Clozing the Gap: Exploring Why Language Model Surprisal Outperforms Cloze Surprisal

Sathvik Nair*
University of Maryland
sathvik@umd.edu

Byung-Doh Oh*
Nanyang Technological University
byungdoh.oh@ntu.edu.sg

Abstract

How predictable a word is can be generally quantified in two ways: using human responses to the cloze task or using probabilities from language models (LMs). When used as predictors of processing effort, LM probabilities outperform probabilities derived from cloze data. However, it is important to establish that LM probabilities do so for the right reasons, since different predictors can lead to different scientific conclusions about the role of prediction in language comprehension. We present evidence for three hypotheses about the apparent advantage of LM probabilities: not suffering from low resolution, distinguishing semantically similar words, and accurately assigning probabilities to low-frequency words. These results call for efforts to improve the resolution of cloze studies, coupled with experiments on whether human-like prediction is also as sensitive to the fine-grained distinctions made by LM probabilities.

1 Introduction

Prediction is a key principle of language comprehension: Words that are more expected given a context are easier to process compared to unexpected words (Ehrlich and Rayner, 1981; Kutas and Hillyard, 1984; Smith and Levy, 2013). Studying this process requires a measure of how predictable a word is in its context to typical human readers (i.e. its *predictability*). The effects of predictability on processing effort have been widely studied and characterized quantitatively (Smith and Levy, 2013; Brothers and Kuperberg, 2021; Szewczyk and Federmeier, 2022; Shain et al., 2024).

The traditional method for estimating such predictability is the cloze task, where participants provide a completion of the next word given an incomplete sentence (Taylor, 1953): Words that are produced more frequently are considered to be more

*These authors contributed equally to this work.

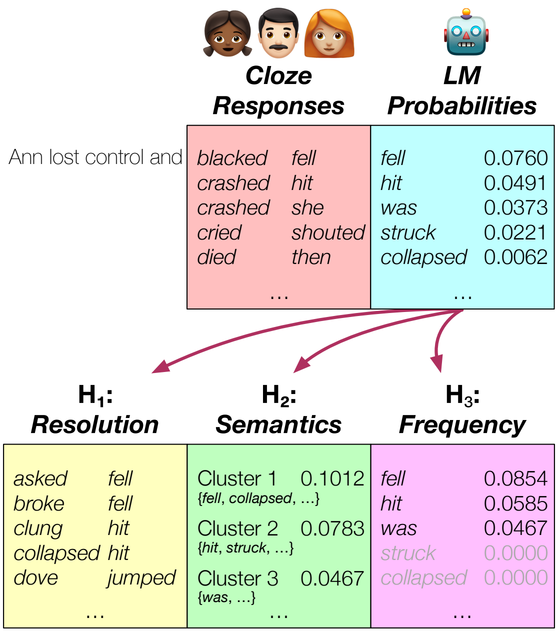

Figure 1: How predictable a word is in its context has traditionally been quantified using the cloze task (red), which is increasingly being replaced with LM probabilities in recent years (blue). After establishing that LM probabilities generally provide a better predictor of reading times than cloze responses (Experiment 1), we test three hypotheses about why (yellow, green, magenta) by conducting a hypothesis-driven manipulation of LM probabilities (Experiment 2).

predictable. Predictability can also be estimated with language models (LMs): Modern LMs are based on neural networks (Radford et al., 2019; Biderman et al., 2023; Grattafiori et al., 2024) that are trained on a cloze-like task of predicting the next word over large-scale examples of human language use, thus providing accurate estimates of words' conditional probabilities.

Recent experiments show that LM surprisal (negative log probability) generally seems to be a superior predictor of comprehension difficulty compared to cloze surprisal (Hofmann et al., 2022; Michaelov et al., 2023; Shain et al., 2024, though

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33193–33210
July 2-7, 2026 ©2026 Association for Computational Linguistics

see de Varda et al., 2024). This stronger fit between LM surprisal and human subject data has led some to claim that LMs should replace the cloze task entirely (Michaelov et al., 2023; Shain et al., 2024).

However, this claim based on fit to data ignores systematic differences between LM and cloze-based estimates of predictability. LMs reflect knowledge of billions of training examples and perfect access to the input text, which enable them to make predictions that average readers cannot (Oh and Schuler, 2023; Oh and Linzen, 2025). Additionally, LM-based estimates of predictability are based *purely* on probabilistic inference and are not explicitly linked to any notion of structure building (Slaats and Martin, 2025; Nair and Phillips, 2026), which limits their ability to model the full range of language processing behavior (Stanojević et al., 2023; Staub, 2025). On the other hand, cloze data directly come from human participants, but the cloze task is an offline task that elicits a single response from each participant, which may evoke different cognitive processes from those that underlie the graded predictions that are made rapidly (Staub et al., 2015; Brothers et al., 2023), and thus do not fully represent human probabilistic intuitions (Smith and Levy, 2011). This difference appears to underlie the qualitatively different next-word predictions—and therefore predictability estimates—from humans and LMs (Jacobs et al., 2024; Shlegeris et al., 2024).

Crucially, different methods of estimating predictability can lead to different scientific conclusions. For example, the conclusion that the effect of predictability is better thought of as a *facilitation* at predictable words (instead of a *cost* at unpredictable words; Brothers and Kuperberg, 2021) changes when GPT2 probabilities are used instead of cloze probabilities to model exactly the same reading time data (Shain et al., 2024). Even within the same class of LMs, how much the slowdown observed at low-frequency words can be reduced to the effect of predictability changes depending on the specific model used to calculate predictability estimates (Oh and Schuler, 2025). Therefore, we argue that it is important to establish where the relative contribution of cloze and LM surprisal in modeling human data comes from, instead of advocating for the use of LM surprisal purely on the basis of stronger fit to data.

To this end, we first re-establish LM surprisal as a superior predictor of reading times compared to cloze surprisal, while additionally exploring how cloze probabilities should be smoothed and transformed. We then adopt a novel approach of intervening on the LM's probabilities to test three hypotheses that may account for the stronger fit of LM surprisal (Figure 1). We find support for all three of our hypotheses: LMs appear to yield stronger predictors of RTs than cloze responses due to their ability to assign highly fine-grained probabilities, distinguish between semantically similar words, as well as between low-frequency words. This calls for efforts to improve the resolution of cloze studies and for experiments on whether humans' predictions are as sensitive to the fine-grained distinctions made by LM probabilities.[1]

2 Experiment 1: Evaluation of Cloze and LM Surprisal as Predictors of RTs

The first experiment compares the fit of cloze probabilities and LM probabilities to by-word reading times (RTs). Using multiple English reading time datasets for which cloze responses are available, we fit regression models that contain either cloze or LM probability as a predictor. We then compare the fit of these models against a regression model that contains both predictors to evaluate whether one predictor subsumes the effect of the other. In order to systematically compare the two measures across datasets, we place an increased emphasis on experimenting with different methods for smoothing and transforming cloze probabilities.

2.1 Response Data: By-Word Reading Times

We analyze four English reading time datasets with available cloze response data. The stimuli text, cloze responses, and reading times of each dataset are as follows:

BK21 SPR (Brothers and Kuperberg, 2021). The BK21 self-paced reading (SPR) dataset contains 216 triplets of sentences that manipulate the preceding context to make the same target word highly predictable (high-cloze), moderately predictable (moderate-cloze), and rarely predictable (low-cloze). The items were validated through a cloze norming study that resulted in about 90 responses per sentence: For the high-, moderate-, and low-cloze conditions, the subjects predicted the correct target word around 91%, 20%, and 1% of the

[1]The code used in this work is available at `https://github.com/sathvikn/cloze-surprisal`.

time respectively. We present the same example triplet from Brothers and Kuperberg (2021):

- High-cloze: Her vision is terrible and she has to wear **glasses** in class.
- Moderate-cloze: She looks very different when she has to wear **glasses** in class.
- Low-cloze: Her mother was adamant that she has to wear **glasses** in class.

The proportion of times the target word (i.e. **glasses**) was produced in the cloze norming study (cf. the full set of responses) is made publicly available. This dataset also contains SPR times collected from 216 subjects who did not take part in the cloze norming study, summed over the target word and two succeeding words (i.e. **glasses** in class).

