, as well as 768-dimensional contextual BERT embeddings and 300-dimensional non-contextual GloVe embeddings. For words which decompose into multiple BERT tokens, only the first token is used.

3 Results

This section presents and briefly discusses the results for each of the datasets[5]. The Llama model did not succeed at responding consistently to the standard prompts, instead sometimes writing code or producing other non-usable output; Llama results are thus omitted from the following analysis, but we note that our approach therefore does not work for all LLMs.

3.1 Word Memorability

Zero-shot and few-shot model predictions are essentially uncorrelated with empirical word memorability scores, while fine-tuned models achieve R^2 values of $0.53 \sim 0.59$ (Figure 2). This exceeds the mean R^2 of 0.28 using the combined baseline predictors of number of meanings, number of synonyms, and frequency.

3.2 Sentence Memorability

Zero-shot and few-shot model predictions are essentially uncorrelated with empirical sentence

[5]The results are available at: link removed for double blind review.

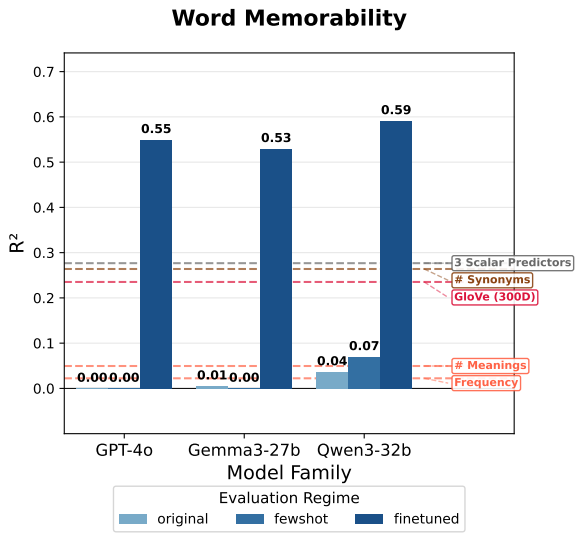

Figure 2: The correlation between model predictions and ground truth norms for word memorability, across 3 model families and in 3 evaluation regimes. For comparison, the mean correlation with the predictions of interpretable baseline predictors is included.

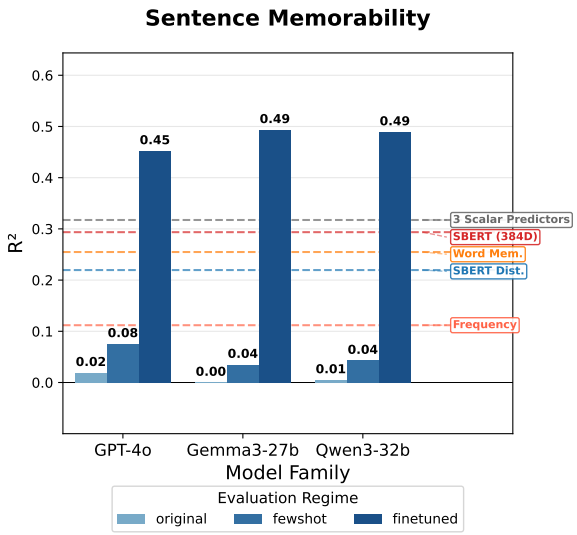

Figure 3: The correlation between model predictions and ground truth norms for sentence memorability, across 3 model families and in 3 evaluation regimes. For comparison, the mean correlation with the predictions of interpretable baseline predictors is included.

memorability scores, while fine-tuned models achieve R^2 values of $0.45 \sim 0.49$ (Figure 3). This exceeds the mean R^2 of 0.32 using the combined baseline predictors of Sentence-BERT distinctiveness, average word-level memorability, and average word frequency.

