

Norms of Error Terms in Transformer Language Models Predict Human Reading Time Beyond Surprisal

Ziyan Cao

Department of Linguistics
The Ohio State University, USA
cao.1610@osu.edu

William Schuler

Department of Linguistics
The Ohio State University, USA
schuler.77@osu.edu

Abstract

It is possible that expectation-based latency in reading time is caused by mechanistic latency in changing internal activation states to match observations; this is a primary motivation for predictive coding models (Ohams et al., 2026). This paper explores measurement of sentence processing effort that is based on the gradient signals of transformer language models during error backpropagation: the 2-norm of the gradient of surprisal w.r.t. the output of intermediate layers, i.e., the norm of the error terms at intermediate layers. Regression experiments on multiple human reading time datasets show that the norm of error terms at the output of intermediate layers of GPT-2 small add predictive power beyond baseline predictors including standard surprisal from the same model.

1 Introduction

Surprisal theory posits that processing difficulty during sentence comprehension is proportional to the negative log probability of the current word given the context (Hale, 2001; Levy, 2008). Levy (2008) formalized this idea in a resource-allocation account where processing effort is related to the cost of updating the comprehender’s belief over a set of complete structures. One limitation of the original KL divergence formulation is that it does not consider similarity relations among structures. It does not distinguish probability mass shifts between similar structures from those between more distant structures. Meister et al. (2024) address a related issue with similarity-adjusted surprisal, which incorporates semantic similarity among possible observed continuations. However, their notion of similarity is defined over words rather than over the latent structural representations considered by Levy (2008).

It is possible that different layers of hidden states learned by neural language models can represent state changes that human comprehenders are sen-

sitive to (Parviz et al., 2011a; Ettinger et al., 2016; Fitz and Chang, 2019). Thus the error propagated to the hidden states could represent direction and magnitude of the adjustments needed to account for the observed word. This magnitude would account for the similarity among belief states that the comprehender is maintaining while surprisal only considers the proportion of probability mass retained by exact match. Such a measure is also related to the temporal predictive coding model of sentence processing (Ohams et al., 2026): under the fixed prediction assumption (when top-down predictions are fixed during inference), the displacement of the hidden states converges to the gradient in backpropagation (Millidge et al., 2022).

In this paper, we explore whether the size of error propagated to hidden layers of transformer language models, as a model of the amount of cognitive state change needed in sentence processing, can serve as predictors of reading time. In particular, we investigate the 2-norm of gradient of surprisal with respect to hidden layers in GPT-2 small. Experiments on multiple reading time corpora show that norms of error provide additional predictive power beyond surprisal.

2 Gradient-Based Update Costs

The error propagation explored in this paper may be considered a generalization of the ‘semantic surprisal’ of Meister et al. (2024). Let $\mathbf{h}_\ell$ be the hidden representation at the output of the ℓ -th transformer layer. Let $\mathbf{h}_L$ be the output of the final transformer layer. Consider the negative log-probability of the observed token, i.e., the surprisal, as the objective $\mathcal{L}$. The error term at $\mathbf{h}_\ell$, i.e., the gradient of surprisal with respect to the output of the ℓ -th layer, is the following:

$$\nabla_{\mathbf{h}_\ell} \mathcal{L} = \left(\frac{\partial \mathbf{h}_L}{\partial \mathbf{h}_\ell} \right)^\top \nabla_{\mathbf{h}_L} \mathcal{L}. \quad (1)$$

Intuitively, the 2-norm of this gradient, $\|\nabla_{\mathbf{h}_\ell}\|$, measures how strongly the model would need to change the representation at layer ℓ to increase the probability of the observed next token.

This paper focuses on the output of each transformer layer. As a concrete instantiation, we consider the GPT-2¹ architecture (Radford et al., 2019) and compute gradients at the output of its 12 transformer layers, as well as at the output of the final layer normalization module. The error term at the output of the final layer normalization is easily interpretable. It can be shown that, at the output of the final layer normalization, $\mathbf{h}_{\text{LN}}$, the norm of the error term corresponds to the Euclidean distance between the embedding of the observed next token and the expected embedding weighted by the model’s estimated next token probability:

$$\|\nabla_{\mathbf{h}_{\text{LN}}}\mathcal{L}\| = \|\mathbf{E}(\mathbf{p} - \delta_{w_t})\| \quad (2)$$