Provo ET (Luke and Christianson, 2018)**.** The Provo eye-tracking (ET) dataset contains 55 short English paragraphs that are about 50 words long, extracted from news articles, science magazines, and works of fiction (total 2,746 words). A total of 478 subjects provided cloze responses to each word in the corpus, resulting in about 40 responses for every word. A separate set of 84 subjects who did not participate in the cloze norming study provided eye-tracking-while-reading data for these paragraphs. From the raw eye fixation data, we derive and analyze two by-word measures; first-pass (FP) duration, which is the time taken between entering a word region from the left (in the case of English) and exiting it to either the left or right, and go-past (GP) duration, which is the time taken between entering a word region from the left and exiting it to the right, including all regressive fixations.

UCL SPR/ET (Frank et al., 2013; de Varda et al., 2024)**.** The UCL SPR and ET datasets contain 361 isolated sentences that were extracted from online novels written by aspiring authors. The sentences were chosen such that they consist fully of frequent English words and could readily be interpreted without the surrounding context or any extra-linguistic knowledge. All 361 sentences were used to collect SPR times from 117 subjects, and a subset of 205 sentences that fit within a single line was used to collect eye-tracking data from 48 subjects (Frank et al., 2013). In a separate study by de Varda et al. (2024), around 80 cloze responses were collected for each word in the smaller subset of 205 sentences. We analyze the SPR times, FP durations, and GP durations provided as part of these datasets.

Data filtering. The reading times in each dataset were filtered in standard ways prior to analysis: Reading times of the first and last word of each sentence (all datasets) and each line (ET datasets) were removed to avoid wrap-up effects. SPR times and GP durations that exceed 3000 ms and FP durations that exceed 2000 ms were also removed. Finally, for the UCL dataset that also provides sentence-level comprehension data, reading times from trials with incorrect responses were removed.

2.2 Predictors: Cloze and LM Probabilities

The main predictors of interest are predictability estimates derived from cloze responses and those calculated from an LM:

Cloze probability smoothing and transform. Partly due to the limitations in the number of cloze responses (ranging from around ~40 to ~90 for our RT datasets), the actual next word is often not produced, leaving their probabilities to be zero. This is problematic for the log transform required for surprisal calculation, and therefore the cloze probabilities need to be smoothed by allocating a small probability to zero-count words.

Additionally, the functional form between cloze probability and RTs has not been widely established in the literature. The linearity of cloze probability on RTs across the high/moderate/low-cloze conditions is reported in Brothers and Kuperberg (2021), but it remains to be seen whether this generalizes to other broad-coverage datasets (cf. the linearity of LM surprisal on RT observed across multiple datasets and languages; Wilcox et al., 2023; Xu et al., 2023; Shain et al., 2024).

To establish a strong predictor based on cloze responses, we first experiment with different smoothing factors and different functional forms to RTs. More specifically, we adopt a form of add-one smoothing as follows:

$$\mathsf{P}(w_t) = \frac{C_{w_t} + 1}{\sum_{w_i} C_{w_i} + V}, \tag{1}$$

where C_{w_i} is the number of times the word w_i was attested as cloze responses. In order to control the probability mass that is assigned to unattested cloze completions, we manipulate the smoothing factor $V \in \{50, 100, 200, 500, 1000, 2000\}$. We also evaluate different functional forms between these cloze probabilities and RTs by transforming them prior to including them in linear mixed-effects regression (Xu et al., 2023). The functional forms we experiment with are raw probability $\mathsf{P}(w_t)$, raw surprisal $\mathsf{S}(w_t) = -\log_2 \mathsf{P}(w_t)$, and various power transforms on surprisal, namely $\mathsf{S}(w_t)^{\frac{1}{2}}$, $\mathsf{S}(w_t)^{\frac{3}{4}}$, $\mathsf{S}(w_t)^{\frac{4}{3}}$, and $\mathsf{S}(w_t)^2$.

LM surprisal calculation. The surprisal of each word was calculated from LM-based predictability estimates. For our LM, we used GPT2 (small; Radford et al., 2019), based on its widespread use for next-word probabilities in psycholinguistic modeling studies (Huang et al., 2024; Giulianelli et al., 2026; Kuribayashi et al., 2025). Each sentence of the BK21 and UCL datasets and each paragraph of the Provo dataset was tokenized and provided to GPT2 to calculate conditional probabilities. The probability of the directly following whitespace was included as part of the word probability (Oh and Schuler, 2024; Pimentel and Meister, 2024) to ensure a proper probability distribution over words.[2]

2.3 Modeling: LME Regression

The ability of cloze probability and LM surprisal to predict reading times is evaluated by their contribution to the log likelihood of a linear-mixed effects (LME; Bates et al., 2015) regression model. For each of the six measures (BK21 SPR, Provo FP/GP, UCL SPR/FP/GP), we first specify a baseline LME model that includes a set of baseline predictors. These baseline predictors are word length in characters, position of the word within each sentence, unigram surprisal as a measure of frequency (all measures), and whether or not the preceding word was fixated (FP/GP measures). Unigram surprisal is calculated using the KenLM toolkit (Heafield et al., 2013) with parameters estimated on ~6.5 billion words of the OpenWebText Corpus (Gokaslan and Cohen, 2019).

To first determine the smoothing and transform of cloze probability that achieves the strongest fit to RTs, we evaluate how much each variant of cloze probability improves the in-sample log likelihood over the baseline LME model. To this end, we fit a set of 36 LME models (6 smoothing factors × 6 transforms) that include the baseline predictors and a variant of cloze probability to about 50% of the observations for each of the six measures.

After choosing the best-fitting variant of cloze probability, the fit of cloze probability is directly compared against that of GPT2 surprisal by comparing nested LME models. Three LME models are fit to each measure; on top of the baseline model, one additionally containing cloze probability, one additionally containing GPT2 surprisal, and one additionally containing both cloze probability and GPT2 surprisal.[3] Their fit to RTs was evaluated using 10-fold cross-validation; after splitting each measure into 10 folds,[4] held-out log likelihood is calculated on the fold that was not used for model fitting. The two comparisons of interest are between the model containing both predictors against that containing only cloze probability and that containing only GPT2 surprisal. Statistical significance is determined by a paired permutation test over the two sets of 10 per-observation log likelihood values. All LME models include a by-subject random intercept; the datasets did not consistently support a richer random-effects structure.

2.4 Results

The results from cloze probabilities with different transforms in Table 1 show that transforming probabilities into surprisal notably improves fit to RT; otherwise, the smoothing factor and the power transforms of surprisal did not have a very large effect. Contrary to the findings of Brothers and Kuperberg (2021), when other datasets are taken into consideration, the effect of cloze probability on reading time does not seem to be linear. Throughout the remainder of this paper, we apply $\mathsf{S}(w_t)^2$ with a smoothing factor of $V = 200$ to all cloze probabilities, as this setup achieved the best fit to the six measures.[5]

[2] Without this correction, for example, if `carpet` is tokenized into subword tokens `car` and `pet`, it always holds that $\mathsf{P}(\texttt{carpet}) \leq \mathsf{P}(\texttt{car})$.

[3] While it is common to include surprisal from preceding words as predictors of the current word's RT to account for 'spillover' effects in reading (Vasishth, 2006), we do not include such spillover predictors for two reasons. The first is that for the BK21 dataset, cloze response data is only available at the target word and not for any other word of the sentence, making it impossible to define spillover predictors for cloze probability. More fundamentally, including spillover predictors may introduce a confound for answering our research question; if an LME model containing multiple versions of cloze probability is compared against that containing multiple versions of LM surprisal, it remains unclear whether the observed modeling benefit is attributable to the main predictor or spillover predictors.

[4] The partitioning is conducted by cycling through each subject-by-sentence combination and assigning observations from that combination to a different partition. See Appendix A for the number of observations in each fold and their total for each RT measure.