3.3 Reading Times

For the Natural Stories Corpus, zero-shot model predictions have low correlations of $0.02 \sim 0.05$ with empirical sentence memorability scores, while fine-tuned models achieve R^2 values of $0.15 \sim 0.21$ (Figure 4). This is considerably lower than the R^2 values attained by fine-tuned models on the memorability data, but higher than the correlation of 0.08 using the combined scalar baseline predictors, and on par with the R^2 value achieved by predicting values using a linear model trained on words' 768-dimensional contextual BERT embeddings.

Interestingly, we observe a qualitatively different pattern for eye-tracking reading times from the OneStop Corpus (Figure 5). The zero-shot model attains R^2 values of 0.27, the few-shot model attains R^2 values of 0.35, while the fine-tuned model attains R^2 values in a wide range from $0.08 \sim 0.57$. For GPT, the fine-tuned model exceeds the predictive power of the baseline predictors, including the combined three scalar predictors of length, frequency, and surprisal, as well as the predictive power of contextual and non-contextual word embeddings. For Gemma and Qwen models, predictions from fine-tuning underperform zero-shot and few-shot predictions, indicating a failure of the fine-tuning process to elicit humanlike reading-time predictions. The moderate predictive power of zero-shot and few-shot predictions, however, point to some latent knowledge that correlates with reading times.

Consistent with known differences between Self-Paced Reading and Eye-Tracking, we find considerable differences in the degree to which the baseline predictors of surprisal, frequency, and length are able to predict reading time measures across the two datasets of Natural Stories and OneStop. We note that our baseline results for OneStop are consistent with the work of Opedal et al. (2024), who found that length and frequency explain a greater deal of variance in reading times, compared to contextual surprisal.

4 Discussion

This section discusses the key insights from the experimental results.

4.1 LLMs can predict sentence-level norms via fine-tuning

The task of generating reading times for each word in a sentence is considerably more complex than estimating, for example, the valence of a single

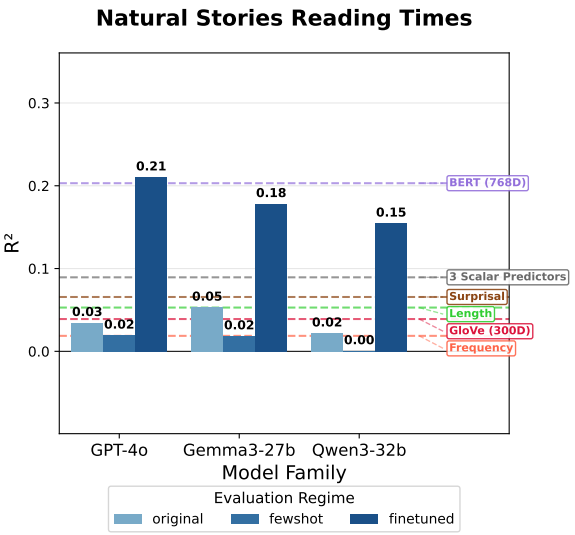

Figure 4: The correlation between model predictions and ground truth norms for self-paced reading times (Natural Stories corpus), across 3 model families and in 3 evaluation regimes. For comparison, the mean correlation with the predictions of interpretable baseline predictors is included.

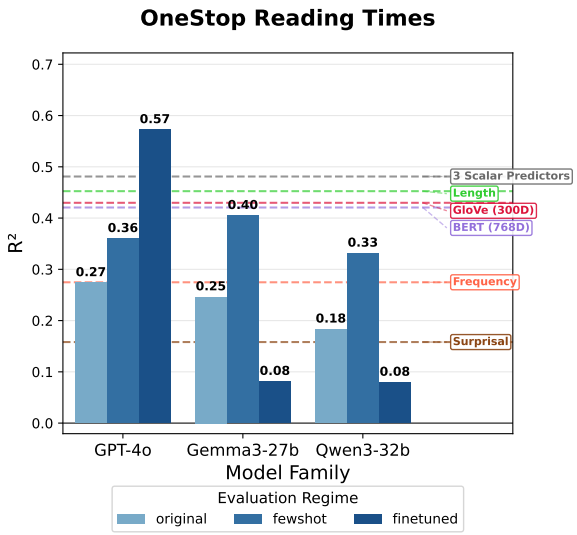

Figure 5: The correlation between model predictions and ground truth norms for eye-tracking reading times (OneStop corpus), across 3 model families and in 3 evaluation regimes. For comparison, the mean correlation with the predictions of interpretable baseline predictors is included.

