Correlating Language Model Surprisal With Cloze and Plausibility: Getting the Best of Both Measures

Kate Rebecca Belcher[1], **Matthew W. Crocker**[2]
[1]Leibniz-Institut für Wissensmedien (IWM), Tübingen, [2]Saarland University, Saarbrücken
k.belcher@iwm-tuebingen.de, crocker@coli.uni-saarland.de

Abstract

Prediction is central to both expectation-based theories of human language processing (such as Surprisal Theory), and the objective of neural network-based causal language models, where upcoming tokens are predicted based on their preceding context. With this similarity in mind, we investigated how language model predictions align with human linguistic prediction measures. We investigated the extent to which small-sized causal LLMs capture two common proxy measures of human surprisal – cloze probability and plausibility – in their predictive patterns. For this analysis, we created a new dataset of 660 sentence pair items with a minimal triplet design, in which target words vary across the full scale of word predictability, and calculate metric alignment by way of Pearson correlation. We find a stronger overall correlation of LM-surprisal with plausibility than with cloze, and, notably, the relationships between LM-surprisal and each of the two offline measures is found to vary depending on the relative predictability of the target word. We conclude that LM-surprisal offers a distinct perspective as a predictability measure than both offline behavioural measures, and that it may offer a useful tool in teasing apart nuances in predictability in certain instances which are not always captured by cloze probability and plausibility alone.

Keywords: Surprisal, expectancy, cloze probability, plausibility, large language models

1. Introduction

The expectation-based theories of language comprehension (Hale, 2001; Levy, 2008), state that the difficulty of processing a word is inversely proportional to its *surprisal*: the negative log probability of the word appearing given the preceding context. Words that are more expected in a given context have a lower surprisal and are typically read more quickly, whereas less expected words have a higher surprisal – conveying more information relative to more expected alternatives – and are read more slowly (Smith and Levy, 2013).

The equation for the calculation of surprisal:

$$Surprisal(x) = \log_2 \frac{1}{P(x|\text{context})} \quad (1)$$

can equally be reframed as:

$$Surprisal_{k+1} = -\log P(w_{k+1}|w_1 \dots w_k) \quad (2)$$

which intuitively expresses how surprisal – as an index of human processing effort – relates to the next word prediction language modeling objective used in causal language models. As (Levy, 2008) notes, surprisal creates a "causal bottleneck", such that whilst surprisal theory assumes a probabilistic model underlying linguistic comprehension, it makes no claims regarding specific mechanisms or linguistic representations. Being model-agnostic, surprisal measures can be derived from all types of probabilistic models, including transformer-based (Vaswani et al., 2017) large language models. Recently there has been considerable interest in determining the extent to which surprisal estimates derived from these language models (LM-surprisal) captures effects of predictability in human comprehension. One common approach is to correlate LM-surprisal with on-line measures of comprehension effort such as reading times, based on the hypothesis that these should index the cognitive effort due to human surprisal.

In this paper, however, we rather focus on the two off-line measures – namely cloze probability and plausibility judgements – that are commonly used in psycholinguistics to estimate the expectancy of a word in context. Cloze probabilities, determined by eliciting multiple human continuations for a given context, provide a direct measure of the predicted probability distribution for the upcoming word before it is encountered – something which cannot be directly revealed by reading times. While such cloze-based expectancy estimates have been shown to correlate well with reading times (e.g. Smith and Levy, 2011), they are limited in their ability to distinguish low probability targets. Plausibility jugdements, by contrast, provide an off-line measure of the integration difficulty of a word once it has been read which has been also argued to reflect surprisal (Brouwer et al., 2021), and can be obtained across the entire range of possible targets. In this paper, we explore the relationship between LM-surprisal derived from Transformer-based models and offline measures, across a broad range of expectancy values. Our findings shed light on both how to best estimate human surprisal, and how these proxy measures systematically align and diverge across the surprisal spectrum.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 99–109
16 May 2026.

1.1. Online vs Offline Measures

In studies of human language processing, online measures such as reading-times and eye-tracking measures are those which record information about how the comprehender processes language in real time as they incrementally read the words of a sentence or discourse, and provide a robust measure of comprehension effort. Such online measures of comprehension effort are thus indirectly related to surprisal, via the hypothesis that the two should be inversely proportional (Hale, 2001). Such online indices have been widely investigated, establishing a robust link of surprisal with both reading times (e.g. Shain et al., 2022; Smith and Levy, 2013; Aurnhammer and Frank, 2019; Merkx and Frank, 2021) and neurophysiological brain responses (Michaelov et al., 2024; Frank et al., 2015; Krieger et al., 2025). Offline measures, on the other hand, provide a complementary means to elicit more introspective responses to various dimensions of a linguistic stimulus – requiring the participant to make an active judgement or supply an answer – also allowing participants more time. In the present investigation, we focus on two human offline measures of word expectancy that are known to correlate with processing difficulty, namely cloze probability (introduced by Taylor (1953)) and plausibility judgements.