where $\mathbf{E} \in \mathbb{R}^{|V| \times d}$ is the output embedding matrix, δ_{w_t} is the Kronecker delta (one-hot) vector for the next token w_t , and $\mathbf{p}$ is the model’s predicted next token probability. It thus measures the semantic mismatch between the observed next token and the model’s expectation. Note that this measurement is closely related to the similarity-adjusted surprisal introduced by Meister et al. (2024):

$$H_z(w_t) = -\log \sum_{w' \in V} z(w_t, w') P(w' | w_{<t}) \quad (3)$$

where $z(w_t, w')$ is some measurement of similarity between w_t and w' . If one chooses the cosine similarity between w_t and w' as the similarity function:

$$z(w_t, w') = \frac{1}{2} \left(\frac{(\mathbf{E} \delta_{w_t})^\top (\mathbf{E} \delta_{w'})}{\|\mathbf{E} \delta_{w_t}\| \|\mathbf{E} \delta_{w'}\|} + 1 \right), \quad (4)$$

then the similarity-adjusted surprisal, $H_z(w_t)$, becomes:

$$\begin{aligned} -\log \left(\frac{1}{2} + \sum_{w' \in V} \frac{\delta_{w_t}^\top \mathbf{E}^\top \mathbf{E} \delta_{w'}}{2 \|\mathbf{E} \delta_{w_t}\| \|\mathbf{E} \delta_{w'}\|} P(w' | w_{<t}) \right) \\ = -\log \left(\frac{1}{2} + \frac{1}{2} \mathbf{p}^\top \tilde{\mathbf{E}}^\top \tilde{\mathbf{E}} \delta_{w_t} \right) \end{aligned} \quad (5)$$

where $\tilde{\mathbf{E}}$ denotes the row-normalized $\mathbf{E}$. Comparing this with the 2-norm of the error term at the

final layer normalization in equation 2,

$$\left(\|\mathbf{E} \mathbf{p}\|^2 + \|\mathbf{E} \delta_{w_t}\|^2 - 2 \cdot \mathbf{p}^\top \mathbf{E}^\top \mathbf{E} \delta_{w_t} \right)^{\frac{1}{2}} \quad (6)$$

one can see that both of them depend on $\mathbf{p}^\top \mathbf{E}^\top \mathbf{E} \delta_{w_t}$. If we assume that all word embeddings have equal norms of r ,² then the norm in equation 6 can be bounded by a monotone function of $\mathbf{p}^\top \mathbf{E}^\top \mathbf{E} \delta_{w_t}$:

$$(2r^2 - 2 \cdot \mathbf{p}^\top \mathbf{E}^\top \mathbf{E} \delta_{w_t})^{\frac{1}{2}} \quad (7)$$

which resembles the similarity-adjusted surprisal except for the difference between logarithmic and square root scaling (see appendix D for the empirical correlation).

The error signals backpropagated more deeply into the network at layer ℓ is:

$$\nabla_{\mathbf{h}_\ell} \mathcal{L} = \left(\frac{\partial \mathbf{h}_{\text{LN}}}{\partial \mathbf{h}_\ell} \right)^\top \mathbf{E}(\mathbf{p} - \delta_{w_t}). \quad (8)$$

Rather than directly capturing the semantic mismatch at the lexical level, they represent how the semantic mismatch is situated within the model’s representation of the preceding context. In the following sections, we explain how we evaluate the psycholinguistic predictive power of the proposed predictors.

3 Evaluating Predictive Power of Error

3.1 Human Behavioral Data

We evaluate the psycholinguistic predictive power of the error gradient norms on human reading times. We use four English reading time datasets, including both self-paced reading (SPR) where participants read a passage word-by-word, and eye tracking (ET) measurement. For SPR datasets, we consider Natural Stories (Futrell et al., 2018) and Brown (Smith and Levy, 2013) corpus. For ET datasets, we consider Dundee (Kennedy et al., 2003) and Provo (Luke and Christianson, 2018) corpus. For eye tracking measurements, we evaluate on both first-past duration (FP) and go-past duration (GP). Each dataset is partitioned into fit, exploratory, and held-out sets based on the sum of subject ID and sentence ID. The fit set is used to estimate the model’s coefficients. The exploratory

¹Although we use GPT-2 to illustrate the idea, the proposed predictor can be derived from any neural language models that have the concept of hidden layers.