[5] Prior work has also identified superlinear effects of LM surprisal (Meister et al., 2021; Xu et al., 2023; Hoover et al.,

V	$\mathsf{P}(w_t)$	$\mathsf{S}(w_t)$	$\mathsf{S}(w_t)^{\frac{1}{2}}$	$\mathsf{S}(w_t)^{\frac{3}{4}}$	$\mathsf{S}(w_t)^{\frac{4}{3}}$	$\mathsf{S}(w_t)^2$
50	91.7	149.2	141.2	145.9	152.1	**153.1**
100	92.1	149.7	143.3	146.8	152.1	**153.6**
200	92.2	150.0	144.7	147.6	152.0	<u>**153.8**</u>
500	92.4	150.1	146.1	148.3	152.0	**153.7**
1000	92.4	150.2	146.9	148.6	151.9	**153.7**
2000	92.6	150.2	147.3	149.0	151.6	**153.4**

Table 1: The increase in regression model log likelihood due to including cloze probability calculated and aggregated over ~50% of the observations, as a function of different smoothing factors (V) and functional form to RT.

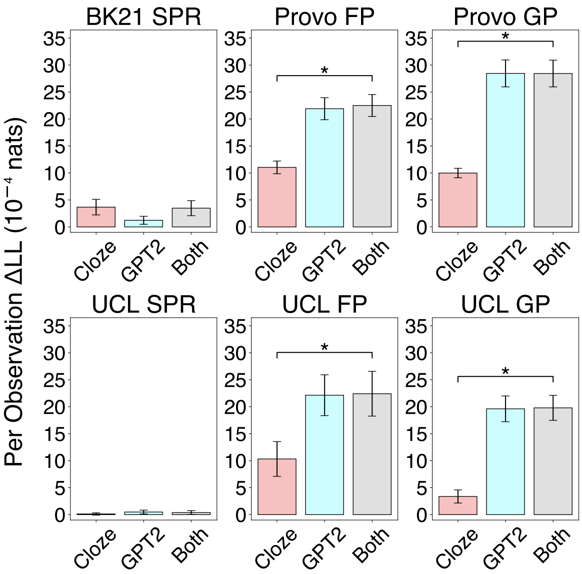

Figure 2: Increase in per-observation log likelihood over the baseline regression models due to including cloze surprisal, GPT2 surprisal, and both predictors, averaged over the 10 folds used in cross-validation. Error bars denote one standard error of the mean (SEM) across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT2 vs. Both), differences that achieve significance at the 0.05 level by a paired permutation test under a 12-way Bonferroni correction (two comparisons on six measures) are marked with an asterisk.

Figure 2 presents the main comparison between cloze and GPT2 surprisal. On four out of six measures, GPT2 surprisal predicts RTs better than cloze surprisal, but not vice-versa, indicating that GPT2 surprisal subsumes the effect of cloze surprisal. This corroborates earlier findings that show LM surprisal is a superior predictor of real-time comprehension data than cloze surprisal (Hofmann et al., 2022; Michaelov et al., 2023; Shain et al., 2024). The two SPR measures that do not show a significant difference between the two predictors were not very well predicted by either cloze or GPT2 surprisal, which is possibly due to task-based differences between SPR and eyetracking.

3 Experiment 2: Why Does LM Surprisal Predict RTs Better?

Having established that GPT2 surprisal predicts RTs over and above cloze surprisal, the second experiment aims to identify where the modeling benefit of GPT2 surprisal comes from. To this end, we manipulate GPT2 probabilities to test three hypotheses that may account for the difference between cloze responses and LM probabilities (Figure 1). The fit of manipulated GPT2 surprisal is compared against that of cloze surprisal following the same modeling setup as Experiment 1. To the extent that each hypothesis is supported, we expect the fit to RT to decrease compared to what is reported in Figure 2.

3.1 Hypotheses and Manipulations

Hypothesis 1 ($\mathbf{H}_1$): Resolution One practical limitation of conducting human studies to collect cloze responses is its cost: It can be prohibitively expensive to collect an ample number of responses over a text corpus. Due to this limitation, cloze probabilities suffer from poor resolution, as they are based on counts from typically fewer than 100 responses, even if their underlying contextual expectations are higher-resolution. In contrast, LMs can estimate probabilities for any arbitrary continuation using their vector representations, and therefore offer probabilities with much higher resolution. To test the hypothesis that the difference in resolution accounts for the difference in the fit

2023). It could be possible that a superlinear effect of cloze surprisal could be associated with production-related pressures, which have been proposed as an explanation for Meister et al.'s (2021) results. However, the small differences between the linear and quadratic transformations of cloze surprisal are not enough to draw conclusions about the functional form of the relationship between surprisal and reading time.

to RTs, we *match* the resolution between cloze and GPT2 surprisal by basing GPT2 probabilities on the same number of samples as cloze responses.

More specifically, instead of directly calculating $\mathsf{P}(w_t \mid w_{1..t-1})$ from GPT2, we sample a set of words from the conditional distribution $\mathsf{P}(W \mid w_{1..t-1})$. For each context, we sample the same number of words as the number of cloze responses (N) for each context. Since LMs' distributions are typically defined over subword tokens rather than words, we iteratively sample N token sequences from $\mathsf{P}(W \mid w_{1..t-1})$. We first sample a token t_0 and its subsequent token t_1. If t_1 marks the end of a word (by containing a leading whitespace or being a punctuation) we take t_0 as a word sample. Otherwise, we sample t_2 conditioned on t_0 and t_1 and check whether t_2 marks the end of a word. We repeat this process up to t_3: If the end-of-word is not reached by t_3, we take the concatenation of t_0, t_1, and t_2 as our sample. Given N samples of words for each context, we calculate resolution-matched probabilities in a similar manner to Equation 1:

$$\mathsf{P}_{\mathrm{H}_1}(w_t \mid w_{1...t-1}) = \frac{C_{w_t} + 1}{N + V}, \qquad (2)$$

where C_{w_t} is the number of times w_t appears in the set of samples. As with cloze probabilities, we apply $\mathsf{S}(w_t)^2$ with a smoothing factor of $V = 200$ prior to regression modeling. To account for variance in the sampling process, we report median regression model performance over five runs.

Theoretically speaking, this manipulation aligns LM-based estimates with Smith and Levy's (2011) view of the cloze task, where each response from a human subject is seen as a sample from their subjective probability distribution.[6]

Hypothesis 2 ($\mathbf{H_2}$): Semantics Another consequence of the LM's ability to consider any arbitrary continuation is that they can make extremely fine-grained distinctions between different words: They assign different probabilities to, for example, *couch* and *sofa*. Human lexical predictions, however, are influenced by shared semantic features across possible alternatives (Federmeier and Kutas, 1999; Federmeier, 2022; Brothers et al., 2023, among others). Recent work making use of cloze responses that share semantic features has been able to predict processing difficulty more effectively than LM probabilities (Arkhipova et al., 2025).

In order to incorporate the influence of shared semantic features into GPT2's probabilities, we first cluster GPT2's vocabulary using k-means clustering with token embeddings. Subsequently, we define the probability of w_t as the probability of its cluster, which is the total probability of its members:

$$\mathsf{P}_{\mathrm{H}_2}(w_t \mid w_{1...t-1}) = \sum_{w_i \in C_{w_t}} \mathsf{P}(w_i \mid w_{1...t-1}), \qquad (3)$$

where C_{w_t} denotes the cluster which w_t belongs to. Under this manipulation, if *couch* and *sofa* are assigned the same cluster, then they will both be assigned the same cluster probability, which is the sum of the probability of *couch*, *sofa*, and other words that belong to that cluster.[7] If w_t consists of multiple subword tokens, we calculate its probability by taking the product of subword cluster probabilities.[8] We perform k-means clustering with $k \in \{20, 40, 80, 100, 500, 1000\}$, again reporting median regression model performance over five runs for each k, to account for the randomness in clustering. We validate our clustering approach and provide representative examples of cluster members in Appendix B.

This manipulation is also inspired by recent work that incorporates semantic proximity between alternative completions into LM predictability estimates (Meister et al., 2024), and clustering LM representations to operationalize semantic predictability (Jacobs et al., 2025).