In a cloze task, participants are presented with a partial sentence prefix, and asked to provide what they judge to be the most likely word to follow, resulting over a sample of participants in a probability distribution of possible continuations. A widely acknowledged limitation of the cloze methodology is that despite providing a good representation of the probability of *likely* continuations, any word that is not mentioned in the study results in a cloze probability of zero. As cloze studies typically include less than 50 participants, this leads to a lack of nuance regarding the probabilities of less expected words, which, according to cloze, are all equally unexpected.

Plausibility is often linked to the idea of 'naturalness' (e.g. Amouyal et al., 2024), 'making sense' (e.g. Federmeier and Kutas, 1999), or how 'likely [it is] to occur in the real world' (Fedorenko et al., 2020). Kauf et al. (2024) link plausibility to world knowledge, concluding that humans judge sentences that align with their knowledge of the world as plausible, and those which violate this world knowledge as implausible. Whilst these definitions indicate a slightly broader scope than cloze, the central idea of expectation and predictability is common, and, as with cloze testing, plausibility judgements are obtained across a sample of people from which an overall judgement is determined for each experimental stimulus.

Though intuitively the relationship between plausibility and cloze probability may seem clear, such that high cloze probability items are typically also highly plausible (Brouwer et al., 2021), the relationship is not linear, and certain scenarios highlight how the concepts diverge, such that unexpected items in terms of cloze may still be judged as plausible (Delogu et al., 2021).

1.2. Comparing Offline Measures and LM-Surprisal

We define language model surprisal (or LM-surprisal) here as the inverse log likelihood of a token's occurrance given the preceding context, as derived from a language model trained with the objective to minimize the perplexity of next word prediction. Tokens refer to word-like or subword strings, 'tokenized' according to a tokenizer algorithm (such as Byte Pair Encoding (BPE) (Wang et al., 2020) or SentencePiece (Kudo and Richardson, 2018)). LM-surprisal is characterized by its data-rich basis and by the possibility to robustly estimate surprisal for any combination of context and continuation, as LM-surprisal estimates are derived from the output layer of the language model, with probabilities generated for all tokens across the entire vocabulary space of the model (Radford et al., 2019; Brown et al., 2020).

Both LM-surprisal and cloze probability serve as metrics to approximate linguistic statistics and word predictability (Smith and Levy, 2011). Rather than representing "true surprisal" values, these measures are better understood as distinct estimates of expectancy derived from different underlying models. A key distinction is that cloze probability is argued to incorporate human-centric factors like world knowledge and metacognitive strategies (Frank et al., 2015, p. 9), which are also often linked to plausibility (Kauf et al., 2024). However, the gap between these metrics is narrowing; Eisape et al. (2020) demonstrate that the correlation between LM-surprisal and cloze probability improves as model architectures become more sophisticated, with Transformer-based models outperforming simpler ones. This relationship is further supported by evidence of significant correlations between LM-surprisal, cloze probability, and plausibility across recent literature (e.g., Michaelov et al., 2024).

We therefore carry out a more systematic investigation into the ways in which LM-surprisal aligns with these two offline correlates of human surprisal, emphasizing the full range of target word predictability. Levy (2008) notes that despite the prominence of cloze probability in work on prediction, it is typically only higher cloze items (with a cloze probability >0.3) that are considered. We target this gap in our work, in which we build a dataset in which target words are designed to cover a broad distribution of word predictability, with a par-

ticular focus on the variation in lower expectation items, given that surprisal theory "relies on difficulty asymmetries between low-probability words" (Levy, 2008, p.42). Specifically, we adapt the design of Federmeier and Kutas (1999), who investigated the effect of target word expectancy based on cloze, and semantic category membership on the N400 ERP brain response. In their study, stimuli consisted of sentence pairs including a context sentence, followed by a continuation that contained the target word. The most expected single-word completion, as determined by cloze testing, determined the target in the high expectancy Condition A. The two target words for Conditions B and C are both unexpected, normed to have a cloze probability of <0.05. Conditions B and C differ crucially, however, in that Condition B is of the same semantic category as the most expected completion, while Condition C is semantically unrelated, belonging to a different basic semantic category. The consequence of this manipulation is that while B and C have similarly low cloze probability, targets in the B condition are generally rated as more plausible – a difference that was also attested in the N400 responses. Federmeier and Kutas also investigate the effect of contextual constraint on plausibility, that is, the extent to which the sentential context leads the comprehender towards a very specific completion of the second sentence, and confirm previous results from Schwanenflugel and LaCount (1988); Schwanenflugel and Shoben (1985) that less constraining contexts license a broader range of possible completions. This led to the finding that low cloze targets in Conditions B and C were judged to be more plausible in low constraint compared to high constraint items.

