Surprising narrative events elicit convergent responses in the brain, subjective reports, and large language models

Ziwei Zhang (zz112@uchcicago.edu)
Department of Psychology, University of Chicago
Institute for Mind and Biology, University of Chicago

Jadyn Park (jadynpark@uchicago.edu)
Department of Psychology, University of Chicago

Isabella Summe (summe@uchicago.edu)
Department of Psychology, University of Chicago

Kruthi Gollapudi (kruthig@uchicago.edu)
Department of Psychology, University of Chicago

Yuan Chang Leong (ycleong@uchicago.edu)
Department of Psychology, University of Chicago
Institute for Mind and Biology, University of Chicago
Neuroscience Institute, University of Chicago

Monica Rosenberg (mdrosenberg@uchicago.edu)
Department of Psychology, University of Chicago
Institute for Mind and Biology, University of Chicago
Neuroscience Institute, University of Chicago

Abstract

Linguistic surprise occurs when incoming linguistic information violates expectations formed from prior context. For example, when we hear a story, we are surprised when unfolding events do not align with our expectations. Here we ask whether large language models (LLMs) represent event-level surprise similarly to humans. To measure LLM surprise in two stories, we asked an LLM to generate text predictions as increasing amounts of context were revealed. For each story event, we operationalized LLM surprise as the dissimilarity between LLM's internal embeddings of the predicted and actual text. We measured human surprise during the same events with self-reported ratings and predictions of a brain-based model of surprise applied to fMRI data. LLM surprise was significantly correlated with self-reported and brain-predicted surprise across events. This suggests that LLMs and humans predict the same events as surprising. Our findings highlight LLMs' potential in modeling human surprise to narrative events.

Keywords: Linguistic surprise; natural language processing; large language models; event perception; network dynamics; fMRI

Introduction

Linguistic surprise occurs naturally as we listen or read. For example, it would be more surprising to hear "you are" than "ice is" at the end of the sentence, "Chicago winters are as cold as….". While LLMs have been shown to encode word-level surprise (Goldstein et al., 2022), humans construct event models during comprehension by integrating linguistic input, and experience surprise when new information contradicts these models. Do LLMs encode event-level surprise in a way similar to humans and thus provide a good model of human linguistic surprise? We approximated model "surprise" for events in two stories by asking an LLM to predict upcoming text and calculating the dissimilarity between the model's internal embeddings of the predicted and actual text. We compared this LLM metric of surprise to (1) human subjective reports and (2) a validated brain network model of surprise that has been shown to track surprise in non-linguistic contexts (Zhang et al., 2024).

Methods

Measuring LLM-surprise. We used Llama-3.1-8B (Touvron et al., 2024) to quantify linguistic surprise at different horizons for two stories (*Paranoia,* 3451 words [Finn et al., 2018]; and *How to draw*, 1951 words [LeBel et al., 2023]). We manually segmented the stories into 61 and 36 events, respectively. Starting with the first 10 words, we incrementally revealed one word at a time. At each step, we asked the model to generate the subsequent N words (top_k=50, temperature=1.2, max_new_tokens=N) for N=1,10,20,30. This was repeated i=30 times to capture the variety in text generation and yielded T time points (corresponding to each word-added step) for each N and i. At each time T, we extracted embeddings from the middle layer (layer 20, 4096 features) and averaged the embeddings across N and i, forming a 4096-by-T matrix E_p for each N. We extracted embeddings for the corresponding N words in the actual story, forming a 4096-by-T matrix E_a. Cosine dissimilarity between predicted (E_p) and actual (E_a) embedding at each time point T served as our measure of LLM-surprise, indicating how much the generated text diverged from the actual text. To capture event-level surprise, we averaged LLM-surprise within each event for each N.

Measuring self-reported surprise. In two behavioral studies, two groups of 30 participants listened to the two stories while providing real-time surprise ratings via a slider (ranging from "completely not surprised" to "completely surprised") and were instructed to adjust the slider whenever their level of surprise changed. To mirror the LLM-surprise analysis, we calculated word-level human-rated surprise by forward filling the surprise of each word between two ratings. We z-scored participants' surprise ratings, averaged them across participants, and computed event-level surprise by averaging within each event.