word. The same applies to estimating the memorability of a sentence. This is because both of these norms involve considering the ways in which multiple words in a sentence interact with each other. For instance, one of the most memorable sentences in the study of Clark et al. (2026) is *Does olive oil work for tanning?*; the memorability of this sentence cannot simply be reduced to the sum of word-level features, but involves the compositional meaning of the words in relation to each other.

Despite this challenge, LLMs are capable of producing estimates that correlate with human behavior when fine-tuned on a small set of supervised data. For both memorability and reading times, fine-tuning achieves a correlation with the ground truth that exceeds that of simple, theoretically motivated baseline predictors. This is consistent with previous results for word-level norms (Conde et al., 2025b), showing that the same general paradigm can be extended to the sentence level. This means that LLMs can potentially be used to predict those norms which are more difficult to collect than word-level norms which rely only on subjective judgments. We also note that the improvement is not attributable simply to high performance on words seen during fine-tuning — for the memorability experiments, the evaluation was conducted on entirely held-out stimuli. For the reading time experiments, all test sentences were held-out, though some individual wordforms were seen in both fine-tuning and testing; model prediction accuracy was high even for words not been seen during fine-tuning.

Crucially, for most datasets, the R^2 value attained by the fine-tuned prompting method exceeds that of both theoretically motivated baseline predictors, and rich semantic embeddings. This demonstrates that the prompting method not only can capture the variance explained by these interpretable predictors, but captures additional variance as well.

On a theoretical level, this may suggest that the rich distributed representations of words produced by language models encode sufficient information about a diverse range of psycholinguistic behavioral measures, such that fine-tuning on a small dataset enables reliable decoding of these features from the existing model representations (as opposed to learning the relationship from scratch).

On a practical level, this method offers a way for psycholinguists to develop predictive models that can estimate psycholinguistic norms for held-out data. While this can introduce challenges in interpretability, the approach may be useful in settings such as the development of experimental materials. One possible use case is ensuring matched control stimuli, as has previously been done with the A-Maze variant (Boyce et al., 2020) of the Maze reading task (Forster et al., 2009).

4.2 Zero-shot and few-shot performance vary widely across domains

In our results, we observed that for word and sentence memorability, zero-shot model predictions were essentially completely uncorrelated from human behavior. We speculate that the features which make words and sentence memorable may not be obvious or transparent, consistent with findings from the memorability literature showing that humans themselves are poor judges of memorability in domains such as images (Isola et al., 2014). Meanwhile, the features that correlate strongly with reading times are relatively simple and interpretable in comparison — basic properties of words such as length and frequency, as well as contextual features such as their predictability given preceding sentential context (Hale, 2001; Smith and Levy, 2013; Wilcox et al., 2023, 2020), which may be particularly well-aligned with the next-word prediction training objective of LLMs. Thus we speculate that zero-shot prompting of reading times from LLMs is more effective in extracting predictions that align with human behavior, compared to memorability.

4.3 Differences between self-paced reading and eye-tracking

We also observe considerable differences in the performance of models when comparing self-paced reading and eye-tracking. Eye-tracking remains the gold standard for reading time data, because participants are able to read naturalistically and because the paradigm allows both high spatial and temporal resolution. In self-paced reading, by comparison, there are well-established "spillover" effects (Rayner, 1998; Smith and Levy, 2013) stemming from the rapid pressing of keys and the latency between reading and motor actions. Additionally, the paradigm is somewhat divorced from reading in the wild. Fine-tuning still yields an improvement in predicting self-paced reading times, compared to zero-shot prompting, but the predictive power may be limited by the noise ceiling of the human data.