A subset of the Federmeier and Kutas dataset was later used by Ettinger (2020) to examine pragmatic inference and common sense reasoning in the BERT language model (Devlin et al., 2019). Specifically, Ettinger (2020) notes that common sense knowledge must be used to establish what is being spoken about in the first sentence of the sentence pair (the context), and pragmatic inference in determining how the second sentence relates to it. The experimental design of Federmeier and Kutas (1999) encourages a range of target word sentence completions from highly expected to highly unexpected, and ensures variation in the relative predictability levels within low cloze items, providing desirable characteristics for our experiment. In addition, the experimental set up allows us to investigate whether the same-category preference found in humans by Federmeier and Kutas (1999) is also reflected in patterns of LM-surprisal. Adapting the terminology used in Ettinger (2020), we refer to the variance of fit to the sentential context within zero-cloze items as the *pragmatic fit*. An example item from our stimuli is shown in Table 1.

1.3. Predictions

In light of the differing sensitivities of cloze probability and plausibility judgements, we explore how LM-surprisal aligns with these offline measures across the full spectrum of word predictability. Specifically, we address whether LM-surprisal correlates more strongly with the predictive nature of cloze tasks or the integrative nature of plausibility ratings. Looking beyond correlations with the entire dataset, we investigate the extent to which LM-surprisal may be differentially sensitive to these metrics depending on target expectancy. We predict that cloze probability will strongly correlate with LM-surprisal for expected words (in particular, Condition A), where human active prediction is most salient. Conversely, we examine whether plausibility is able to provide a more robust fit with LM-surprisal for less expected or "zero-cloze" items (Conditions B and C), as it captures the fine-grained pragmatic fit and world knowledge that cloze tasks fail to distinguish. Finally, we investigate if LM-surprisal captures the same interaction between contextual constraint and plausibility observed in humans, particularly in less constraining contexts where a broader range of completions is licensed.

2. Dataset Creation

Dataset creation followed four steps: stimuli generation; cloze testing for the generated stimuli; arrangement of target words for 'unexpected' conditions; and plausibility testing of all items. As in the original work by Federmeier and Kutas (1999), the language of the dataset is English.

2.1. Stimuli Generation

ChatGPT (version GPT 3.5) (Brown et al., 2020) was used for the initial generation of stimuli following the design of Federmeier and Kutas (1999). The model was prompted via the web interface to generate similar sentence pair examples, with examples from the original published stimuli by Federmeier and Kutas (1999) used as examples in a few-shot prompt. The final stimuli items were then compiled by manually reviewing and filtering the resulting synthetic stimuli according to the design criteria.

As in Federmeier and Kutas (1999), target word manipulations were centered around category-based rotations, following the concept of "basic category" as outlined by Rosch et al. (1976). A basic category is described as a category of concrete noun, which carries the most information to distinguish itself from other basic categories. An

Example stimulus	*The power went out during the storm, leaving the whole house in darkness. Susan rummaged in the drawer in search of a ...*				
Condition	Target	Cloze	Plausibility	Av. Cloze (Full dataset)	Av. Plaus. (Full dataset)
A	*torch*	0.8	6.72	$0.591_{\pm 0.223}$	$6.361_{\pm 0.465}$
B	*bulb*	0	3.11	$0.008_{\pm 0.032}$	$3.611_{\pm 1.529}$
C	*screw*	0	1.79	$0.0002_{\pm 0.002}$	$2.747_{\pm 1.326}$

Table 1: Average and example cloze and plausibility statistics

example would be "dog", for which a person can typically name several category members (e.g. "chihuahua", "dachshund", "corgi"), but all category members share the majority of features, such that the basic category "dog" is maximally distinguishable from other basic categories (e.g. "bird", "rodent" etc.). In total, target words stem from 58 distinct basic categories, belonging to 9 higher-level "super-categories" (meaning higher categorical levels such as "animals", "plants", or "tools and household objects").

2.2. Cloze Testing

Stimuli were divided into two lists, stratified such that all target semantic categories were equally represented across the lists. Stimuli were presented one at a time with the final word of the second sentence left blank. Participants, recruited via the crowdsourcing website *Prolific*, responded to each question by typing their completion into the empty text field. 30 participants were obtained for each list, and participants could not participate in both lists, leading to 60 unique participants. All participants were aged between 18 and 65 years old (mean age 39.5) and reported both their *primary language* and *first language* as English. After collection, data was minimally processed through spelling error correction, resolution of number agreement marking, and the unification of spelling variations. Cloze probabilities were then calculated from this data, and ranged from 0.033 (where only one person gave a specific response) to 1.0 (where all 30 participants gave the same response).