Hypothesis 3 ($\mathbf{H_3}$): Frequency Another potential difference between cloze and LM surprisal is that human subjects may not ever produce low-frequency continuations to a context (Smith and Levy, 2011), while effects of word frequency during reading are well documented (Kliegl et al., 2004; Shain, 2024, among others). In contrast, LMs assign probabilities to arbitrary low-frequency continuations, which may account for the improved

[6] Hao et al. (2020) claim neural LMs' probability estimates may be closer to humans' subjective probability distributions compared to n-gram probability estimates, based on their higher fit to RTs and correlation with cloze probabilities.

[7] An anonymous reviewer points out that this implementation may be over-simplistic. An alternative could be to adjust probabilities such that words in larger clusters gain probability, while words in smaller clusters lose probability. However, LM probabilities would then make fine-grained distinctions between words, therefore making it inappropriate to test H_2.

[8] The impact of subword tokenization is likely to be negligible, as very few words in the corpora were split by the GPT2 tokenizer (5.1% for BK21, 5.9% for Provo, and 1.9% for UCL). See Nair and Resnik (2023) and Giulianelli et al. (2024a) for more general discussions.

fit of GPT2 surprisal over cloze surprisal. We test this hypothesis by constraining GPT2 to assign probabilities to only high-frequency tokens.

First, we separate the subword vocabulary of GPT2 into two non-overlapping subsets V_F (frequent) and V_I (infrequent) depending on whether the token meets some frequency threshold. Subsequently, all probability of tokens in V_I are set to zero, and the probability of tokens in V_F is renormalized to sum to one. Finally, a form of add-one smoothing is applied to allow log transform:

$$\mathsf{P}_{\mathrm{H}_3}(w_t \mid w_{1\ldots t-1}) = \begin{cases} \frac{\mathsf{P}(w_t|w_{1\ldots t-1})}{\sum_{w_i \in V_F} \mathsf{P}(w_i|w_{1\ldots t-1})} \times \frac{|V_F|}{|V_F|+1} & \text{if } w_t \in V_F \\ \frac{1}{|V_F|+1} & \text{if } w_t \in V_I. \end{cases} \quad (4)$$

If w_t consists of multiple subword tokens, we calculate its probability by multiplying the probabilities of its subword tokens under Equation 4. We test three different frequency thresholds based on `wordfreq` (Speer, 2022); 10^3 occurrences per billion words, 10^4 occurrences per billion words, and 10^5 occurrences per billion words.

We note that while the three hypotheses are motivated by different aspects of the cloze task, they are not necessarily mutually exclusive.

3.2 Results

The aggregate results in Figure 3 show that all three manipulations result in a significant decrease in fit to human RTs, providing support for all three hypotheses.[9] In other words, GPT2 appears to yield stronger predictors of RTs than cloze responses due to its ability to assign highly fine-grained probabilities (H_1) and distinguish between semantically similar words (H_2) and low-frequency words (H_3).

The by-measure results in Figure 4 show that compared to results from unaltered GPT2 surprisal in Figure 2, the decrease in fit to RT is the most apparent on the Provo ET measures, with cloze surprisal now explaining RTs over and above surprisal from the manipulated GPT2 variants. We speculate that the effect is not as large on the UCL ET measures due to the different characteristics of the corpus. The UCL corpus is a collection of short, isolated sentences that consist of all high-frequency words; therefore, most notably, our manipulation based on H_3 is unlikely to change the surprisal predictors by much.

[9]For H_2 and H_3, which required decisions about the number of clusters and the frequency threshold respectively, we report results using 80 clusters and a threshold of 10^4 per billion as representative examples of the similar trend observed across settings. Results based on different numbers of clusters and frequency thresholds, alongside results from the five individual runs for H_1 and H_2, are reported in Appendix C.

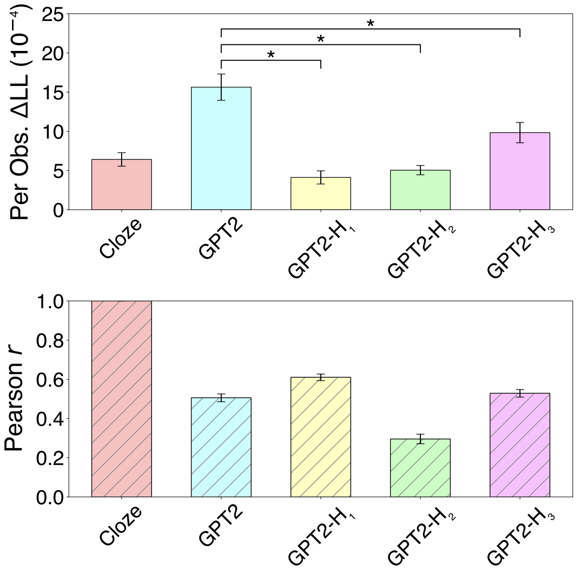

Figure 3: (**Top**) Increase in per-observation log likelihood over the baseline regression models due to including cloze surprisal, GPT2 surprisal, and manipulated variants of GPT2 surprisal, averaged over all 60 folds (10 folds of six measures) used in cross-validation. Error bars denote one SEM across all 60 folds. Among the three comparisons between GPT2 and its manipulated variants, differences that achieve significance at the 0.05 level by a paired permutation test under a 3-way Bonferroni correction are marked with an asterisk. (**Bottom**) Pearson correlation between cloze probabilities and each set of GPT2-based probabilities, calculated over the three text corpora. Error bars denote 95% confidence intervals derived by a permutation test.

4 Experiment 3: Towards Combining Cloze Responses With LM Probabilities

The previous experiment showed that GPT2's ability to assign fine-grained, 'high-resolution' probabilities to the actual observed word w_t provides an advantage for modeling human RTs. In this experiment, we combine cloze responses with GPT2 probabilities through the link of similarity-adjusted surprisal (SA surprisal; Meister et al., 2024, c.f. cooccurrence-based smoothing; Essen and Steinbiss, 1992). SA surprisal is based on the average probability of likely alternatives to the observed word w_t, which is weighted by their distance to w_t. Unlike H_2, SA surprisal considers attested alternative completions for the context, rather than words related in meaning to w_t. It can therefore alleviate issues with the poor resolution of cloze

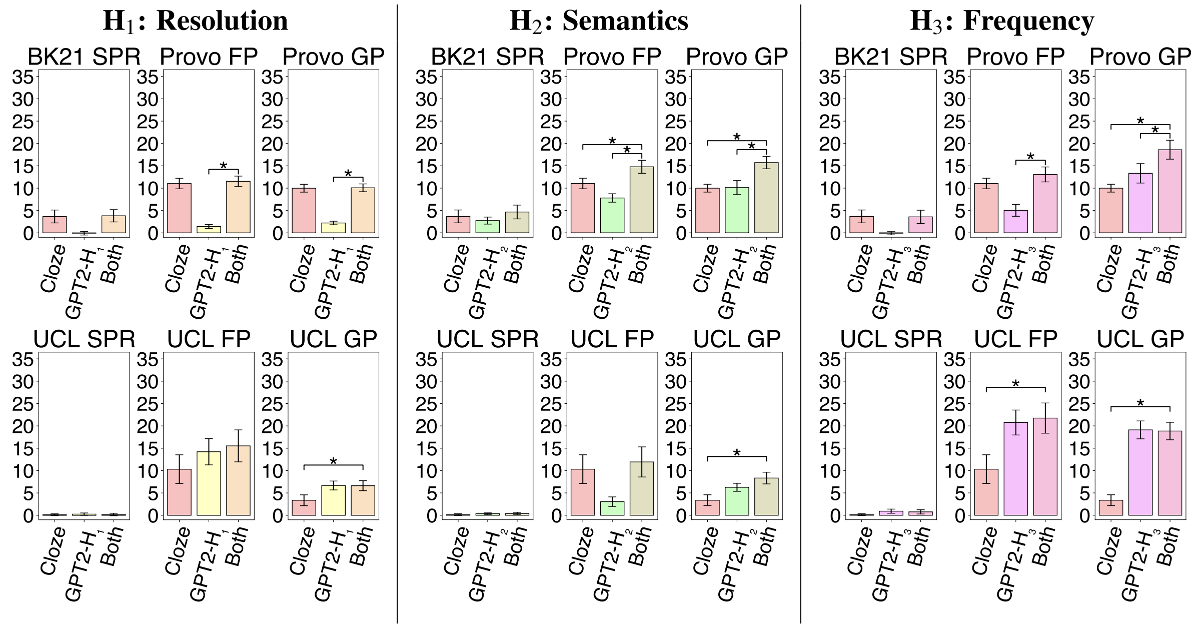

Figure 4: Increase in per-observation log likelihood in 10^{-4} nats over the baseline regression models due to including cloze surprisal, manipulated GPT2 surprisal, and both predictors, averaged over the 10 folds used in cross-validation. Error bars denote one SEM across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT2-$\text{H}_{\{1,2,3\}}$ vs. Both), differences that achieve significance at the 0.05 level by a paired permutation test under a 12-way Bonferroni correction (two comparisons on six measures) are marked with an asterisk.

responses; $\mathsf{P}(w_t)$ can be calculated even if w_t itself is not attested in the set of cloze responses.