²Though note that strictly this is not true and see Oyama et al. (2023) and Kobayashi et al. (2020) for discussions of relations between norm of word embedding and the word’s lexical properties.

partition is for exploratory analysis and development. The held-out set is reserved for the final significance testing.

3.2 Predictors

Norm of error To compute the norm of error at a word w_t , assuming that the sequence of tokens is $(w_1, w_2, \dots, w_t)$, we compute $\|\nabla_{\mathbf{h}_\ell} - \log P(w_t | w_{<t})\|_2$.³ For this experiment, we consider the gradient at the output of all 12 hidden layers of GPT-2 small as well as the gradient at the output of the final layer normalization layer. We choose GPT-2 small since it was shown that the surprisal derived from GPT-2 small has the best psycholinguistic predictive power among the GPT-2 model family (Oh and Schuler, 2023). This allows us to compare the norm of gradient with the surprisal derived from the same language model. For words composed of multiple tokens, we take the sum of the token-level measurement. When the complete material does not fit in the context window of GPT-2, we use a sliding window with 50% overlap.

Surprisal We include surprisal estimated by GPT-2 small as one of the baseline predictors for all the regression experiments. The surprisal estimator is derived using the same approach as the norm of error except that whitespace-trailing decoding is used to derive the word-level surprisal (Oh and Schuler, 2024).

Other baseline predictors We include word length (number of character) and word position in sentence as baseline predictors for all regression experiments. We also include unigram surprisal (Heafield et al., 2013) as one of the baseline predictors.

3.3 Evaluation

We evaluate the predictive power of error norms by first fitting two linear regression models that predict reading time (RT) using item-level predictors on the fit partition. The baseline model includes all the baseline predictors described in section 3.2. The full model includes all the baseline predictors and the norm of error term at a hidden layer. To account for spillover effects, we include each predictor (except word position) at the previous position in the

³ $\mathbf{h}_\ell$ here is the hidden representation at token $t - 1$.

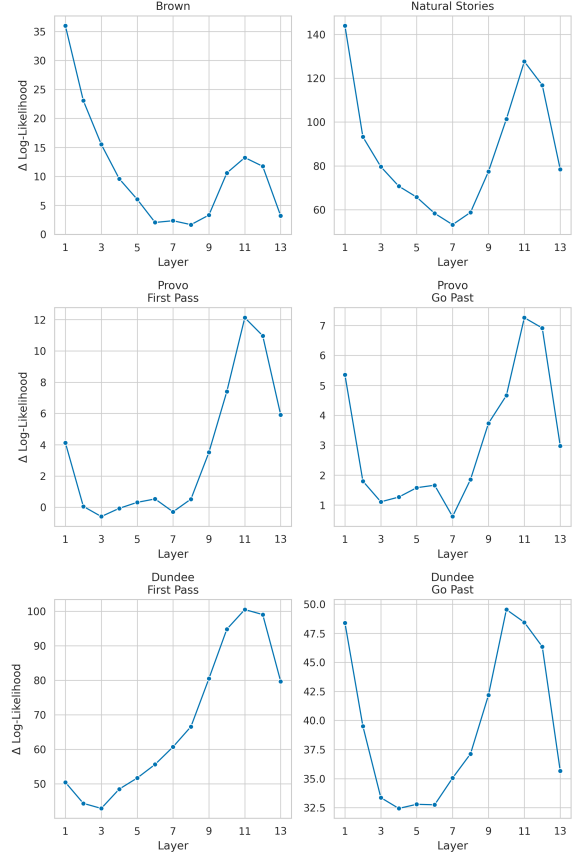


Figure 1: Δ Log-likelihood of adding the norm of the error terms at different layers of GPT-2 small as predictors of human RTs over surprisal baseline on the exploratory partition. Layer 1 means the output of the first transformer module. Layer 12 means the output of the last transformer module. Layer 13 means the output of the final layer normalization.

regression.⁴

The predictive power of the proposed predictors is measured by the difference in log-likelihood between the baseline regression model and the full regression model that additionally includes the proposed predictor on the held-out partition of each dataset:

$$\Delta LL = LL_{\text{full}} - LL_{\text{base}}. \quad (9)$$

The log likelihood of the held-out partition is estimated by assuming Gaussian noise around the model prediction with variance estimated on the fit partition:

$$\sum_{(y, \mathbf{x}) \in \mathcal{D}_{\text{held}}} \log \mathcal{N}(y | \beta^\top \mathbf{x}, \sigma^2) \quad (10)$$