2.3. Target Word Selection

The response with the highest cloze probability (i.e. given with the highest frequency) determined the "best completion" target word for high predictability Condition A. Every word appeared as a target in each condition, helping to mitigate differences in conditions arising from lexically specific effects. For Condition B, target words were rotated within their basic semantic category to appear with another sentence pair (i.e. completions for the category 'dog' were rotated among other 'dog' stimuli, etc.). Condition C target words belonged to a different basic category, but the same super-category (i.e. at the next hierarchical level). This was accomplished by rotating target words from conditions A or B of another item which satisfy this criteria. For example, a target word from the basic category 'dog' would appear with a stimulus item from a different basic category of the super category 'animals'. Zero-cloze words were chosen as preference, and if this was not possible, the target word was as low-cloze as possible. For both Condition B and C, small morphosyntactic changes, such as resolving number agreement differences, were allowed to ensure that the resulting sentences were always grammatical, regardless of condition. This resulted in a fully counter-balanced design, such that each target word appeared with a sentence pair in each of the three conditions, resulting in a final dataset comprising of 220 sentence pairs, each with three possible endings.

2.4. Plausibility Ratings

Plausibility ratings were collected for all sentence pair contexts in all three conditions (660 items in total, split across six lists). A modal average of 18 individual responses (mean av. 19.67[1]) were collected for each item. The final plausibility rating for an item was the average of all responses. As with the cloze study, all participants were native English-speaking adults (mean age 37.9), recruited via *Prolific*. Judgements were made on a Likert scale with numbered points and labels at the extreme ends, from 1 (*"not plausible at all"*)' to 7 (*"very plausible"*). We asked participants to judge *"How plausible is the last sentence in the context of the whole text?"* Formulation in this way allowed participants to focus on the entire text and use all information available in their judgements. As the target word was always the final word in the stimulus, the plausibility judgement was always made directly after encountering the target word. In both the cloze and plausibility studies, six attention check items were included throughout the study. Correctly responding to at least four of the six items was a prerequisite for inclusion in the study. No par-

[1] For one list, 28 responses were collected due to a technical error. Post-hoc analysis showed no statistical difference in mean rating with 28 responses compared to a random 18 responses, so all participants were included in the final data.

ticipants were excluded on the basis of the attention checks, so all data was included in the study.

2.5. Dataset Characteristics

The resulting dataset contains 220 sentence pairs, each with three different target word conditions, to give a total of 660 experimental items. The average cloze probabilities and plausibility ratings for each condition are shown in Table 1. The averages by condition suggest a clear trend: Condition A target words are both the most expected in terms of cloze probability, and rated most plausible. While average cloze probability for Condition B is slightly higher than Condition C, both B and C were far less expected than A. Average plausibility similarly decreases from Condition A to Condition C. However, average plausibility for Condition B sits in the middle of the 1-7 scale at 3.611, and Condition C almost one point lower at 2.747, while neither are highly implausible. These more graded average plausibility ratings of Conditions B and C relative to A are consistent with our basic category membership manipulation – same category targets have plausibility ratings closer to A, than different category targets.

At the item-level, Condition A is rated as most plausible in nearly 95% of stimuli, demonstrating a strong human preference for the high expectation word in context. Between conditions B and C, Condition B is rated as more plausible than Condition C in 72.7% of items, leaving just over one quarter of items where Condition C, although semantically further from the best completion, is perceived as more plausible than the within-category alternative. The relationship between cloze and plausibility is shown in the swarm plots in Figure 1. Strikingly, we observe little variation in plausibility for higher cloze probability items, however, in low cloze items, the full range of the plausibility scale is utilized. This clearly demonstrates the divergence of the metrics, and highlights the aforementioned limitation of cloze in capturing the full extent of word expectancy. The full dataset with cloze and plausibility ratings can be found at: https://osf.io/x3wzs/.

3. Language Model Analyses

We correlate LM-surprisal against the human-derived cloze probabilities and plausibility ratings using Pearson correlations. Although simple correlational analyses do not allow us to control for other factors such as lexically specific effects, these potential issues are in part mitigated by the cross-condition counterbalancing design. LM-surprisal measures were obtained by calculating token-level output probabilities of target words, given the context of the sentence pair. Where a target word

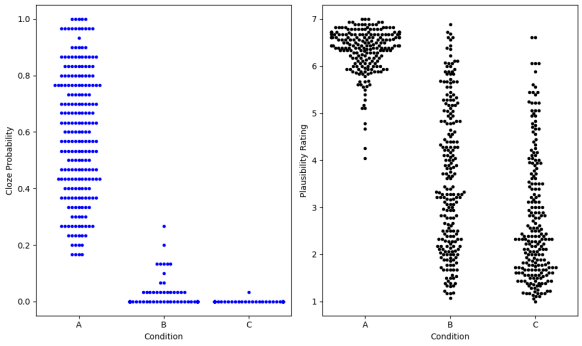

Figure 1: Swarm plots displaying the distribution of plausibility ratings and cloze probability across the dataset, by condition