By evaluating SA surprisal calculated based on the set of cloze responses (SA cloze surprisal), we ask two questions. The first is whether SA cloze surprisal results in a different fit to RTs compared to 'count-and-divide' cloze surprisal. For example, SA cloze surprisal may capture the fact that a reader's reading may be facilitated by *sofa*, even if they produced only *couch* as their cloze response. If so, this may suggest the need for a different method to derive surprisal from cloze responses.

The second comparison is against SA surprisal calculated based on the set of GPT2's responses (SA GPT2 surprisal). By holding the underlying distance space and probabilities constant, this comparison asks whether the set of alternatives that GPT2 considers (i.e. words that might be predicted instead of w_t) also provides a modeling benefit over the set of cloze responses.

4.1 Methods

SA surprisal is defined as:

$$\mathsf{P}_{\mathsf{S}}(w_t \mid w_{1\dots t-1}) = \sum_{w' \in R_{w_{1..t-1}}} z(w_t, w')\mathsf{P}(w' \mid w_{1\dots t-1}), \tag{5}$$

where $z(\cdot)$ is a similarity function and $R_{w_{1..t-1}}$ is a multiset of responses given the context $w_{1..t-1}$. For $z(\cdot)$, we used normalized cosine distance between token embeddings from GPT2. For $R_{w_{1..t-1}}$, we used the set of cloze responses to calculate SA cloze surprisal, and the set of GPT2 samples matched in number (i.e. samples collected to evaluate H_1) to calculate SA GPT2 surprisal. For words that consist of multiple tokens, we mean-pooled across their embeddings to calculate their distance to w_t, following Giulianelli et al. (2024b). The conditional probability $\mathsf{P}(w' \mid w_{1\dots t-1})$ is calculated from GPT2. Protocols for regression modeling and statistical significance testing follow those of the previous experiments.[10]

4.2 Results

Figure 5 shows that both versions of SA surprisal are poor predictors of RTs, indicating that count-and-divide is not an unreasonable method to convert cloze responses into probabilities. Results are inconclusive about the quality of alternatives between cloze responses and GPT2 samples. We leave the exploration of methods for combining cloze responses with LM-based estimates to future work.

[10] We do not report results on the BK21 dataset as it does not provide the raw cloze responses to be used as $R_{w_{1..t-1}}$.

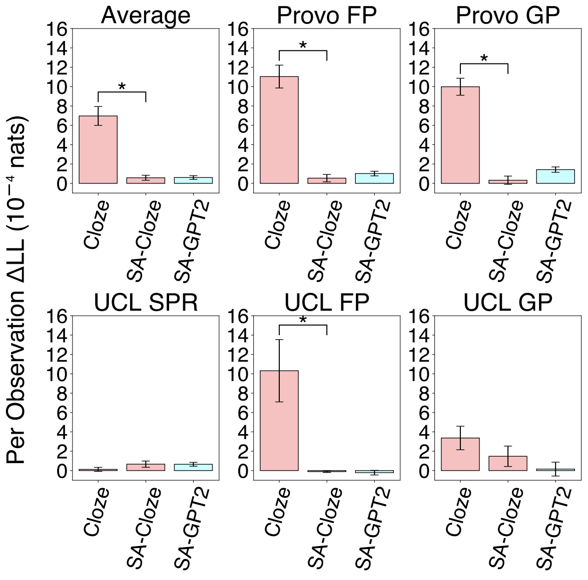

Figure 5: Increase in per-observation log likelihood, averaged over all folds used in cross-validation. Error bars denote one SEM across all folds. Differences that achieve significance at the 0.05 level under a 10-way Bonferroni correction (two comparisons on five measures) are marked with an asterisk.

5 Discussion and Conclusion

To study how linguistic prediction shapes language comprehension, we need to quantify how predictable a word could be to a comprehender. This work compares two very different methods for doing so against word-by-word reading times—using cloze responses and LM probabilities. Consistently with previously reported findings, we first re-establish over multiple datasets that predictability estimates from LMs predict RTs better than those from cloze responses, at least for eyetracking data.[11]

However, instead of treating cloze responses as inferior purely on the basis of relatively poorer fit to RTs, we question where the modeling advantage of LM surprisal comes from. To this end, we tested three hypotheses by manipulating GPT2 probabilities accordingly; namely that GPT2 does not suffer from poor resolution like the cloze task, and that its probabilities can distinguish between semantically similar words and low-frequency words. Our results support all three hypotheses.

Perhaps the most noteworthy finding is that when LMs are used to generate cloze-like estimates from an underlying distribution (i.e. the resolution is lowered), GPT2 surprisal is no longer a stronger predictor of RTs than cloze surprisal. This suggests that cloze surprisal that is based on more responses than are typically collected may be a stronger predictor of RTs than the cloze surprisal evaluated in this work. Nonetheless, simply collecting more responses will not address the inherent limitations of the cloze task, which stem from the fact that it is an untimed production task. As such, we call for alternatives to the cloze task, such that it can better capture predictions during real-time language processing. For example, in addition to the traditional cloze task, timed versions of the cloze task (Staub et al., 2015) or a maze-like variant may help control for factors like conscious reflection that can influence the responses.

At the same time, it remains to be seen whether human-like prediction is as sensitive to the fine-grained distinctions made by LM probabilities. Future experiments should target whether humans' expectations differentiate between e.g. semantically related words or low-frequency words—perhaps using stimuli informed by LM probabilities. This will help ascertain whether LM surprisal indeed predicts RTs because it captures human-like aspects of prediction. While neural network-based LMs offer a lot of convenience for applications in cognitive modeling, what drives their next-word predictions is not fully understood. Therefore, there remains a possibility that their probabilities capture human data during real-time processing due to reasons unrelated to human-like prediction.

From a theoretical perspective, it is also important to understand the empirical limitations of prediction in language comprehension (e.g. Huang et al., 2024; Timkey et al., 2025; Staub, 2025). Therefore, instead of using LM probabilities as a potential one-size-fits-all explanation of all measures of processing effort during language comprehension, we encourage future modeling approaches to be more explicit about the links between various stages of probabilistic inference and individual measures. An example of such class of models is predictive coding models, which not only implement a neurobiologically plausible approach to inference (Clark, 2013), but also offer cognitively interpretable measures that can be linked to different aspects of processing behavior when applied to language comprehension (Nour Eddine et al., 2024; Ohams et al., 2026).

[11] See Buggy et al. (2026) for more arguments about distinct cognitive processes recruited during self-paced reading and eyetracking and their consequences on predictability effects.

Limitations

This work relied on language models trained on English text and data from human subjects that are native speakers of English. Therefore, it remains to be seen whether the findings will generalize to other language models and data collected in other languages. Although multilingual language models and reading time datasets exist, we are not aware of any cross-linguistic datasets of cloze responses aligned to reading time data. In our manipulations for Hypotheses 2 and 3, we relied on multiplying token-level probabilities, instead of applying corrections used for unaltered word probabilities, since these corrections are not be likely to introduce major qualitative changes to our results (Oh and Schuler, 2024; Pimentel and Meister, 2024).

Ethical Considerations

This work used reading time data collected as part of previously published research. We refer readers to the respective publications for the data collection and validation procedures. This work used these datasets for their intended purpose—to study human sentence processing—and therefore we foresee no potential risks associated with this work.