⁴The complete formula is $\text{rt} \sim \text{sentpos} + \text{wlen} + \text{wlenS1} + \text{unigramsurp} + \text{unigramsurpS1} + \text{totsurp} + \text{totsurpS1} + \text{gradnorm} + \text{gradnormS1}$

	Δ LL	Δ MSE
Natural Stories	67.9	-18.1**
Brown	20.8	-37.2*
Dundee (First Pass)	81.9	-59.5***
Dundee (Go Past)	50.0	-132.5***
Provo (First Pass)	5.5	-14.5 ^{ns}
Provo (Go Past)	8.3	-103.0 ^{ns}

Table 1: Δ Log-likelihood and Δ mean squared errors of adding the norm of the error terms at layer 11 of GPT-2 small as predictors of human RTs over surprisal baseline on the held-out set. Statistical significance is determined by permutation testing on squared errors.

where y and x are the reading time and the predictors, β is the regression coefficient, and σ^2 is the variance estimated on the fit partition. We also compute the difference in the mean squared error, Δ MSE, on the held-out partition for significance testing.

3.4 Results

Figure 1 presents the Δ LL due to including the norm of gradient at different layers beyond the baseline predictors on the exploratory partition of each dataset. Note that, except for Provo first pass duration, the norms of gradient at all layers improve RT prediction. The two self-paced reading datasets and the go past duration of the two eye tracking datasets show a peak at layer 1 and then another peak around layer 11. On the other hand, the two go-past duration datasets show smaller effects at early layers and peak at layer 11. We suspect the reason is that the two first-pass measures, which mainly reflect early processing, depend less on the representation of the local context. Note that the norm of the gradient closest to the output, which resembles the similarity-adjusted surprisal calculated in Meister et al. (2024), does not have the strongest predictive power for human RTs.

Table 1 presents the Δ LL and Δ MSE due to including norm of gradient at layer 11 as the additional predictor. Including the proposed predictor improves RT prediction in all four datasets.⁵ According to paired permutation tests on squared

⁵The log-likelihoods of the surprisal baselines are -196665, -1253875, -307939, -338910, -178172, and -197657 for Brown, Natural Stories, Dundee FP, Dundee GP, Provo FP, and Provo GP respectively. The gains of surprisal over the other control predictors are 96, 175, 225, 163, 50, and 54 for Brown, Natural Stories, Dundee FP, Dundee GP, Provo FP, and Provo GP respectively. Likelihood ratio tests show that the differences are significant.

errors, we observed that the improvement is significant except for Provo FP and GP.⁶ To determine whether the effect generalizes across datasets of this kind, we applied a binomial test on the number of significant datasets. The probability of observing three or more significant results by chance is about 0.00048.

To determine whether surprisal and norm of gradient explain complementary variance of human reading times, we conduct a set of reverse hierarchical regressions (by treating the norm of gradient as the baseline predictor and surprisal as the target predictor) and find that surprisal still contributes beyond the norm of error baseline. When comparing norm of gradient and surprisal head to head, we observe that norm of gradient produces better higher log-likelihood for Brown, Natural Stories, Dundee FP, and Dundee GP but produces lower log-likelihood for Provo FP and GP.⁷

We compare the norm of gradient at layer 11 to the similarity-adjusted surprisal (Meister et al., 2024) via permutation testing on the held-out set. The norm of gradient at layer 11 shows significance improvement on Natural Stories and Dundee.⁸ We observe numerical improvement but non-significant difference in Brown and Provo.⁹

4 Related Work

Prior work has explored the relation between gradient-based saliency and the pattern of fixation duration in human reading and found that the relative importance of tokens measured by token saliency in masked language models (MLM) is correlated with human reading patterns (Hollenstein and Beinborn, 2021). This is different from our work in that they directly measure how important each individual token is in predicting the context while we measure how much the representation of the context would have to change at different layers of an autoregressive language model.

Prior work has illustrated the role of predictive coding in language comprehension (Bornkessel-

⁶ $p < 0.01$ for Natural Stories, $p < 0.05$ for Brown, $p < 0.001$ for Dundee FP, and $p < 0.001$ for Dundee GP.