consisted of multiple sub-word tokens, the respective sub-word token probabilities were summed, following the commonly used methodology for this scenario (Krieger et al., 2025; Kauf et al., 2024). We extracted LM-surprisal values from several small decoder-only language models. All models were open-source, so that token probabilities could be accessed, and were base models (rather than instruction-tuned), to ensure the same language modeling objective (next word prediction). We focus on four comparable models of between 1-2 billion parameters: GPT2-XL (Radford et al., 2019), GPT-Neo (Black et al., 2021), Falcon-RW-1B (Penedo et al., 2024) (henceforth, Falcon-1B), and Qwen3-1.7B-base (Yang et al., 2025) (henceforth, Qwen3-1B). Recent works suggest that models in this smaller size range are more appropriate for modeling human behavioural measures than larger models (Oh and Schuler, 2023; Oh et al., 2024), thus, they are the focus of our investigation. We compare the correlation of LM-surprisal with both cloze probability and plausibility, evaluating the entire dataset, individual conditions, and the modulation of contextual constraint level on the observed patterns. All reported correlations are significant to the 0.01 level unless otherwise indicated.

3.1. Alignment of Cloze and LM-Surprisal

Across the entire dataset, the relationship between cloze and surprisal is clearly observable, with all models achieving a moderately strong negative correlation. Falcon-1B shows the strongest correlation of -0.689. A result of the conditional design of the dataset is the clustering of items in the zero cloze area, which is clearly observable in Figure 2. Therefore, we look at how the correlation strength varies at the individual condition level, to give greater insight into where correlation stems from item level differences, as opposed to the contrast alone of zero-cloze and non-zero cloze items between con-

ditions. The by-condition correlations are reported in Table 2. We observe moderate correlations between LM-surprisal and cloze across the models for Condition A, and, despite the skew towards low-cloze items in Condition B, correlations are comparable to, or even exceed the strength of Condition A correlations across models. Given the intended lack of cloze variation for Condition C items, there is no observable correlation between cloze and surprisal for Condition C.

3.2. Alignment of Plausibility and LM-Surprisal

Table 3 reports the Pearson correlation between LM-surprisal and plausibility for the evaluated models. For the entire dataset, the overall correlation is strong across models, indicating that LM-surprisal accounts well for human-rated plausibility variation. The strongest correlation is observed for Qwen3-1B, slightly exceeding Falcon-1B. At the individual condition level, we observe significant correlations for all conditions, with the weakest correlation in Condition A, where -0.15 $< r <$ -0.25. Correlations in the unexpected (B/C) conditions are significantly stronger, ranging between -0.41 and -0.51, representing moderate negative correlations, which are stronger in Condition C than Condition B. In addition to strong overall alignment, Qwen3-1B shows the strongest correlation for the intermediary Condition B items, however is outperformed by Falcon-1B in Conditions A and C, suggesting that Falcon-1B is better able to approximate predictability at the more extreme ends of the scale. The correlation for Falcon-1B is plotted with all conditions together in Figure 3.

3.3. Constraint Level Analysis

In an additional analysis, we evaluate the effect of contextual constraint in modulating target word plausibility and corresponding LM-surprisal. The data are split into two subsets according to the constraint level of the context. Following Federmeier and Kutas (1999), we use the median cloze probability of the Condition A target word as the threshold. Items where the top cloze word has a cloze probability of 0.6 or higher form the 'high constraint' group.

Table 4 shows average human plausibility by condition and constraint level, with the highest of each constraint pair marked in bold. Higher constraining items are judged as more plausible than less constraining items in Condition A, however in Conditions B and C, we observe the opposite effect, such that less constraining items are viewed as *more* plausible than those with more highly constraining contexts. These between-constraint differences were significant (Mann-Whitney-U test), and reflect the same patterns found by Federmeier and Kutas (1999), further verifying the similarity in characteristics between the two datasets, and indicating that less constraining contexts license a broader range of possible completions (Schwanenflugel and Shoben, 1985; Schwanenflugel and LaCount, 1988). The corresponding constraint-modulated correlations between surprisal and plausibility are also reported for the two best performing models of the four models tested. In Condition A, we observe no significant correlation when broken down by constraint level. In Condition B, for three of the models, we observe a pattern whereby the correlation with high constraint items exceeds the equivalent low-constraint correlation. Falcon-1B is an exception to this, where a stronger correlation is observed for low constraint items. In Condition C, a stronger correlation between LM-surprisal and plausibility is