Acknowledgments

We thank Philip Resnik, Samer Nour Eddine, Colin Phillips, Cassandra Jacobs, members of the UMD psycholinguistics community, as well as the audience at the 39th Annual Conference on Human Sentence Processing for valuable discussion, and the ARR reviewers for constructive feedback. This material is based on work supported by the NSF GRFP (No. DGE 2236417) to Sathvik Nair. Research for this paper was made possible by a Start-Up Grant (No. 03INS002748C420) from Nanyang Technological University, Singapore to Byung-Doh Oh.

References

Yana Arkhipova, Alessandro Lopopolo, Shravan Vasishth, and Milena Rabovsky. 2025. When meaning matters most: Rethinking cloze probability in N400 research. *bioRxiv*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2397–2430.

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.

Trevor Brothers, Emily Morgan, Anthony Yacovone, and Gina Kuperberg. 2023. Multiple predictions during language comprehension: Friends, foes, or indifferent companions? *Cognition*, 241:105602.

Ryan Buggy, Stephanie Cho, Cory Shain, and Adrian Staub. 2026. Predictability effects in natural reading are logarithmic, not linear: Evidence from an eye-movement replication of Brothers and Kuperberg (2021). *PsyArXiv*, Kt3sw_v1.

Andy Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

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

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Ute Essen and Volker Steinbiss. 1992. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 161–164.

Kara D. Federmeier. 2022. Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, 59(1):e13940.

Kara D. Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024a. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024b. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. 2026. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Journal of Memory and Language*, 148:104715.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783v2.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Markus J. Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Jacob Louis Hoover, Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Cassandra L. Jacobs, Loïc Grobol, and Alvin Tsang. 2024. Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned. *arXiv preprint*, arXiv:2410.12057v2.

Cassandra L. Jacobs, Ryan J. Hubbard, Loïc Grobol, and Kara D. Federmeier. 2025. Uncovering patterns of semantic predictability in sentence processing. *Journal of Memory and Language*, 144:104653.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

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.

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

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.

James A. Michaelov, Seana Coulson, and Benjamin K. Bergen. 2023. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, 15(3):1033–1042.

Sathvik Nair and Colin Phillips. 2026. Across the levels of analysis: Explaining predictive processing in humans requires more than machine-estimated probabilities. *arXiv preprint*, arXiv:2604.09466.

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

Samer Nour Eddine, Trevor Brothers, Lin Wang, Michael Spratling, and Gina R. Kuperberg. 2024. A predictive coding model of the N400. *Cognition*, 246:105755.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint*, arXiv:2510.05141v1.

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 and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472.

Byung-Doh Oh and William Schuler. 2025. Dissociable frequency effects attenuate as large language model surprisal predictors improve. *Journal of Memory and Language*, 143:104645.

Chiebuka Ohams, Sathvik Nair, Shohini Bhattasali, and Philip Resnik. 2026. A predictive coding model for online sentence processing. *Journal of Memory and Language*, 146:104705.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Cory Shain. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

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

Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. 2024. Language models are better than humans at next-token prediction. *Transactions on Machine Learning Research*.

Sophie Slaats and Andrea E. Martin. 2025. What's surprising about surprisal. *Computational Brain & Behavior*, 8:233–248.

Nathaniel J. Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, pages 1637–1642.

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

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

Miloš Stanojević, Jonathan R. Brennan, Donald Dunagan, Mark Steedman, and John T. Hale. 2023. Modeling structure-building in the brain with CCG parsing and large language models. *Cognitive Science*, 47(7):e13312.

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

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

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

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

William Timkey, Kuan-Jung Huang, Byung-Doh Oh, Grusha Prasad, Suhas Arehalli, Tal Linzen, and Brian Dillon. 2025. Eye movements reveal a dissociation between prediction and structural processing in language comprehension. *PsyArXiv*, eq2ra_v1.

Shravan Vasishth. 2006. On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In *Proceedings of the International Conference on Linguistic Evidence*, pages 96–100.

Ethan Gotlieb 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.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

A Number of Observations for Each RT Measure

Table 2 outlines the number of observations that were analyzed for each RT measure.

B Validation of H_2 Clustering

To determine the stability of the cluster assignments used for H_2, we compared the labels assigned to the tokens in GPT2's vocabulary under different clustering settings. For each value of k, we computed a pairwise Adjusted Rand Index (Hubert and

Measure	Total	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
BK21 SPR	45 840	4578	4601	4592	4575	4573	4591	4575	4574	4602	4579
Provo FP	105 958	10 773	10 483	10 505	10 579	10 782	10 512	10 575	10 500	10 575	10 674
Provo GP	105 775	10 749	10 463	10 490	10 564	10 762	10 494	10 565	10 470	10 561	10 657
UCL SPR	99 865	9836	10 232	9923	9733	10 104	10 015	9827	10 365	9858	9972
UCL FP	41 014	4130	4108	4064	4147	4155	4163	4061	4089	4051	4046
UCL GP	40 993	4129	4101	4064	4145	4152	4162	4058	4087	4049	4046

Table 2: The number of observations in each of the 10 folds and their total for each measure.

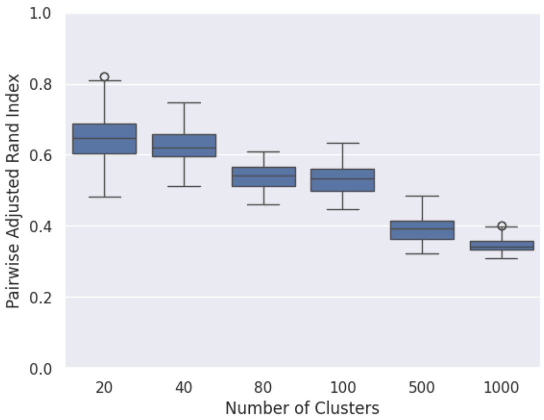

Figure 6: Adjusted Rand Index over all pairs of cluster assignments for different values of k used in H_2.

Arabie, 1985) across all five clustering runs (Figure 6). This measures how often different items were assigned to the same cluster, correcting for random chance variation. Although the stability of the clustering decreases as more clusters are used, all results are far above zero, which would be the value of the Adjusted Rand Index under a random baseline.

We also show an illustrative example of the results of clustering GPT2's static token embeddings in Table 3. This presents select items from five clusters under one of the runs of k-means clustering, where k = 80, reported in Figure 4. Since this is a qualitative example, we also provide manual characterizations of the clusters.

C Complete Results of Experiment 2

Figure 7 shows the performance of GPT2-H_2 surprisal for different numbers of clusters used in k-means clustering. We report the median results over five runs for each value. The trends in statistical significance remain the same for $k \geq 80$, as reported in the main text. When $k = 40$, cloze surprisal predicts RTs over and above GPT2-H_2 surprisal on BK21 SPR and UCL FP, and there is no significant difference between the two on UCL GP. When $k = 20$, GPT2-H_2 surprisal predicts RTs over and above cloze surprisal on BK21 SPR.

Figure 8 shows the performance of GPT2-H_3 surprisal for different frequency thresholds. We report results for 10^4 per billion words (center) in the main text, which shows the same trend as when 10^3 per billion words is used as the threshold (left). When the threshold is increased to 10^5 per billion words (right), GPT2-H_3 surprisal no longer predicts RTs over and above cloze surprisal on Provo GP.

Figures 9 and 10 show results based on individual sampling runs for GPT2-H_1 surprisal and clustering runs for GPT2-H_2 surprisal respectively. Across both hypotheses, there is no substantial variability in the magnitude of increase in per-observation log likelihood or the pattern of statistical significance.

Examples	Characterization
concerned, *angry*, *worried*, *interested*, *aware*, *proud*, *sick*, *surprised*, *tired*, *convinced*	Emotions
day, *month*, *weekend*, *annual*, *Nov*, *May*, *afternoon*, *Thursday*, *night*, *February*	Time
leg, *face*, *foot*, *heart*, *hands*, *wings*, *bones*, *tongue*, *stomach*, *lung*	Body parts
mother, *boy*, *father*, *adult*, *twin*, *teenagers*, *grandfathers*, *youth*, *orphan*, *aunt*	Family
operates, *extends*, *participates*, *condemns*, *remains*, *uses*, *appears*, *becomes*, *disappears*, *collects*	Agentic verbs

Table 3: Examples from select k-means clusters ($k = 80$) over GPT2's static embeddings and their characterization, used in Run #5 on the UCL corpus in our evaluation of H_2.