⁷ Δ MSE = -3.2 for Brown, Δ MSE = -12.1 for Natural Stories, Δ MSE = -33.4 for Dundee FP, Δ MSE = -63.5 for Dundee GP, Δ MSE = 28.4 for Provo FP, Δ MSE = 103.7 for Provo GP.

⁸ Δ MSE = -10.7, $p = 0.03$ for Natural Stories; Δ MSE = -23.9, $p = 0.005$ for Dundee FP; Δ MSE = -77.7, $p = 0.0005$ for Dundee GP.

⁹ Δ MSE = -18.5 for Brown; Δ MSE = -19.4 for Provo FP; Δ MSE = -94.5 for Provo GP.

Schlesewsky and Schlesewsky, 2019; Kuperberg and Wlotko, 2018; Lewis and Bastiaansen, 2015; Michaelov et al., 2023; Eddine et al., 2023; Wang et al., 2022). Another line of work explored the relationship between change in latent state and prediction error and processing difficulty (Aurnhammer and Frank, 2019; Brouwer et al., 2016; Ettinger et al., 2016; Fitz and Chang, 2019; Michaelov and Bergen, 2023; Parviz et al., 2011b; Rabovsky et al., 2018; Rabovsky and McRae, 2014; Venhuizen et al., 2019). Most of that work focus on controlled materials while we explore broad-coverage naturalistic reading.

5 Conclusion

This paper explores the psycholinguistic predictive power of the norm of the error backpropagated to different layers of a transformer LLM, as a model of effortful state change during sentence processing. Regression experiments on reading time data show that the norm of the gradient at intermediate layers improve the fit to human reading time data numerically and the improvement is significant on three out of four datasets that we explored.

Limitations

This paper only explores reading time data in English. Further work is needed to investigate whether the effect of the norm of gradient generalize to other languages.

References

- Christoph Aurnhammer and S. Frank. 2019. [Evaluating information-theoretic measures of word prediction in naturalistic sentence reading](#). *Neuropsychologia*, page 107198.
- Ina Bornkessel-Schlesewsky and Matthias Schlesewsky. 2019. [Toward a neurobiologically plausible model of language-related, negative event-related potentials](#). *Frontiers in Psychology*, 10.
- Harm Brouwer, Matthew W. Crocker, Noortje J. Venhuizen, and John C. J. Hoeks. 2016. [A neurocomputational model of the n400 and the p600 in language processing](#). *Cognitive Science*, 41:1318 – 1352.
- Samer Nour Eddine, Trevor Brothers, Lin Wang, Michael W. Spratling, and Gina R. Kuperberg. 2023. [A predictive coding model of the n400](#). *Cognition*, 246:105755 – 105755.
- Allyson Ettinger, Naomi H Feldman, Philip Resnik, and Colin Phillips. 2016. [Modeling n400 amplitude using vector space models of word representation](#). *Cognitive Science*.
- Hartmut Fitz and Franklin Chang. 2019. [Language erps reflect learning through prediction error propagation](#). *Cognitive psychology*, 111:15–52.
- Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. [The natural stories corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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*.
- 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 (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.
- Nora Hollenstein and Lisa Beinborn. 2021. [Relative importance in sentence processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.
- Alan Kennedy, Robin Hill, and Joël Pynte. 2003. [The dundee corpus](#). In *Proceedings of the 12th European conference on eye movement*.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. [Attention is not only a weight: Analyzing transformers with vector norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.
- Gina R. Kuperberg and Edward W. Wlotko. 2018. [A tale of two positivities and the n400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation](#). *Journal of cognitive neuroscience*, 32:12 – 35.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Ashley Glen Lewis and Marcel C. M. Bastiaansen. 2015. [A predictive coding framework for rapid neural dynamics during sentence-level language comprehension](#). *Cortex; a journal devoted to the study of the nervous system and behavior*, 68:155–68.
- 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, Miami, Florida, USA. Association for Computational Linguistics.
- James A. Michaelov, Megan D. Bardolph, Cyma K Van Petten, Benjamin K. Bergen, and Seana Coulson. 2023. [Strong prediction: Language model surprisal explains multiple n400 effects](#). *Neurobiology of Language*, 5:107 – 135.
- James A. Michaelov and Benjamin K. Bergen. 2023. [Ignoring the alternatives: The n400 is sensitive to stimulus preactivation alone](#). *Cortex; a journal devoted to the study of the nervous system and behavior*, 168:82–101.
- Beren Millidge, Alexander Tschantz, and Christopher L Buckley. 2022. Predictive coding approximates backprop along arbitrary computation graphs. *Neural Computation*, 34(6):1329–1368.
- 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, Miami, Florida, USA. Association for Computational Linguistics.
- Chiebuka Ohams, Sathvik Nair, Shohini Bhattachali, and Philip Resnik. 2026. A predictive coding model for online sentence processing. *Journal of Memory and Language*, 146:104705.
- Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. 2023. [Norm of word embedding encodes information gain](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2108–2130, Singapore. Association for Computational Linguistics.
- Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. 2011a. [Using language models and latent semantic analysis to characterise the n400m neural response](#). In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46, Canberra, Australia.
- Mehdi Parviz, Mark Johnson, Blake W. Johnson, and Jon Brock. 2011b. [Using language models and latent semantic analysis to characterise the n400m neural response](#). In *Australasian Language Technology Association Workshop*.
- Milena Rabovsky, Steven Stenberg Hansen, and James L. McClelland. 2018. [Modelling the n400 brain potential as change in a probabilistic representation of meaning](#). *Nature Human Behaviour*, 2:693 – 705.
- Milena Rabovsky and Ken McRae. 2014. [Simulating the n400 erp component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning](#). *Cognition*, 132 1:68–89.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.
- Noortje J. Venhuizen, Matthew W. Crocker, and Harm Brouwer. 2019. [Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience](#). *Discourse Processes*, 56:229 – 255.
- Lin Wang, Lotte Schoot, Trevor Brothers, Edward J. Alexander, Lena Warnke, Minjae Kim, Sheraz Khan, Matti S. Hämäläinen, and Gina R. Kuperberg. 2022. [Predictive coding across the left fronto-temporal hierarchy during language comprehension](#). *Cerebral cortex*.