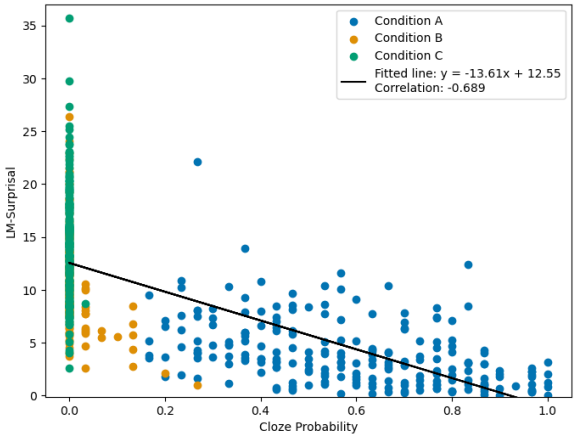

Figure 2: Correlation between Falcon-1B surprisal and cloze probability with conditions colour-coded

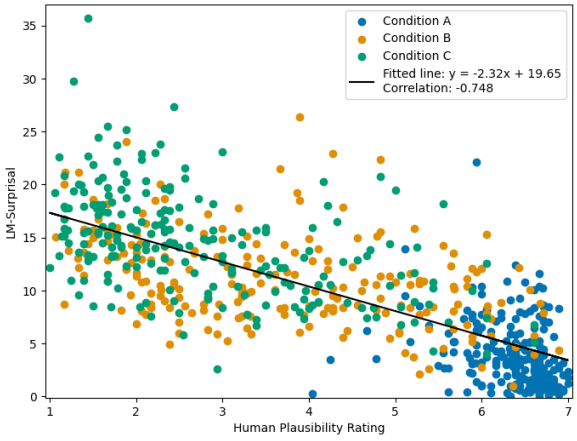

Figure 3: Correlation between Falcon-1B surprisal and plausibility with conditions colour-coded

LM-surprisal vs Cloze	GPT2-XL	GPT-Neo	Falcon-1B	Qwen3-1B
Full data (ABC)	-0.62	-0.625	**-0.689**	-0.671
Condition A	-0.307	-0.296	**-0.386**	-0.279
Condition B	-0.344	-0.358	**-0.375**	-0.35
Condition C	-0.058^{ns}	-0.109^{ns}	-0.078^{ns}	-0.062^{ns}

Table 2: Pearson correlation with cloze for all data and by condition (^{ns} indicates correlation is **not** significant, **boldface** indicates the strongest correlation across models)

LM-surprisal vs Plausibility	GPT2-XL	GPT-Neo	Falcon-1B	Qwen3-1B
Full dataset (ABC)	-0.693	-0.696	-0.748	**-0.753**
Condition A	-0.208	-0.207	**-0.245**	-0.164
Condition B	-0.43	-0.434	-0.415	**-0.466**
Condition C	-0.494	-0.468	**-0.504**	-0.502

Table 3: Correlations of LM-surprisal with plausibility across all data and by condition, **boldface** indicates the strongest correlation across models)

found for *low* constraint items consistently across models. This mirrors the interaction between constraint and plausibility found in Conditions B and C in the human plausibility ratings. The between-constraint differences in the surprisal-plausibility correlations for Conditions B and C did not reach significance, however, the qualitative pattern suggests that LM-surprisal captures plausibility best in the case of *less* expected, and *less* contextually constrained items.

3.4. Comparing Cloze and Plausibility Correlations

We exemplify the side-by-side correlations with cloze and plausibility for Falcon-1B and Qwen3-1B in Table 5. We observe that, in both cases, surprisal provides a statistically significantly better correlation for human plausibility than cloze ($p<0.05$). Comparing the correlations for each condition side-by-side, we find that in Condition A, for expected items, surprisal provides a qualitatively better fit for cloze than plausibility ($p=0.505$ for Falcon-1B). However, for the less expected items in Conditions B and C, surprisal provides a better fit for plausibility than for cloze for both conditions, which is statistically significant in Condition C. Thus, we find that whilst surprisal shows correlations with both cloze probability and plausibility ratings, the strength of correlation varies with both the expectancy of the final word, and the pragmatic expectancy of this word within its sentential context.

Furthermore, for the Falcon-1B estimates, which are overall in greatest alignment with the human measures, a pattern arises whereby, with cloze, the strength of the correlation decreases over the conditions with relative expectancy. The strongest correlation is found in Condition A, with a slightly weaker correlation in Condition B, and no significant correlation at all in Condition C. Conversely, with plausibility, surprisal offers the weakest correlation in Condition A, with an increase in strength for Condition B, and further increase to the strongest correlation in Condition C. Thus, we see an inverse relationship between surprisal and cloze, and surprisal and plausibility as a function of a word's expectancy (whether measured by cloze or plausibility). This across-condition pattern is stable across models for the surprisal-plausibility correlation, and, whilst the corresponding pattern with cloze is less defined for Qwen3-1B, such that the strongest correlation with cloze is in Condition B rather than Condition A, the inverse relationship observed at the extreme ends of the predictability scale still holds, with Conditions A and C showing contrasting patterns for both variables.