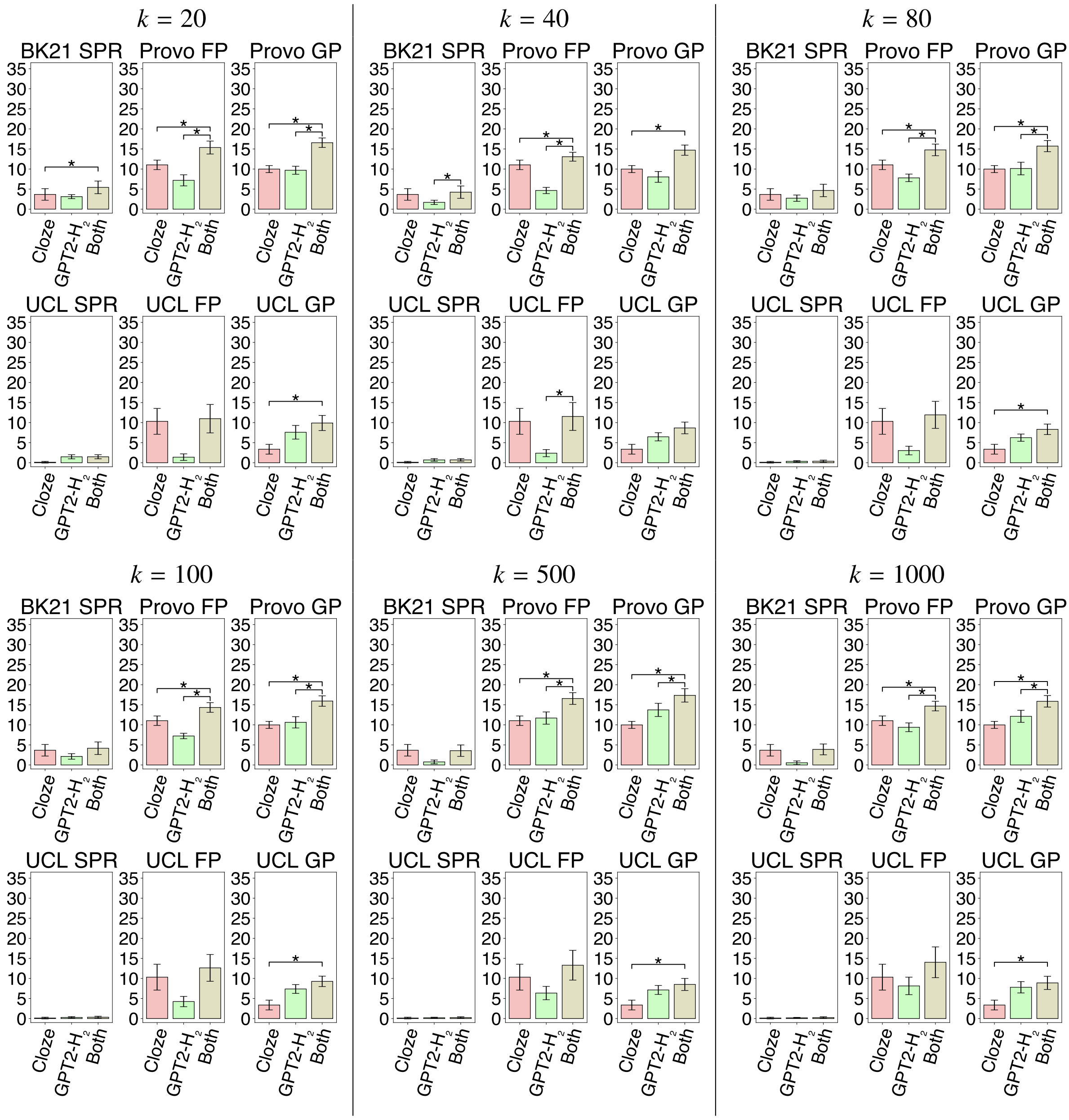

Figure 7: Increase in per-observation log likelihood in 10^{-4} nats over the baseline regression models due to including cloze surprisal, GPT-H_2 surprisal based on different numbers of clusters, and both predictors, averaged over the 10 folds used in cross-validation. Error bars denote one SEM across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT2-H_2 vs. Both), differences that achieve significance at the 0.05 level by a paired permutation test under a 12-way Bonferroni correction (two comparisons on six measures) are marked with an asterisk.

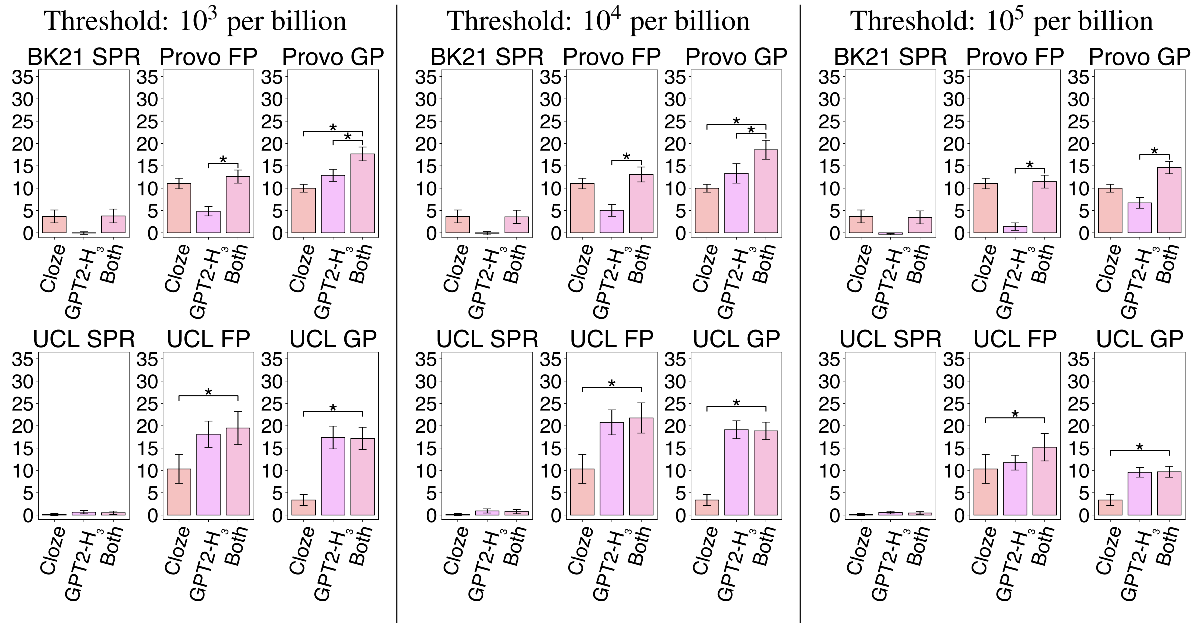

Figure 8: Increase in per-observation log likelihood in 10^{-4} nats over the baseline regression models due to including cloze surprisal, GPT-H_3 surprisal based on different frequency thresholds, and both predictors, averaged over the 10 folds used in cross-validation. Error bars denote one SEM across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT2-H_3 vs. Both), differences that achieve significance at the 0.05 level by a paired permutation test under a 12-way Bonferroni correction (two comparisons on six measures) are marked with an asterisk.