	Δ Log-likelihood
Natural Stories	72.6**
Brown	50.8***
Dundee (First Pass)	140.5**
Dundee (Go Past)	69.7**
Provo (First Pass)	6.4 ^{ns}
Provo (Go Past)	6.3 ^{ns}

Table 2: Δ Log-likelihood of adding the norm of the error terms at layer 11 of GPT-2 small as predictors of human RTs over surprisal baseline using cross-validation. All predictors at the previous position except word position are included to account for spillover. Δ Log-likelihood values are sum of the 10 folds. Statistical significance is determined by paired permutation tests on log-likelihood across folds.

A Evaluation with cross-validation

In this section, we evaluate the predictive power of the 2-norm of gradient using 10 fold cross-validation to replicate Meister et al. (2024). The datasets, predictors, and models are kept the same as in section 3. For each dataset, we take the average RT across subjects for each item in the dataset and split the items into 10 folds randomly. In each iteration of cross-validation, one of the folds serves as the held-out set and the remaining folds serve as the training set. This produces 10 pairs of held-out log-likelihood for the baseline model and the full model. We can then compute the ΔLL due to including the norm of gradient by taking the difference between the sum of log-likelihood across folds for the full and baseline model. The statistical significance of the improvement can be determined by paired permutation tests on the log-likelihood of the 10 folds.

Table 2 presents the ΔLL due to including the norm of gradient at layer 11. The results are consistent with the results reported in section 3. The error norm at layer 11 provides significant predictive power beyond surprisal on Natural Stories, Brown, and Dundee. The predictor improves the fit on Provo numerically although the improvement is not statistically significant. Figure 2 presents the trends of predictive power at different layers. The trends are also similar to those presented in Section 3.