4. Discussion

In the present paper we investigated the nature of the relationship between LM-suprisal and offline behavioural measures of expectancy. Our results suggest that LM-surprisal is able to capture expectancy variations at both ends of the scale: In high expectation items, which are more common in active prediction, surprisal reflects some of the nuance in expectation captured by cloze. Moreover, for less expected completions, LM-surprisal still provides a graded picture of word expectation in context, which corresponds to the pragmatic fit of that word in its context. For the highly expected words in Condition A, we found that surprisal was more strongly correlated with cloze probability than plausibility. However, in the low expectation conditions, correlation strength was significantly higher between surprisal and plausibility. Thus, for predictions of a language model, the whole spectrum of relative predictability can be captured more effectively than by either one of cloze or plausibility taken alone. Returning to the predictions in Section 1.3,

	Condition A		Condition B		Condition C	
	H	L	H	L	H	L
Human plausibility	**6.484**	6.232	3.394	**3.837**	2.506	**2.996**
Falcon-1B LM-surp. vs plaus.	-0.11^{ns}	-0.187^{ns}	-0.365	**-0.449**	-0.476	**-0.504**
Qwen3-1B LM-surp. vs plaus.	-0.139^{ns}	-0.064^{ns}	**-0.479**	-0.435	-0.489	**-0.523**

Table 4: Constraint effects on human plausibility and LM-surprisal-plausibility correlation strength. (^{ns} indicates a non-significant correlation, **boldface** indicates the strongest correlation for each H/L constraint pair)

	Falcon-1B		**Qwen3-1B**	
	Surp. vs Cloze	Surp. vs Plaus.	Surp. vs Cloze	Surp. vs Plaus.
Entire dataset	-0.689	**-0.748**	-0.671	**-0.753**
Condition A	**-0.386**	-0.245	**-0.279**	-0.164
Condition B	-0.375	**-0.415**	-0.35	**-0.466**
Condition C	-0.078^{ns}	**-0.504**	-0.062^{ns}	**-0.502**

Table 5: Relationships of LM-surprisal (Falcon-1B/Qwen3-1B) with cloze probability and plausibility (^{ns} indicates lacks of significant correlation, **boldface** indicates the strongest correlation for each pair)

these findings support the idea that LM-surprisal is differentially sensitive to the two measures in accordance with relative expectancy. This pattern proved fairly robust across the models tested.

Our results highlight how the differences between the human offline prediction metrics are heightened when considering words from the extreme ends of the scale in terms of word probability. Cloze probability allows for a more fine-grained distinction in expectancy among high-expectation targets than plausibility ratings, which instead demonstrate ceiling effects such that the probability of high-expectation words in lower constraining contexts is overestimated, relative to those in higher constraining contexts. Conversely, plausibility ratings offer a more graded differentiation of less expected targets, possibly reflecting the importance of pragmatic and world knowledge over purely linguistic constraints in estimating surprisal for these conditions. The constraint analysis offered further support to this analysis, demonstrating a clear trend across all models that for the least expected completions (Condition C), the correlation of surprisal and plausibility is stronger for less constraining items than highly constraining items. Our results pattern with those of both Federmeier and Kutas (1999) and Schwanenflugel and LaCount (1988); Schwanenflugel and Shoben (1985) for human-rated plausibility, with the additional insight that LM-surprisal patterns appear also to follow this pattern for the least expected items.

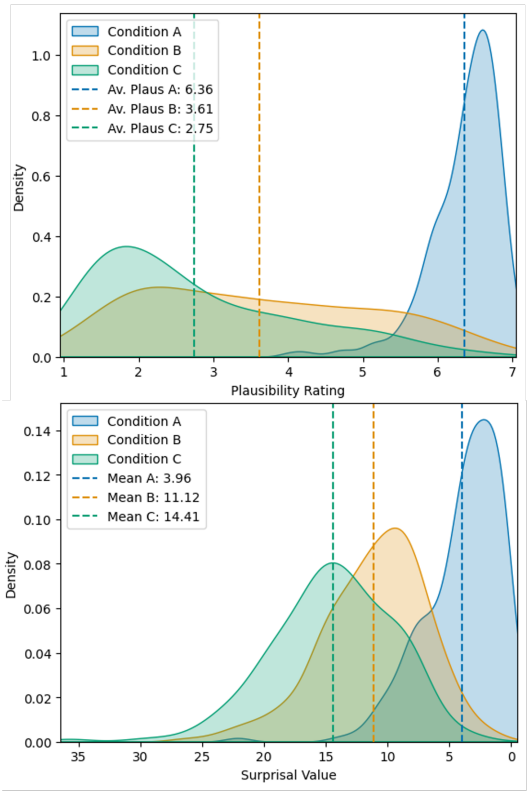

Figure 4: Density plots for human plausibility (upper) and Falcon-1B surprisal (lower) by condition. Surprisal is plotted with lowest surprisal on the right of the scale.