Measuring a neural signature of surprise. Previously, we developed the surprise edge-fluctuation-based predictive model (EFPM), a brain network model whose interactions, measured via fMRI, predicted surprise in both active learning and passive viewing tasks (Zhang et al., 2024). We analyzed existing fMRI data collected as 22 different

participants listened to *Paranoia*. In contrast to the adaptive learning task (McGuire et al., 2014) performed by participants in the fMRI dataset used to train the model, this passive listening task did not involve visual input or motor responses. We thus excluded visual and motor brain regions from our original EFPM. We applied this lesioned EFPM to every fMRI time point from each participant following our previous work (see *Methods* in Zhang et al., 2024) and averaged the resulting EFPM score time courses across participants. We averaged the scores within event to get one score per event.

Relating LLM, self-reported, and brain-predicted surprise. To test whether LLM-surprise tracks human surprise measured both behaviorally and neurally, we calculated the Pearson's correlation *r* between event-level LLM-surprise and event-level human-rated and EFPM-predicted surprise for each horizon *N.* Significance was assessed non-parametrically by generating null correlation distributions from phase randomizing one of the variables 1000 times before averaging into events and computing *r*. Two-tailed *p* was calculated as (1 + number of abs(null values) ≥ abs(observed value)) / (1 + 1000).

Results

LLM-surprise predicts self-reported surprise. We approximated surprise by asking LLM to generate word predictions of varied lengths *N*. LLM-surprise correlated with human-surprise in both stories (Table 1), suggesting that LLMs are sensitive to self-reported surprise. Moreover, this relationship is consistent across stories when the model predicts *N*=10-30 future words (*N*=20 reported here) but not when it predicts at the single word level (*N*=1).

Brain model of surprise (EFPM) predicts self-reported- and LLM-surprise. The lesioned surprise EFPM predicted human-rated surprise (*r*=0.355, *p*=0.018; Figure 1) as well as LLM-surprise when the model was asked to generate 10 and 20 words (*r*=0.325, *p*=0.021; *r*=0.321, *p*=0.036), but not when *N*=1 or 30 (*r*=0.165, *p*=0.359; *r*=0.279, *p*=0.091) in *Paranoia*. This suggests that a brain network defined to predict surprise in an independent learning task generalized to predict the surprise level of story events rated by humans and LLMs. These findings provide converging evidence that LLM-surprise aligns with human behavioral and neural measures of surprise.

Discussion

We found that LLM-surprise predicts event-level surprise measured from both human ratings and brain-based models. This suggests that LLMs encode expectations over multi-word chunks and represent surprise similarly to humans by integrating expectation violations in an event. Moreover, the surprise EFPM procedure can be used as a domain-general neural predictor of surprise in linguistic and non-linguistic tasks across modalities. These results suggest that LLMs offer a computational framework for investigating how people build and update event representations. By capturing event-level surprise, LLMs may help reverse-engineer the neurocognitive mechanisms underlying narrative expectation and comprehension. More broadly, this alignment opens up new opportunities to link language models with theories of narrative and event representation.

Tables

Table 1. Event-level LLM and human surprise.

Story	Horizon (N)	*r* value	two-tailed *p*
Paranoia	N=1	0.379*	0.022
	N=20	0.383*	0.015
How to draw	N=1	-0.462*	0.001
	N=20	0.689*	0.003

Figures

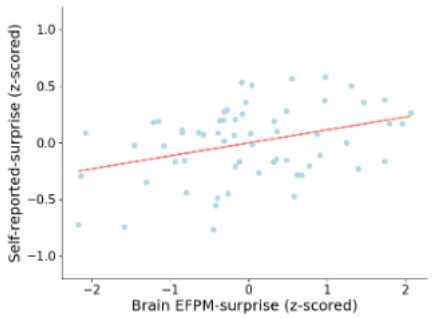

Figure 1: Correlation (*r*) between event-level EFPM score and human surprise in *Paranoia*.

Acknowledgments

This work was supported by resources provided by the University of Chicago Research Computing Center.

References

Zhang, Z., & Rosenberg, M. D. (2024). Brain network dynamics predict moments of surprise across contexts. *Nature Human Behaviour*, 1–15. https://doi.org/10.1038/s41562-024-02017-0

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models | Nature Neuroscience. *Nature Neuroscience*, 25(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Touvron, H., Martin, L., Lu, K., Bhosale, S., Dettmers, T., Ott, M., et al. (2024). *LLaMA 3: Open Foundation and Instruction-Tuned Language Models*. Meta AI. https://ai.meta.com/blog/meta-llama-3/

Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A., & Constable, R. T. (2018). Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nature Communications*, 9(1), Article 1. https://doi.org/10.1038/s41467-018-04387-2

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., & Huth, A. G. (2023). A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1), 555. https://doi.org/10.1038/s41597-023-02437-z

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, 84(4), 870–881. https://doi.org/10.1016/j.neuron.2014.10.013