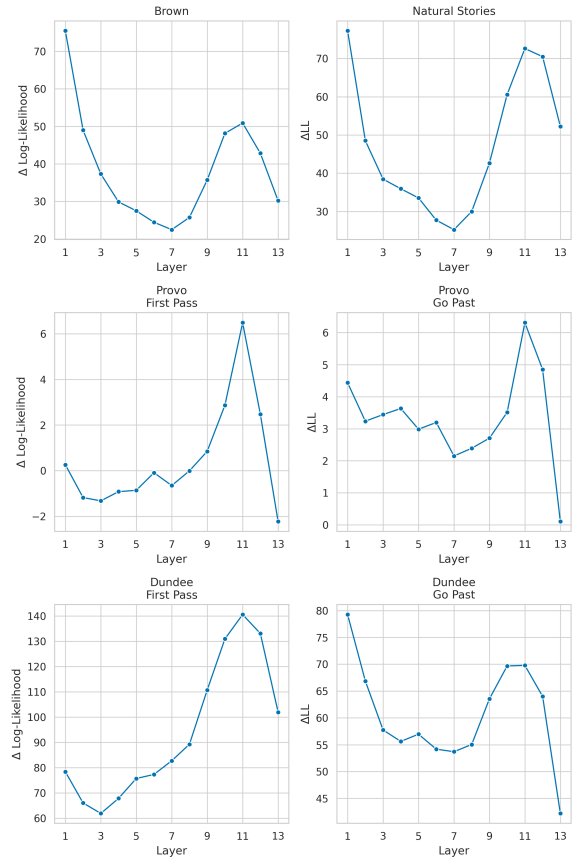


Figure 2: Δ Log-likelihood of adding the norm of the error terms at different layers of GPT-2 small as predictors of human RTs over surprisal baseline using linear regression. The ΔLL s are sums of the 10 folds using cross-validation. The layer indexing follows the definitions in Figure 1.

B Evaluation with linear mixed-effects regression

In this section, we evaluate the predictive power of the norm of gradient at layer 11 using linear mixed-effects regression and cross-validation.

Baseline predictors We include word length and word position in sentence as baseline predictors for the self-paced reading datasets. For eye tracking datasets, we additionally include saccade length in words and whether the previous word was fixated as baseline predictors.

Linear mixed-effects regression We include random intercept by subject. For self-paced reading datasets, we include random slopes for word length, surprisal, and norm of gradient. For eye tracking datasets, we additionally add random slopes for saccade length. The complete formulas are in appendix C.

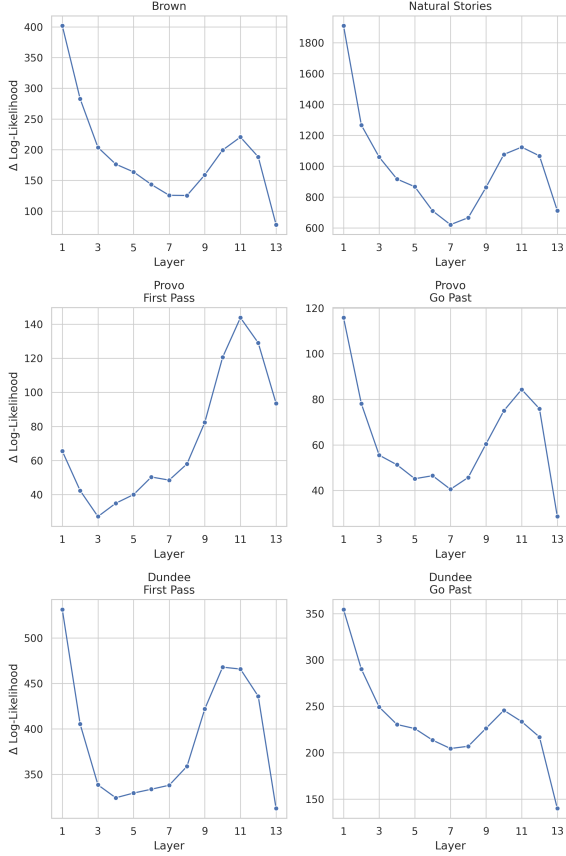


Figure 3: Δ Log-likelihood of adding the norm of the error terms at different layers of GPT-2 small as predictors of human RTs over surprisal baseline using linear mixed-effects regression (no spillover). The Δ LLs are sums of the 10 folds using cross-validation. The layer indexing follows the definitions in Figure 1.

Cross-validation Each dataset is partitioned into 10 folds according to sum of subject ID and sentence ID modulo 10. The Δ LL is computed by taking the difference between the sum of log-likelihood of the 10 folds of the full model and that of the baseline model. The statistical significance of the improvement is determined by paired permutation tests on the log-likelihoods across folds.

Results Table 3 presents the Δ LL due to including the norm of gradient at layer 11 compared to the surprisal baseline. Paired permutation tests on log-likelihoods across folds show that it provides significant predictive power beyond a surprisal baseline on all four datasets. Figure 3 shows the trends of predictive power of the error norms at different layers.

	Δ Log-likelihood
Natural Stories	1123.1***
Brown	220.5*
Dundee (First Pass)	465.7***
Dundee (Go Past)	233.5***
Provo (First Pass)	143.8**
Provo (Go Past)	84.3***

Table 3: Δ Log-likelihood of adding the norm of the error terms at layer 11 of GPT-2 small as predictors of human RTs over surprisal baseline using linear mixed-effects regression. Δ Log-likelihood values are sum of the 10 folds. Statistical significance is determined by paired permutation tests on log-likelihood across folds.

C Linear mixed-effects regression formula (in appendix B)

The baseline model for self-paced reading datasets (Natural Stories and Dundee) is:

$$\begin{aligned}
 rt &\sim z.(sentpos) + z.(wlen) + \\
 &\hookrightarrow z.(wlenS1) + z.(totsurp) + \\
 &\hookrightarrow z.(totsurpS1) + (1 + z.(wlen)) + \\
 &\hookrightarrow z.(wlenS1) + z.(totsurp) + \\
 &\hookrightarrow z.(totsurpS1) \mid \text{subject}) + 1.
 \end{aligned}$$

The full model for self-paced reading datasets is:

$$\begin{aligned}
 rt &\sim z.(sentpos) + z.(wlen) + \\
 &\hookrightarrow z.(wlenS1) + z.(totsurp) + \\
 &\hookrightarrow z.(totsurpS1) + z.(errornorm) + \\
 &\hookrightarrow z.(errornormS1) + (1 + z.(wlen)) \\
 &\hookrightarrow + z.(wlenS1) + z.(totsurp) + \\
 &\hookrightarrow z.(totsurpS1) + z.(errornorm) + \\
 &\hookrightarrow z.(errornormS1) \mid \text{subject}) + 1.
 \end{aligned}$$

The baseline model for eye tracking datasets (Dundee and Provo) is:

$$\begin{aligned}
 rt &\sim z.(sentpos) + z.(wlen) + \\
 &\hookrightarrow z.(wdelta) + z.(prevwasfix) + \\
 &\hookrightarrow z.(totsurp) + z.(totsurpS1) + \\
 &\hookrightarrow z.(wlenS1) + z.(wdeltaS1) + (1 + \\
 &\hookrightarrow z.(wlen) + z.(wdelta) + \\
 &\hookrightarrow z.(totsurp) + z.(totsurpS1) + \\
 &\hookrightarrow z.(wlenS1) + z.(wdeltaS1) \mid \\
 &\hookrightarrow \text{subject}) + 1
 \end{aligned}$$

The full model for eye tracking datasets is:

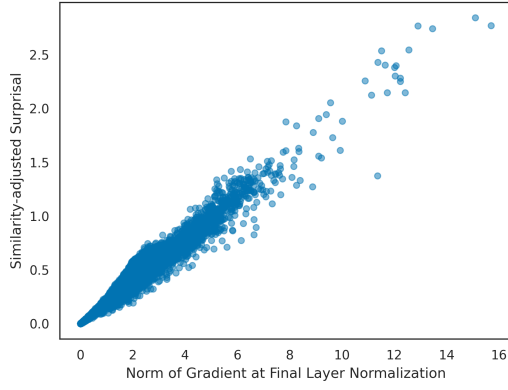


Figure 4: The relationship between similarity-adjusted surprisal derived from GPT-2 small’s embeddings and the error norm at final layer normalization layer on the Natural Stories corpus.

```

rt ~ z.(sentpos) + z.(wlen) +
  ↪ z.(wdelta) + z.(prevwasfix) +
  ↪ z.(totsurp) + z.(totsurpS1) +
  ↪ z.(gradnorm5) + z.(gradnorm5S1)
  ↪ + z.(wlenS1) + z.(wdeltaS1) + (1
  ↪ + z.(wlen) + z.(wdelta) +
  ↪ z.(totsurp) + z.(totsurpS1) +
  ↪ z.(gradnorm5) + z.(gradnorm5S1)
  ↪ + z.(wlenS1) + z.(wdeltaS1) |
  ↪ subject) + 1

```

D Empirical correlation between similarity-adjusted surprisal and error norms at the final layer normalization

Figure 4 shows the relationship between norms of gradient at final layer normalization of GPT-2 small and similarity-adjusted surprisal using GPT-2 small’s embedding-based cosine distance on the Natural Stories corpus. The Pearson correlation coefficient between them is 0.968.