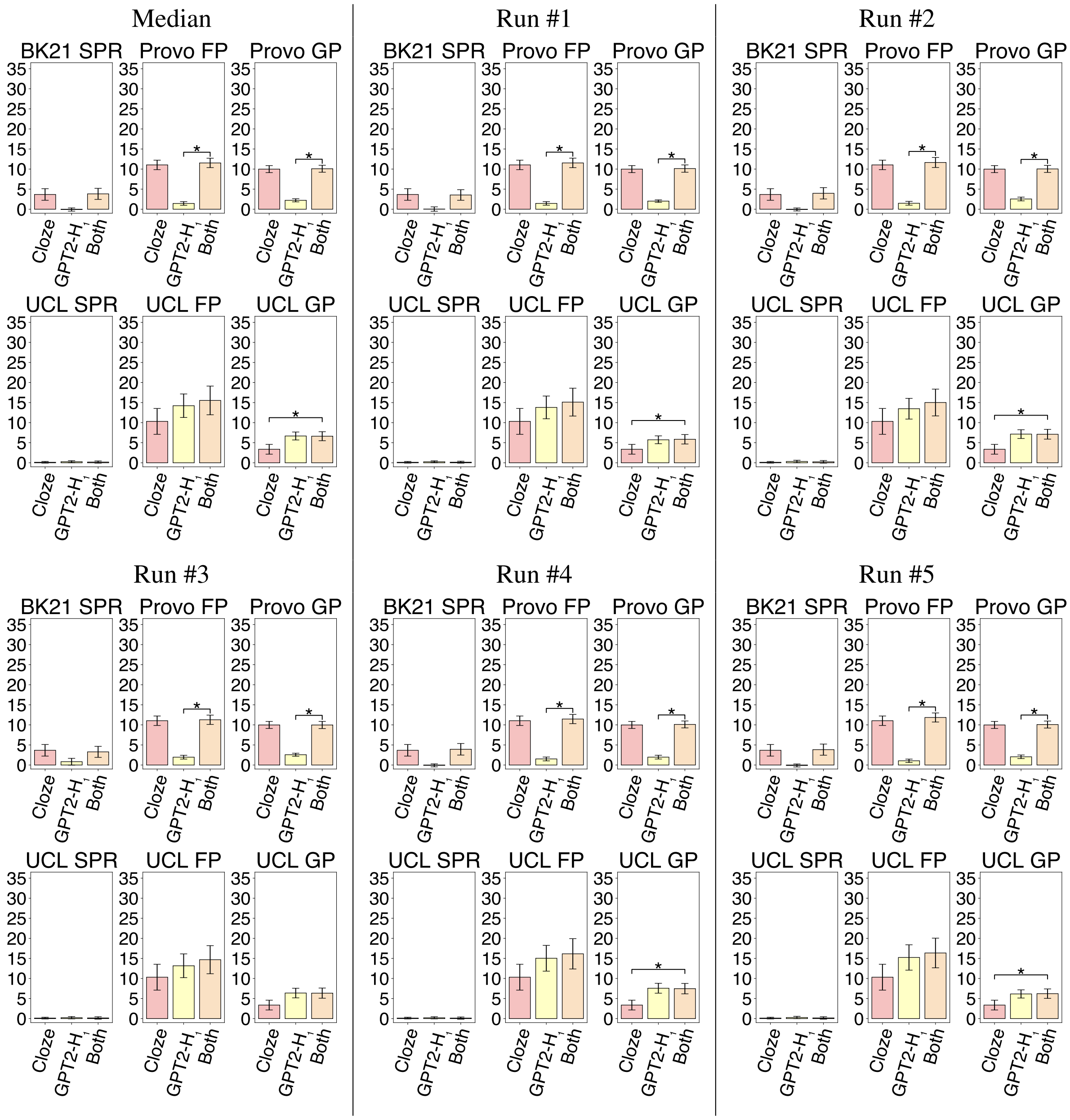

Figure 9: Increase in per-observation log likelihood in 10^{-4} nats over the baseline regression models due to including cloze surprisal, GPT-H$_1$ surprisal from each sampling run, and both predictors, averaged over the 10 folds used in cross-validation. Results based on the median log likelihood on each fold are repeated for comparison. Error bars denote one SEM across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT2-H$_1$ vs. Both), differences that achieve significance at the 0.05 level by a paired permutation test under a 12-way Bonferroni correction (two comparisons on six measures) are marked with an asterisk.

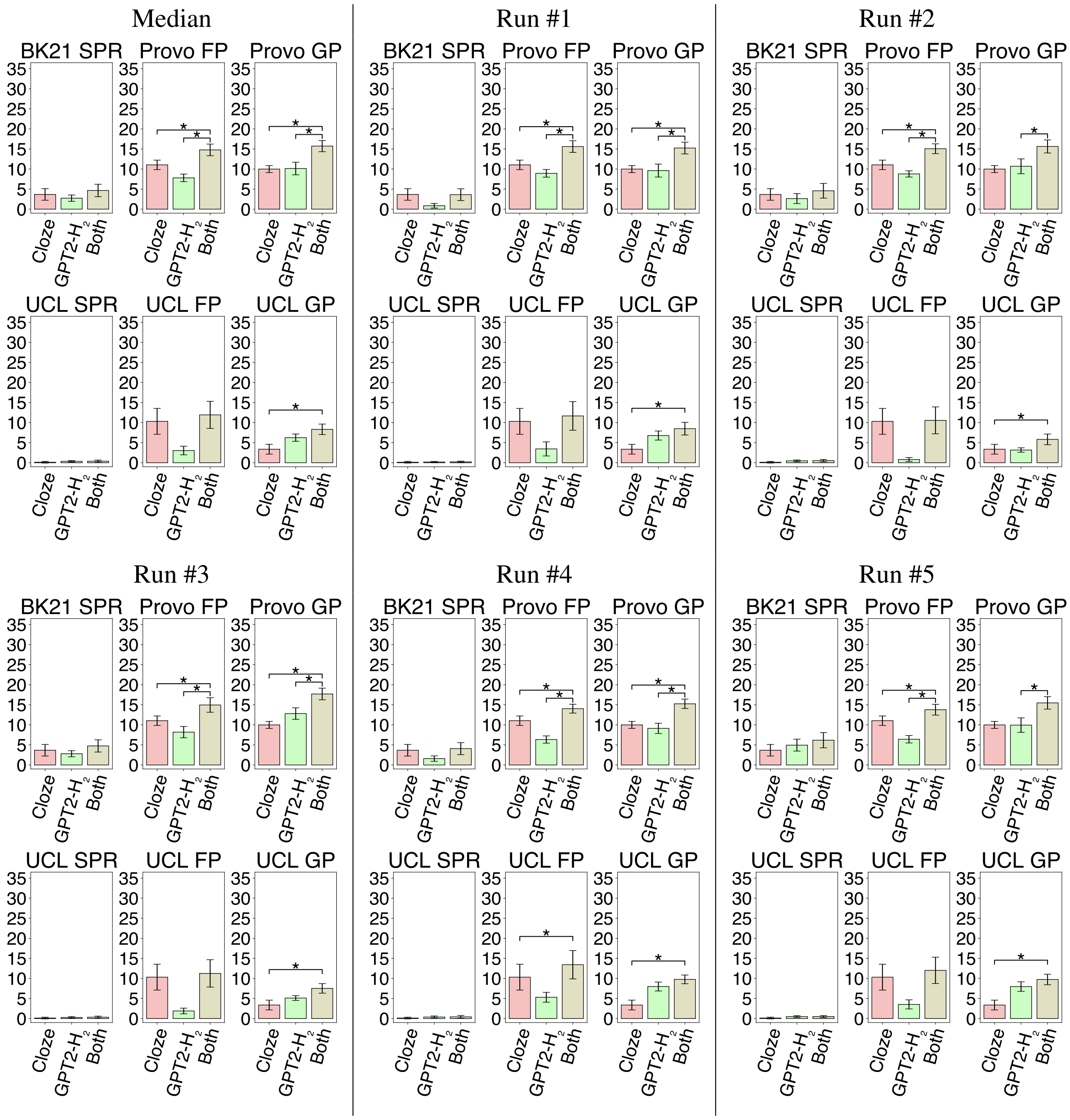

Figure 10: Increase in per-observation log likelihood in 10^{-4} nats over the baseline regression models due to including cloze surprisal, GPT-H_2 surprisal based on 80 clusters from each clustering run, and both predictors, averaged over the 10 folds used in cross-validation. Results based on the median log likelihood on each fold are repeated for comparison. Error bars denote one SEM across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT2-H_2 vs. Both), differences that achieve significance at the 0.05 level by a paired permutation test under a 12-way Bonferroni correction (two comparisons on six measures) are marked with an asterisk.