Should language models replace the cloze task forever?

Sathvik Nair[1] (sathvik@umd.edu) and Byung-Doh Oh[2] [1]UMD, [2]NTU Singapore

Prediction is a key principle of language comprehension: Words that are more expected given a context are easier to process compared to unexpected words [1–3]. A quantitative understanding of this process requires a measure of how predictable a word is in its context to human readers (i.e. its *predictability*). The traditional method for estimating predictability is the cloze task [4]: Words that are produced more frequently in a particular context are considered to be more predictable. In lieu of conducting human cloze studies, predictability can also be estimated with language models [LMs; 5–7]: These neural networks learn the probability of each word as a result of being trained to predict the next word. LM surprisal (negative log probability) generally seems to be a superior predictor of comprehension difficulty compared to cloze surprisal, which has led to claims that LMs should replace the cloze task altogether [8–10, but see 11]. However, 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 [12, 13]. On the other hand, 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 during real-time processing [14, 15]. Given that this difference results in qualitatively different next-word predictions by humans and LMs [16, 17], it is important to establish where advantages of LM surprisal comes from.

To this end, we analyzed reading times (RTs) in four English self-paced reading (SPR) and eye-tracking (ET) datasets for which cloze responses are available [BK21 SPR, Provo ET, UCL SPR, UCL ET; 11, 18–20]. We first compared the fit of cloze and LM surprisal to held-out RTs, with an increased focus compared to previous work on exploring different methods of smoothing (i.e. assigning probabilities to unobserved responses) and transforming (i.e. assuming different functional forms to RT) cloze probabilities. After determining the smoothing setup that achieves the best fit to about 50% of the data points, four linear mixed-effects (LME) models were fit to each dataset: One containing only the baseline predictors (word length, word position, unigram surprisal [frequency], and whether the previous word was fixated [ET only]), one additionally containing either cloze or GPT-2 [6] surprisal, and one additionally containing both cloze and GPT-2 surprisal (Table 1). Their fit to RTs was evaluated using ten-fold cross-validation; after splitting each dataset into 10 folds, held-out log-likelihood was calculated on the fold that was not used for model fitting. GPT-2 surprisal predicted RTs over and above cloze surprisal, but not vice-versa, indicating that GPT-2 surprisal subsumes the effect of cloze surprisal (Figure 1).

Where does the benefit of GPT-2 surprisal over cloze surprisal come from? We adopt a novel approach of intervening on the LM's probabilities, which can be used to test hypotheses about this observation. We test the following three hypotheses reflecting biases in the cloze task: LM surprisal has 'higher resolution,' since it is based on neural network representations (H_1), LM surprisal can distinguish between semantically related words (H_2), and LM surprisal can be reliably calculated for low-frequency words (H_3). To test H_1, we match GPT-2's resolution by sampling the same number of words as the number of responses that were collected in the corresponding cloze dataset. For H_2, we generate clusters of the tokens in GPT-2's vocabulary and sum the probabilities of individual items in a given word's cluster. For H_3, we limit GPT-2's vocabulary to high-frequency tokens. Evaluation following the same methods shows that this manipulation substantially decreases fit to RTs, relative to unadjusted GPT-2 surprisal (Figure 2).

Overall, when cloze-like constraints are imposed on LMs, the modeling advantage of LM surprisal goes away. Further experimentation is needed to improve the resolution of cloze data and to determine if humans are indeed sensitive to fine-grained differences in LM probabilities.

Datasets	LME Model	Formula
SPR	Baseline	`RT ~ unisurp + length + index + (1 | subject)`
	Cloze	`RT ~ clozesurp + unisurp + length + index + (1 | subject)`
	GPT-2	`RT ~ GPT-2surp + unisurp + length + index + (1 | subject)`
	Cloze + GPT-2	`RT ~ GPT-2surp + clozesurp + unisurp + length + index + (1 | subject)`
ET	Baseline	`RT ~ unisurp + length + index + pfix + (1 | subject)`
	Cloze	`RT ~ clozesurp + unisurp + length + index + pfix + (1 | subject)`
	GPT-2	`RT ~ GPT-2surp + unisurp + length + index + pfix + (1 | subject)`
	Cloze + GPT-2	`RT ~ GPT-2surp + clozesurp + unisurp + length + index + pfix + (1 | subject)`

Table 1: LME formulae used in the experiments. The analyzed datasets did not consistently support a richer random-effects structure. `unisurp`: unigram surprisal, `index`: word position within the sentence, `pfix`: whether the previous word was fixated. All predictors were z-transformed.