Figure 6 shows the R^2 values between model-predicted reading times and the ground truth values as a function of word position within sentence, for the NaturalStories and OneStop corpora, using the GPT model. For eye-tracking data (OneStop), the R^2 between model predictions and the ground truth remains above close to 0.6 even for word positions far into a sentence. Meanwhile, for the noisier self-paced reading data, R^2 values drop close to 0 for positions beyond approximately 25 words into a sentence. We note that the fine-tuning data naturally contains fewer examples of late word positions than early word positions. Thus, a single prompt can generate model predictions that correlate with human eye-tracking RTs across a long timescale, but for noisier self-paced reading RTs, model predictions become noisier with word position.

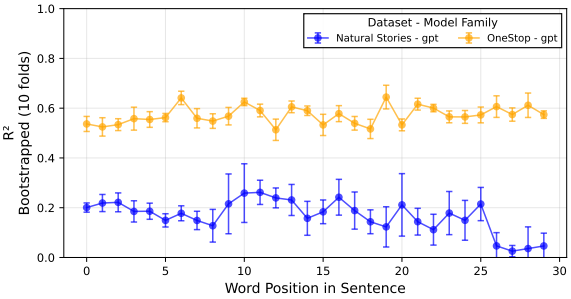

Figure 6: Correlation of fine-tuned model prediction to ground truth for different word positions in a sentence.

5 Conclusion

In conclusion, this work presents several novel findings. First, this work demonstrates that prompting LLMs for psycholinguistic norms can be extended to sentence-level norms — the memorability of an entire sentence, or the reading time of a word in context — not just word-level features, by using a simple and straightforward supervised fine-tuning strategy. We also showed that the reliability of zero-shot LLM predictions for psycholinguistic norms varies considerably by domain, with the predictions for both word and sentence memorability being uncorrelated with empirical norms. Across domains, fine-tuning on a few hundred examples is generally able to align the model predictions with empirical values, even beating strong, theoretically motivated baselines.

While the applications of this method in psycholinguistics are promising, our results also suggest that practitioners must be wary of trusting the zero-shot predictions of LLMs on psycholinguistic norming tasks (Hu and Levy, 2023), and careful validation against human data is recommended (Conde et al., 2025b). Future work may target a comprehensive application of this method to other language model families and additional human languages, in order to evaluate the generalizability of these results.

Limitations

In this section, we acknowledge several limitations of our study.

A limitation of the present study is the exclusive focus on English. It remains an open question whether prompting works as well in lower-resource languages or languages with very different linguistic properties, such as agglutinative languages or those with non-alphabetic writing systems. This limitation also reflects the broader WEIRD bias (Western, Educated, Industrialized, Rich, Democratic) highlighted in cognitive science and NLP, where research has tended to focus on a narrow set of languages and populations that are not representative of global diversity.

Finally, we echo existing calls for caution (Gao et al., 2025; Dentella et al., 2023) regarding the use of LLM outputs as a substitute for, rather than a noisy estimate of, human psycholinguistic norms, especially in zero-shot settings without careful validation against ground-truth values.

References

Doris Aaronson and Hollis S. Scarborough. 1976. Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance*, 2(1):56–70. Place: US Publisher: American Psychological Association.

Wilma A. Bainbridge. 2017. The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(5):706–716. Number: 5 Place: US Publisher: American Psychological Association.

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. OneStop: A 360-Participant English Eye Tracking Dataset with Different Reading Regimes.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, and 21 others. 2025. A foundation model to predict and capture human cognition. *Nature*, pages 1–8. Publisher: Nature Publishing Group.

Veronica Boyce, Richard Futrell, and Roger P. Levy. 2020. Maze Made Easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Trevor Brothers and Gina R. Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Marc Brysbaert, Gonzalo Martínez, and Pedro Reviriego. 2024. Moving beyond word frequency based on tally counting: Ai-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, 57(1):28.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Thomas Hikaru Clark, Greta Tuckute, Bryan Medina, and Evelina Fedorenko. 2026. A distinctive meaning makes a sentence memorable. *Journal of Memory and Language*, 146:104700.

Javier Conde, Miguel González, María Grandury, Gonzalo Martínez, Pedro Reviriego, and Mar Brysbaert. 2025a. Psycholinguistic Word Features: a New Approach for the Evaluation of LLMs Alignment with Humans. *arXiv preprint*. ArXiv:2506.22439 [cs].

Javier Conde, María Grandury, Tairan Fu, Carlos Arriaga, Gonzalo Martínez, Thomas Clark, Sean Trott, Clarence Gerald Green, Pedro Reviriego, and Marc Brysbaert. 2025b. Adding LLMs to the psycholinguistic norming toolbox: A practical guide to getting the most out of human ratings. *arXiv preprint arXiv:2509.14405*.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210. Number: 2.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120. Publisher: Proceedings of the National Academy of Sciences.

Kenneth I. Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41(1):163–171.

Michael C. Frank and Noah D. Goodman. 2025. Cognitive Modeling Using Artificial Intelligence. *Annual Review of Psychology*. Publisher: Annual Reviews.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina. *arXiv preprint*. ArXiv:2410.19599 [econ].

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2024. Machine Psychology. *arXiv preprint*. ArXiv:2303.13988 [cs].

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Horace He and Thinking Machines Lab. 2025. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint*. ArXiv:2305.13264 [cs].

Phillip Isola, null Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What Makes a Photograph Memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482. Number: 7.

Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pages 2429–2437, Red Hook, NY, USA. Curran Associates Inc.

Alan Kennedy, Joël Pynte, Wayne S. Murray, and Shirley-Anne Paul. 2013. Frequency and predictability effects in the Dundee Corpus: an eye movement analysis. *Quarterly Journal of Experimental Psychology (2006)*, 66(3):601–618.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177. Place: Netherlands Publisher: Elsevier Science.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3):1271–1291.

Gonzalo Martínez, Juan Diego Molero, Sandra González, Javier Conde, Marc Brysbaert, and Pedro Reviriego. 2024. Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, 57(1):5.

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.

D. C. Mitchell and D. W. Green. 1978. The Effects of Context and Content on Immediate Processing in Reading. *Quarterly Journal of Experimental Psychology*, 30(4):609–636. Publisher: SAGE Publications.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the Role of Context in Reading Time Prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3058, Miami, Florida, USA. Association for Computational Linguistics.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422. Place: US Publisher: American Psychological Association.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint*. ArXiv:1908.10084 [cs].

Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. 2019. The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270.

Eneko Sendín, Javier Conde, Pedro Reviriego, Juan Haro, Pilar Ferré, José A Hinojosa, and Marc Brysbaert. 2025. Combining the power of large language models with finetuning based on strategically collected human ratings: A case study about age-of-acquisition estimates of spanish words. *Psicológica*.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, and 12 others. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

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

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language Models Fail to Introspect About Their Knowledge of Language. *arXiv preprint*. ArXiv:2503.07513 [cs].

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Sean Trott. 2024. Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Greta Tuckute, Kyle Mahowald, Phillip Isola, Aude Oliva, Edward Gibson, and Evelina Fedorenko. 2025. Intrinsically memorable words have unique associations with their meanings. *Journal of Experimental Psychology. General*.

Ethan G. Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-TimeComprehension Behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 42(0).

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470. Place: Cambridge, MA Publisher: MIT Press.

A Use of Artifacts and Models

We utilize the `GPT-4o-mini-2024-07-28` language model via the OpenAI API.

We also utilize existing datasets from Tuckute et al. (2025) (MIT License), Clark et al. (2026) (MIT License), Futrell et al. (2021) (CC BY-NC-SA 4.0 license.), and Berzak et al. (2025) (CC BY 4.0), which are publicly available via GitHub or OSF. We use this data purely for research purposes.

We acknowledge the use of the ChatGPT and Cursor AI assistants for help with code development.

B Risks

We do not foresee any risks to this research, as it is focused on the evaluation of existing models using a new approach.