We also observe where LM-surprisal deviates from the two offline metrics. The density plots in Figure 4 illustrate that the categorical design of the dataset resulted in clearly distinguishable peaks in surprisal across conditions. The categorical membership appeared to be less influential for participants judging the plausibility of these continuations, as demonstrated by the wide, evenly spread distribution across the plausibility scale for Condition B. Additionally, although the by-constraint correlation of surprisal and plausibility reflected the same patterns as human plausibility for Condition C, the equivalent result was only found with one model for Condition B, appearing to be less consistent with

the human plausibility results.

5. Conclusion

In this study, we investigated the extent to which LM-surprisal aligned with two common offline human predictability metrics, with a particular focus on low expectation items, which, while crucial for surprisal theory, are often lesser explored in the literature (Levy, 2008). Our dataset design deliberately encouraged the full use of the probability scale for target items, confirming the fact that low expectation in terms of cloze probability does not always correspond to low plausibility. We empirically showed that LM-surprisal is *differentially sensitive* to the two metrics, providing a closer fit for cloze probability than plausibility when words are highly expected, and a closer fit for plausibility than cloze when words are less expected in context. In the lowest expectancy items, the correlation between surprisal and plausibility also displayed similar constraint-based patterns as human plausibility data. These results suggest that the two behavioural metrics provide complementary estimates of human surprisal, with LM-surprisal capturing the strengths of each. Our findings thus emphasize that, while both offline measures offer a valuable means for estimating human surprisal – and thus evaluating language models – neither measure fully captures the entire expectancy spectrum. Our results pave the way for the more nuanced use of these measures both in future evaluations of more sophisticated models, as well as in experimental psycholinguistic investigations. Future experimental work may wish to combine both offline human measures, especially depending on the range of expectancies in the data. Further language model analyses may seek to investigate how LM-surprisal correlates with plausibility as model size increases, and whether the patterns found by Oh et al. (2024) with naturalistic corpora are also applicable for low-expectancy stimuli such as those in our study. Overall, our results confirm that while cloze is a well-established measure for distinguishing more and less expected words, it fails to capture the relative expectancy of less expected and implausible items. Plausibility offers a potential alternative human measure to quantify expectancy for these items. Importantly, however, language models demonstrate the potential to provide a robust estimate of human surprisal across the whole distribution of expectancy, modeling the best elements of both measures.

6. Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany —Project-ID 232722074—SFB 1102. The first author carried out this work while at Saarland University and the University of Malta, and gratefully acknowledges the support of the Erasmus Mundus European Masters Programme in Language and Communication Technologies (EU grant no. 2019-1508), and funding from the Joachim Herz Stiftung through the ALEE project. We kindly thank the anonymous reviewers for their comments.

7. Limitations

We acknowledge several limitations in the study. The conclusions drawn about LM-surprisal can only be drawn in relation to the specific models tested, acknowledging the nature of the models as causal language models trained on next word prediction, and as small LLMs, the largest of which has 1.7 billion parameters. This is notably smaller than the largest and most state of the art language models now available. However, larger models have been shown to offer a reduced ability to model human behavioural measures compared to those in the range evaluated here (Oh and Schuler, 2023), and, such models are typically closed source, which hinders the analysis of output probabilities, and may be trained with additional training objectives, such as instruction-tuning, reducing the link between the original motivation for the study. While language models were used both in the stimuli generation process and in the evaluation of the stimuli, the types of LMs used for each task have fundamentally distinct training objectives, with the instruction-tuned chatbot model ChatGPT used in stimuli generation, and base language models – pretrained on next-token prediction only – for the evaluation. Although we use a counter-balanced dataset design, we acknowledge that using Pearson correlation as our primary analysis method may not fully reveal the relationship of the predictors we consider here. Further analyses, such as regression analyses, could help to tease apart these factors.

8. Ethics Statement

All participants in the data collection studies for the creation of the dataset used in this research gave their informed consent and received remuneration for their participation. We believe any broader ethical concerns of the conducted analyses to be minimal.

9. Bibliographical References

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In

Findings of the Association for Computational Linguistics: EACL 2024, pages 166–181, St. Julian's, Malta. Association for Computational Linguistics.

Christoph Aurnhammer and Stefan Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-TensorFlow.

Harm Brouwer, Francesca Delogu, Noortje J. Venhuizen, and Matthew W. Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Francesca Delogu, Harm Brouwer, and Matthew W. Crocker. 2021. When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, 1766:147514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

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.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

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*.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W. Crocker. 2025. On the limits of llm surprisal as a functional explanation of the n400 and p600. *Brain Research*, 1865:149841.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, page 12–22, Online. Association for Computational Linguistics.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

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

EMNLP 2023, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data only. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

P. J. Schwanenflugel and K. L. LaCount. 1988. Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:344–354.

Paula J Schwanenflugel and Edward J Shoben. 1985. The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, 24(2):232–252.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

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

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9154–9160.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.