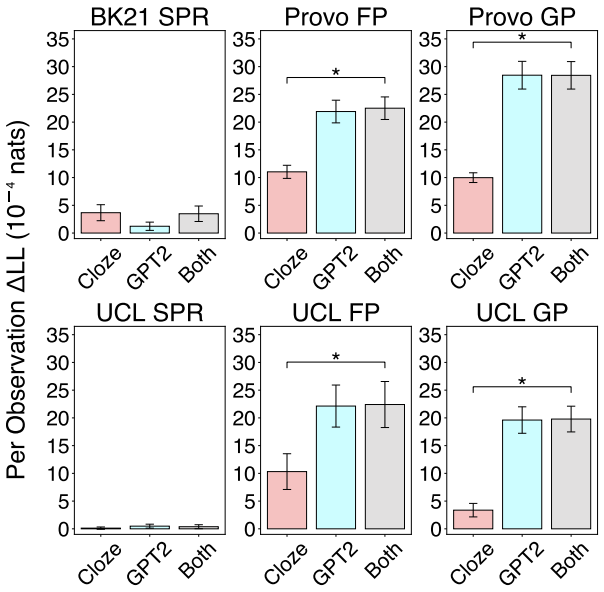

Figure 1: Increase in per-observation log likelihood over the baseline regression models due to including cloze surprisal, GPT-2 surprisal, and both predictors, averaged over the 10 folds used in cross-validation. Error bars denote one standard error of the mean (SEM) across the 10 folds. Among the two comparisons of interest (Cloze vs. Both; GPT-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.

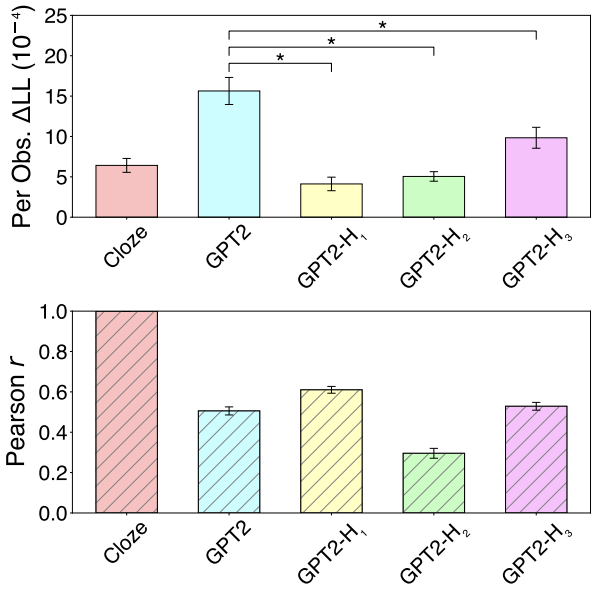

Figure 2: (**Top**) Increase in per-observation log likelihood due to including cloze surprisal, GPT-2 surprisal, and its manipulated variants, averaged over all folds used in cross-validation. Error bars mark one SEM. Asterisks mark significant differences between GPT-2 and its manipulated variants under a paired permutation test, where significance is at the 0.05 level under a 3-way Bonferroni correction. (**Bottom**) Pearson correlation between cloze probabilities and each set of GPT-2 probabilities, calculated over the three text corpora.

[1] Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading.
[2] Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association.
[3] Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic.
[4] Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability.
[5] Jozefowicz, R. et al. (2016). Exploring the limits of language modeling.
[6] Radford, A. et al. (2019). Language models are unsupervised multitask learners.
[7] Biderman, S. et al. (2023). Pythia: A suite for analyzing large language models across training and scaling.
[8] Hofmann, M. J. et al. (2022). Language models explain word reading times better than empirical predictability.
[9] Michaelov, J. A. et al. (2023). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements.
[10] Shain, C. et al. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time.
[11] de Varda, A. G. et al. (2024). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data.
[12] Oh, B.-D. and Schuler, W. (2023). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?
[13] Oh, B.-D. and Linzen, T. (2025). To model human linguistic prediction, make LLMs less superhuman.
[14] Staub, A. et al. (2015). The influence of cloze probability and item constraint on cloze task response time.
[15] Brothers, T. et al. (2023). Multiple predictions during language comprehension: Friends, foes, or indifferent companions?
[16] Jacobs, C. L. et al. (2024). Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned.
[17] Shlegeris, B. et al. (2024). Language models are better than humans at next-token prediction.
[18] Frank, S. L. et al. (2013). Reading time data for evaluating broad-coverage models of English sentence processing.
[19] Luke, S. G. and Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms.
[20] Brothers, T. and Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension.